ACTES DU XIIème CONGRES INTERNATIONAL DES SCIENCES PHONETIQUES

19-24 août 1991 - AIX-EN-PROVENCE, FRANCE

Vol 1

PROCEEDINGS OF THE XIIth INTERNATIONAL CONGRESS OF PHONETIC SCIENCES

AIX-EN-PROVENCE, FRANCE - August 19-24, 1991

TEXTES / *PAPERS*

Symposium
"Auditory integration"

Peter W. Jusczyk
(Bertrand Delgutte, Robert E. Remez*,
James R. Sawusch*, Jean-Luc Schwartz*)

Conference

Victor W. Zue*

(Renato de Mori*, John Laver, Michael Studdert-Kennedy)

CONFERENCE

SYMPOSIUM

"Links between speech production and acoustics."

CONFERENCE

THE INTEGRATION OF PHONETICS AND PHONOLOGY

JOHN J. OHALA

University of Alberta, Edmonton, Canada
and
University of California, Berkeley, California, USA

ABSTRACT

For two millennia the phonetic sciences have been united in an attempt to understand the structure and behavior of speech. Questions and methods related to language history, speech pathology, and speech technology freely mixed. In the early 20th century a split developed between phonetics and phonology. In this paper I argue that the split was ill-conceived and based on a misapprehension of the aims and potential of phonetics (re-named "integrative phonology" here). None of the sub-disciplines in the phonetic sciences are on such a sure footing as to have the luxury of "going it alone." A reconciliation should be based on a frank admission that the great questions common to all the phonetic sciences remain unanswered and need a cooperative effort for their resolution.

1. INTRODUCTION

In the beginning there were no disciplines -- only people who asked questions and wondered about the make-up and workings of the universe, including the universe of speech: How did speech originate? What is the nature of speech? How are spoken words made different from each other? What is the origin of different languages or how does it happen that the "same" word is pronounced differently by different speakers or even by the same speaker in different contexts? How can one best learn another language? How can speech defects be corrected? How does sound come be associated with meaning? How can we control and extend the power of speech: evoke its sense with writing, transmit it over great distances, make inanimate objects respond to the spoken command? The biblical stories of Adam naming the beasts and the tower of Babel story are candidate answers to these implicit questions. The story of Ali Baba gaining access to the cave with the spoken command "open, sesame" reflects Man's desire to control machines using speech. The various writing systems of the world -- some having considerable antiquity -- testify to the ability of untutored people to analyze words into the sound elements that make them different.[1] Other cultures and other ancient texts offer different candidate answers to the same questions. Panini's grammar of Sanskrit, written some 25 centuries ago, gives answers to the question of what the nature of speech is; he specifies its articulatory correlates, a descriptive method we use to this day. Greek speculations on language dealt with the development of words and the association of sound and meaning.

How did the phonetic sciences develop from such questions and analyses? In this paper I propose to give not a formal history of the phonetic sciences but a few historical vignettes which will serve to remind us of some of the roots of our field and especially to shed some light on the emergence of phonetics and phonology as separate enterprises.

2. VIGNETTES FROM THE HISTORY OF THE PHONETIC SCIENCES

It is fascinating to discover the diverse origins of any field. Geology, for example, can trace its beginnings to biblical interpretations, the study of gems, minerals, and fossils, mining, cartography and astronomy, as well as traditional descriptions of the earth's surface by travellers. I doubt, though, that there can be few other fields with such a diversity of parent disciplines as the phonetic sciences. These include medicine covering also anatomy, physiology, and speech pathology; physics and engineering; zoology and ethology; language teaching; music and voice training; philology (the study of the history and interpretation of texts); grammar and rhetoric; psychology (including developmental studies); archeology and anthropology; stenography and spelling reform.

Ancient (and much modern) literature is filled with purely speculative answers to the above questions about speech. Some of these speculations are impressive for their ingenuity and occasionally for their congruence with modern findings (not that we should uncritically take that as a measure of success). But it is true in the phonetic sciences as in all others: many theories are offered, few receive empirical support. Significant advances require speculation coupled with supporting evidence.

Many of the notable early studies of speech were done by medical people because physicians were inclined to be empirical in their work, drawing conclusions based on direct experience with their patients. Unlike others, their livelihood depended on their being able to get results, not just elegantly turned arguments. As Galen, the 2nd century AD Greek physician and anatomist, remarked on questions of anatomy and physiology, Aristotelian philosophers preferred disputation to dissection [10].

Galen is perhaps the earliest "hands on" practitioner of the phonetic sciences known to us. Apparently on the basis of first-hand observations, he elucidated the respiratory element of speech and discovered the cerebral source of the recurrent nerve (the principal motor nerve of the larynx) which had previously been thought to come from the heart [60].

Among other notable medical people who made contributions to the phonetic sciences are the Italians Hieronymous Fabricius, also known as Fabricius ab Aquapendente, (ca. 1533-1619) [17] and his student Julio Casserio (ca. 1552-1616) [9]; the Englishman William Holder (1616-1698) [30]; the Swiss, Johann Conrad Amman (1669-1724) [1, 2] and Albrecht von Haller (1708-1777) [28]; the Germans Johannes Müller (1801-1858) [51], Emil Du Bois-Reymond (student of Müller; 1818-1896) [15], Hermann von Helmholtz (1821-1894) [29], and Ernst Brücke (1819-1892) [7][2]; the Czechs, Jan Purkyně (1787-1869)[3], Johann Nepomuk Czermak (1828-1873) [12]; the Frenchmen Denis Dodart (1634-1707) and Antoine Ferrein (1693-1769) [19]; the Dutch F. C. Donders (1818-1889) [14] and Hendrik Zwaardemaker (1857-1930) [83, 84].

Of these, Holder and Amman and the English mathematician John Wallis (1616-1703) were motivated in their study of speech by their attempts to teach the deaf to speak [1, 2, 30, 79].[4] Their pioneering works were quite influential for more than a century afterwards. Amman's work, which was translated into English, French, and German, exhibits some remarkably novel observations; for example, regarding the lateral "semi-vowel" [1], he notes [1] (pp. 52-53) that it

> is formed when the Tongue is so applied to the Roof, and the upper Teeth, that the Voice cannot, but by a small Thred, as it were, get forth by the Sides of the Tongue; for if

you compress the Cheeks to the Grinders, you stop up the Passage of the Voice, and it will be very difficult for you to pronounce this Letter,..."

This easily replicable do-it-yourself experiment demonstrates clearly the role of the buccal sulcus (the space between the cheeks and the molars) as a resonating cavity in speech (at least in some speakers). Amman is one of the first to attribute voice to the modulation imparted to the air stream passing through the glottis by the vibrations of the vocal cords [p. 29]. These vibrations he considered the 'substance' of speech; the 'form'[5] was imparted by "the various configurations of those hollow channels, thorough which they pass..." [p. 26]. This is one of the earliest and clearest expression of what we would now call the 'source - filter' model of speech. He also establishes an elementary binary, hierarchical classification of phonetic features which incorporates certain notions that might well be considered seriously by modern phonologists, e.g., that manner features dominate place of articulation features [p. 66]. He considered his system as a 'natural' hierarchical taxonomy and comments that substitutions of sounds (e.g., in pathological speech) involve similar sounds at the lowest strata of the hierarchy, not the highest, i.e., a dental 'semi-vowel' like l is substituted for another, r, or one nasal for another, i.e.., we don't see substitutions of vowels for consonants, etc. [pp. 66-67] It was also Amman (like Wallis before him) who made what might now be considered phonological observations: "If any word terminates in **n** and the following word begins with **b** or **p**, ... then in pronouncing the **n** we unconsciously change it, for the sake of euphony, into **m**, ..."

Amman was aware of the discrepancies between pronunciation and spelling but considered this primarily a fault of pronunciation. From our point of view this may be regarded as a confusion of spelling and sound but before we adopt a superior attitude, let us be sure we ourselves do not suffer from vestiges of the same confusion [57].

It was also a medical doctor, Christian Gottlieb Kratzenstein,[6] a German who lived and worked in Denmark, (1723-1795) who in 1780 was among the first to attempt the synthesis of speech and publish the results [37, 38].[7] Even though it concerned just isolated steady-state vowels, he did not yet have a clear idea of resonance, and his resonators bore little resemblance to the vocal tract (and thus didn't clarify how human vowels came about), he at least showed that mechanical synthesis of some speech sounds was possible.

It was, however, Wolfgang von Kempelen (1734-1804) a Hungarian engineer and a native of Vienna (part of the Austro-Hungarian Empire) who in 1791 [34] made one of the most influential pioneering contribution not only to speech synthesis but to phonetic science in general. His work *Mechanismus der menschlichen Sprache* gave complete blueprints (actually woodblock prints, and splendid ones at that) detailing the construction of a speaking machine [16]. It must be emphasized that it was not the speaking machine by itself which had such an impact on the field. Rather it was the combination of the machine plus the book he wrote describing it which had such great repercussions. His efforts represented a kind of step-function increase in the detailed attention given to all aspects of speech production. The book gives an impressive review of contemporary knowledge and speculation on speech and language. He discusses, among other things, animal communication, the sign language of the deaf, the origin of speech and language. He reviews the earlier work of Galen, Amman, van Helmont, Dodart, Ferrein, Haller, Herder, de Brosses, Court de Gebelin, Lord

Monboddo, Adelung, Abbé de l'Epée, and Kratzenstein. He gives a phonological comparison between languages, not only on their segment inventory but also with respect to their phonotactics (permissible clusters).

Erasmus Darwin (1731-1802), grandfather of Charles Darwin, a erudite, imaginative, and progressive "gentleman scientist" of the Enlightenment, dabbled in speech synthesis and constructed a mechanical synthesizer along the lines of von Kempelen although simpler in that it was capable only of labial sounds **p**, **b**, **m**, and the single vowel **a** ([13], pp. 119-120). In what must be one of the earliest proposed applications of phonetics to speech technology he suggested that his machine, "... if built in a gigantic form, might speak so loud as to command an army or instruct a crowd." In fact, this plan never would have worked because resonant frequencies are inversely proportional to the length of the vocal tract. A gigantic mouth would have had resonances so low and so close together (in frequency) that it is doubtful human ears could resolve them or recognize them as speech-like sounds. (However, it could have been possible in principle to make a speaking machine speak loud enough to address crowds by keeping the vocal tract the normal length but augmenting the lung force.) Darwin, although he was apparently not unaware of previous efforts by other writers, conducted his own analysis of the sounds of languages of the world and concluded that some 32 or 33 separate sounds might be recognized, including the Welsh **ll**. He also proposed that these sounds could be represented more simply by employing only 13 unary features which included the basic three places of articulation, oral resonance[8], nasal resonance, voiceless frication, voiced frication, etc. Since he found it difficult to determine the exact 'place of articulation' of vowels via kinesthesia, he devised a simple palatograph: "I rolled up some tin foil into cylinders about the size of my finger; and speaking the vowels separately through them [that is, inserting the cylinders into his mouth], found by the impressions made on them [that is, where they were dented], in what part of the mouth each of the vowels was formed...[p. 119]". This is one of the earliest instrumental phonetic studies performed on a live, intact, speaker.

One person seldom celebrated in the history of our field but who made several interesting contributions is the Englishman Thomas Young (1773-1829), also trained in medicine but who is most well known in the physical sciences for his demonstration of the wave nature of light. His minor dissertation written in Göttingen in 1795-1796 -- now lost -- was on the topic of universal phonetics: he proposed that all languages could be written phonetically using just 40 to 50 distinct letters. He was the first to decipher Egyptian hieroglyphics, a task completed for the most part by François Champollion. In an undeservedly-neglected paper of 1818, Young gave a mathematical account of the need to find several cognate words between languages in order to establish a family relationship. It is also to Young that we owe the coinage of the term 'Indo-European' (in a review of Adelung's *Mithradates*).

Robert Willis (1800-1875), a Cambridge professor of mechanics (engineering we would call it today) in his 1830 work "On the vowel sounds" [81] specified quantitatively the vocal tract resonances of vowels and claimed that the major determinant of vowels' characteristic acoustic patterns was vocal tract length. He also claimed that there were infinite vowel sounds and that one vowel faded gradually and imperceptibly into its neighbor in the series [i e a o u]. He remarked that with some refinement of his investigations he should be able to provide "philologists with a correct measure for the shades of differences in the pronunciation of the vowels by different nations." Although his single resonance model of vowels is not supported today it is reminiscent of the notion that one can specify a single "characteristic" resonance of most vowels and that this is equivalent to a weighted average of F2 and F3 for front vowels and is approximately F2 for low and back vowels [18].[9]

One of the more interesting things about Willis' work is a subsequent, paper it inspired by T. Hewitt Key (1799-1875) first professor of Latin, then professor of comparative philology, at London University (now University College). Key, trained in medicine and mathematics (and a teacher of math at the newly formed University of Virginia from 1825 to 1827) contributed several papers to the Philological Society of London on various specific sound changes and sound change types. His paper "On vowel-assimilation, especially in relation to Professor Willis's experiment on vowel-sounds" appeared in the Transactions of the Philological Society for the year 1852 (but which was published in 1855) [35]. In this paper Key tries to explain vowel harmony and umlaut by invoking Willis' notion that vocal tract length is the main articulatory difference between vowels. This would not be judged a successful attempt in the light of current knowledge but let us not engage in what's called 'Whig' history (historical events judged according to modern standards and tastes): it is an admirable effort at applying the latest phonetic theories to phonological problems. It also has some memorable and still pertinent quotes:

> [some scholars of language] have allowed themselves .. to be led astray by paying more attention to the symbols of sound than to sounds themselves. ... Scholars seldom unite the love of classical and scientific pursuits; and a paper [i.e., Willis'] of the highest value for philology might well fail to meet with all the attention it deserved from the students of language, when published in a series of treatises [*Transactions of the Cambridge Philosophical Society*] almost exclusively of a mathematical character; not but that the paper has an indisputable claim to such a position, since it treats the problem with the accuracy of modern physics.

Hermann Grassmann (1809-1877), Sanskritist and the discoverer of the well known Greek and Sanskrit dissimilatory sound changes which are named after him [27], devoted most of his energies in his prime to mathematics, not to philology which was a pursuit in his later life. This hero of the comparative method and inspiration for the neo-grammarians, also made a significant (but now generally neglected) contribution to acoustic phonetics apparently being the first person to declare that some vowels had two distinct resonances, not just one as taught by Willis. He determined these resonances by purely auditory means by identifying the number (and thus the pitch) of the prominent harmonics of intoned vowels much as so-called harmonic singers can manipulate individual harmonics of their voice. This work was published in 1854 [26], nine years before Helmholtz published similar findings using instrumental means.

Another well known comparative philologist who saw no bar to integrating physical studies of speech with philology is Karl Verner (1846-1896], discoverer of the famous sound law that bears his name [77]. Verner's Law states that medial voiceless fricatives became voiced unless the accent fell on the preceding syllable. In his later years Verner was interested in trying to find out how and why accent could influence segments in this way. He constructed on his own an elaborate optical device which permitted him to enlarge the speech tracks on an Edison phonograph cylinder and to project them on the wall such that he could make hand tracings of them and then measure and analyze them. In essence he measured periods to derive the pitch and did a Fourier analysis of the signal. As it turned out, he didn't get any results he thought worth publishing. His research wasn't made public until after his death [21, 33, 78].

Abbé Pierre-Jean Rousselot (1846-1924), often called the father of experimental phonetics, continued to some extent the tradition of physiological studies of movement initiated by E. J. Marey, physician and pioneer in the study of locomotion and the one who perfected the kymograph (with his invention known as "Marey's capsule"). In general, it would not be inappropriate to say that Rousselot attempted to do for speech what Helmholtz attempted to do for vision and hearing, i.e., reduce their function to known physical physiological principles. Indicative of his view of the broad integrative character of the phonetic sciences are two of his major works, one, his dissertation [66] which was an attempt in part to give an instrumental phonetic account of the sound changes which shaped the dialect spoken in his home town, and, two, the application of phonetics to the problems of the deaf [67].

Even more than individual effort, what really demonstrates the existence of a continuing *tradition* mixing physics, physiology, and philology is the way that different authors built on the work of others, as in the case of T. Hewitt Key applying Robert Willis' theory of vowel production to vowel harmony. Many other examples of this exist including the following two.

Von Kempelen's work was widely known and extremely influential throughout the 19th century; it was cited in virtually every subsequent major work on voice and speech. Wilhelm Jacobi (1816-?) in his 1843 [31] work on the history of the German language attempted to give an account of German ablaut by a complex quasi-mathematical scheme based on von Kempelen's description of the articulation of various vowels. Other philologically-oriented writers incorporating the best contemporary phonetics into their philological work include H. E. Bindseil (1803-1876) [4], Karl Moritz Rapp (1803-1883) [63], Rudolf von Raumer (1815-1876) [64], and Friedrich Techmer (1843-1891) [74].

A further potentially far-reaching chain of influence from von Kempelen and Helmholtz to Alexander Graham Bell (1845-1922) is well known [22]. Crucial links in this chain were, first, Sir Charles Wheatstone (1802-1875) who demonstrated to the young Bell his replica of von Kempelen's machine and loaned him his copy of von Kempelen's book and, second, Alexander J. Ellis (1814-1890) who was a friend and associate of Alexander Melville Bell (1819-1905), Graham Bell's father. Ellis tried to explain to Alexander Graham and his older brother Melville how Helmholtz had discovered the principal resonances of vowels and synthesized them using tuning forks. Alexander Graham, while still a teenager, along with his brother, constructed a speech synthesizer roughly along the lines of von Kempelen's, although incorporating more realistic anatomical detail. This experience along with the extensive knowledge of articulatory phonetics that he learned from his father, author of the influential system of self-interpreting physiologically-based phonetic transcription [3], gave Graham Bell the confidence to think that it should be possible to break speech down into some simpler form and transmit it across great distances. The rest, as they say, is history.

3. THE UNITY OF THE PHONETIC SCIENCES

What conclusion can be drawn from these snapshots from the early history of phonetic sciences? The conclusion I draw is that there had not yet been any hardening of the division of the phonetic sciences into largely separate sub-disciplines of phonetics and phonology and their applications in speech pathology and speech technology. Certainly those who studied speech pursued their research primarily in the way they were used to, depending on their background and training: medical, mathematical, physical, or philological, but with many interesting and enlightened excursions from one domain to another. There seemed to be a genuine belief in an idea that we tend to give only lip service to today: the underlying unity of all science--or at least of the phonetic sciences.

It is generally recognized that the separation of phonology and phonetics occurred as a result of the rise of structuralism, taught initially by Ferdinand de Saussure (1857-1913) and Jan Baudouin de Courtenay (1845-1929) but fully developed in phonology by the Prague School. N. S. Trubetzkoy (1890-1938) [75, 76], a leader of the Prague School, differentiated between "... the study of sound pertaining to the act of speech (phonetics) ... and the study of sound pertaining to the system of language (phonology)." Since the proper study of all of structural linguists was the system of language it followed from this (and is commonly believed today) that phonetics is not part of linguistics.[10] The emphasis on system or the relationship between speech sounds rather than on the substance of those sounds represented a new concern and one which seemed at the same time to open up new frontiers for phonological study and to liberate the study of speech sounds from physical phonetics and all the burdens of its natural sciences methods.

I admire and draw inspiration from the phonological work of Trubetzkoy and other phonologists in the tradition initiated by the Prague School. Indeed, some of Trubetzkoy's phonological generalizations were based on intuitive phonetic grounds (though he felt he had to apologize and explain at some length how this didn't imply that he thought precise phonetic correlates of sound contrasts mattered). But Trubetzkoy's conception of phonetics was something of a cartoon stereotype:

> La phonétique actuelle se propose d'étudier les facteurs matériels des sons de la parole humaine: soit les vibrations de l'air qui leur correspondent, soit les positions et les mouvements des organes qui les produisent. ... Le phonéticien est nécessairement atomiste ou individualiste ... Chaque son de la parole humaine ne peut être étudie qu'isolement, hors de tout rapport avec les autres sons de la même langue. ([75], pp. 232-233)

A similar stereotype applied to astronomy would characterize its proper activity as merely looking at and cataloguing stars. No mention would be made of cosmology, astrophysical theory, etc., i.e., attempts to generalize about the birth, development and death of stars, the formation of galaxies, the origin of the universe. This is the fallacy of equating the immediate, visible object of study to the ultimate object of study. Though the immediate object of study in phonetics (and in the psychological study of speech) may be the sounds and articulations of speech, the ultimate objects of study are the underlying causes of speech sound behavior, where "behavior" includes the same broad domain that Johann Amman studied three centuries ago, how laterals are produced, the assimilation of nasals to the place of articulation of following stops, the patterns of substitution of one speech sound for another.[11]

A possible advantage of the split of phonology from phonetics was the freedom of the phonologist to address issues more of a psychological or functional than a strictly physical phonetic nature. Also, it was possible to bring in a host of new ostensibly non-phonetic factors as the causes of speech sound behavior, e.g., structural "pressure" (the existence or non-existence in the language of similar contrasts).

But to carry through with such a program it would have been necessary to embrace some of the methods and concepts of psychology or perhaps certain aspects of the theory of communication. Unfortunately this was not done. Rather phonology was practiced as if it were an autonomous discipline owing little or nothing to other scientific domains.

And it was not just the domain of inquiry that phonology left behind after its divorce from phonetics; it also abandoned phonetics' approach to argumentation, i.e., its manner of bringing evidence to bear on theoretical claims. Over the decades the phonetic sciences had established a respectable degree of *accountability* in the way that generalizations and theories were proposed and defended. If anything, the degree of accountability in the field has been improved and tightened since then. As a result there is a relatively continuous and cumulative tradition on which to develop and refine both methods and theories. To give just one example, and one which has far-reaching implications for phonology and for the behavioral sciences in general: careful phonetic studies spanning a century have demonstrated, the tremendous amount of variation -- essentially infinite in character -- that exists in the speech signal [55, 59].

In contrast, autonomous phonology has yet to develop a tradition of accountability: it has enlarged the list of causal factors which it can cite to account for given phonological behavior -- structural pressure; maintenance of equilibrium in the total phonological system; striving for simplicity, naturalness or unmarkedness, learnability, etc. (and this is a positive move) -- but it has not enlarged its repertory of ways to insure the *quality* of evidence offered in support of its claims. Actually, by abandoning phonetic methods and by not adopting those from psychology, it has depleted its methodological arsenal. Freed from what it regards as the confinement of an "empiricist and mechanistic" approach to speech sounds, it can not only propose a completely new range of theories but even those which contradict phonetic findings: voiceless sounds can be called voiced, nasalized vowels can be called oral, distinctively aspirated stops can be treated as redundantly aspirated, closed syllables can be called open. None of this is inherently bad; throughout the history of science, claims which seem to fly in the face of common sense have proven their worth, e.g., that matter consists primarily of empty space. Nevertheless, at some point this and all claims must impinge on the tangible world, even if indirectly, e.g., (to continue the preceding example) by showing that most subatomic particles pass through metal sheets without being deflected. However necessary and valuable simplicity and generality of individual claims are and the degree to which they fit into a larger self-consistent theoretical framework, these properties by themselves never substitute for empirical support. It is disappointing is to see the almost complete disinterest of autonomous phonologists in the possible relevance for their claims of phonetic or psychological findings. For example, linguistics textbooks continue to characterize aspiration on /p t k/ in English as redundant and, to my knowledge, have never paid any attention to, or attempted to contradict, the evidence that aspiration is the principal auditory cue differentiating them from /b d g/ in initial position [44].

It may be objected that in spite of phonologists' statements about the difference between phonology and phonetics, there is a sense in which all phonological work in fact incorporates some phonetics insofar as it uses terms such as 'obstruent', 'voice', etc. However, I would like to differentiate between two forms of phonetics [56], one I call 'taxonomic' phonetics (for lack of a better term) and the other 'scientific' phonetics. Taxonomic phonetics has provided us with traditional phonetic terms and symbols used to describe and classify speech sounds and has remained essentially unchanged since the formation of the International Phonetic Association a century ago. Scientific phonetics, on the other hand, continues to change. It constantly expands its horizons; it develops new data, concepts, and methods; it rejects or revises earlier beliefs shown to be deficient, and, to the extent that these beliefs or theories have congruence with the universe, it has practical payoff, e.g., in language teaching, speech pathology, and speech technology. Of course, it also has payoff in phonology: how would we be able to make sense of the inherent tendency of obstruents towards voicelessness [54] if Husson's neurochronachtic theory of vocal cord vibration had not been effectively refuted. While autonomous phonology embraces taxonomic phonetics, for the most part it excludes scientific phonetics. A good bit of what is called and taught as "phonetics" in many universities -- if it is taught at all -- is exclusively taxonomic phonetics.

This is a pity because scientific phonetics is the intellectually most exciting form of the field-- and one of the most successful and rigorous within linguistics (if one allows, of course, that it is part of linguistics). It addresses issues of fundamental importance for phonology: how sounds differ from each other [39, 44, 70, 71], how sounds vary thus leading to sound change [24, 54, 55]. It is even possible in many cases to give principled reasons why sounds change in one way but not in others. Insofar as the causes of change can be located in the physical phonetic domain, it calls into question the common practice of assigning change to the grammar [57, 59].

The development of divisions and specialized branches of scholarly disciplines is common enough in the history of science, e.g., the basic division between statistics and pure mathematics. This happens naturally as the body of knowledge and methods in one area becomes too large for individual practitioners to master. This happened with organic and inorganic chemistry. Splits also occur as new questions arise. This happened in nuclear and (classical) physics. But in examining the causes of the split of phonology and phonetics, I conclude that it was based on a complete misunderstanding of what was termed "phonetics": an inability to see the forest for the trees.

4. INTEGRATIVE PHONOLOGY

What of the body of scholarship that autonomous phonology split off from -- that body of work that was decreed not to be phonological and by some not even part of linguistics? What shall we call it? 'Phonetics'? No, it was and is more than that. This tradition never really acquiesced to the claim that traditional phonological concerns -- to explain the behavior of speech sounds -- could or should be approached in an isolated, autonomous fashion. Therefore, I'll call it "integrative phonology". As I tried to argue and demonstrate, integrative phonology was accepted and practiced up to and throughout the 19th century. In spite of the supposed separation of phonetics and phonology triggered by the structuralist revolution, it was also practiced in this century by Zipf [82], Stetson [68, 69], Zwirner and Zwirner [85], Menzerath [48, 49], Grammont [25] -- to pick a few out of many such figures active in the first half of this century. In the last half of this century, we see the same principles in (and I must be forgiven for the brevity and unavoidable selectivity of the following list) Jakobson, Fant, and Halle's pivotal work, *Preliminaries to Speech Analysis* [32], in the work coming out of Haskins Laboratories [45] and the Pavlov Institute of Physiology [36] (especially the research on syllable structure), the Institute of Phonetics in Copenhagen [20], the work on speech sound universals by Ladefoged [39] and Maddieson [46], as well as contributions by Lehiste [40], Lindblom [41, 42, 43], Stevens [70, 71], Rossi [65], and Browman and Goldstein [6, 24].

Integrative phonology does not accept its proclaimed banishment from linguistics. It has not surrendered phonological questions to those who would pursue them in isolation of phonetics, psychology, and many other disciplines that can assist. In fact, in spite of Trubetzkoy's claim to the contrary, phonetics *has* developed methods and theories which address the functioning of speech sounds as elements of a system [41, 42, 44, 71]. The dividing line between phonetic, phonological, and psychological studies of speech sounds is quite blurred in much current research, e.g., that of Pisoni [61], Fowler [23], Massaro [47], Nearey [52, 53].

Integrative phonology does not solve problems by the unchecked proliferation of novel theoretical entities; rather, it attempts to keep the theoretical entities to a minimum and draws most of the building blocks of its theories from the realm of the previously established -- often that which has substantial empirical support. Its theories tend to contain within them an indication of how they could be tested and for the most part the first test is offered by the author of the theory.

I also think integrative phonologists have more fun with their research: they retain a kind of child-like curiosity about speech and like children often get their hands dirty and insert odd objects into their mouths and noses.

5. THE RECONCILIATION OF INTEGRATIVE AND AUTONOMOUS PHONOLOGY

The legacy of this divorce of autonomous phonology from integrative phonology six decades later is that a considerable gap has developed between them [11, 39]. An expression of this, perhaps inadvertent, is the frequently encountered collocation 'the interface between phonetics and phonology', where, as I have argued elsewhere [56], the term 'interface' incorrectly implies that the two disciplines are largely independent and autonomous. But if there is an apparent irreconcilable chasm between the two, even though both are trying to understand the same phenomenon, speech, we should entertain the possibility that one or both of them espouses unrealistic and indefensible positions. Perhaps there really isn't such an unbridgeable gap if we could just drop the extravagant claims. I make this proposal seriously: even astrology and astronomy could be reconciled if the empirically indefensible claims made by one side or another could be thrown out.

To start such a rapprochement between the two approaches to phonology I suggest that both sides should admit to the things that they are really not sure about.

What does integrative phonology know and not know about speech? Considerable lore about speech has accumulated over the centuries which permit reasonably complete descriptions of particular instances of speech. As a result it is possible to do speech synthesis by rule and to some extent speaker-independent speech recognition based on feature extraction. In spite of these successes, however, it must be admitted that we do not yet have a truly general theory of speech production and perception. For example, although there is strong belief that there is *some* fundamental concatenative unit underlying speech there is not much agreement on what it is. Various proposals exist: the phoneme, the phone (the same size as the phoneme but drawn from a much larger set since there is no posited functional identity between all the phones), the diphone, the demisyllable, the syllable, etc. It is even possible that more than one of these units are operable at different stages of speech production and speech perception [58]. Shockingly, there is not even complete agreement on the acoustic-auditory correlates of vowel quality: most believe that formant frequencies matter but acknowledge that absolute formant frequencies can't be crucial since these vary between speakers and even within speakers between different contexts. Much research is being done on trying to discover higher-order relationships between the formants [50]. There is some evidence that time-varying formant frequencies are important cues to vowel quality [52, 53, 73], i.e., that vowels consist of a trajectory through the vowel space rather than static points. Some reject formant frequencies and advocate whole-spectrum measures [5]. Related to this is a much more fundamental dispute over whether there are any truly context-invariant phonetic correlates of linguistic distinctions as opposed to context-sensitive cues [43, 72]. There is also no clear consensus on the causes of universals in speech sound systems, although there is informed speculation on this topic [39, 41, 42, 43, 70, 72]. The list of disputed issues is quite large.

It might be thought that if integrative phonology is unsure about such fundamental points then clearly it is in a weak position vis-a-vis autonomous phonology. But I take the controversies as a sign of strength and honesty; it would be much worse if the community of integrative phonologists just gullibly accepted claims based on their superficial plausibility, mere internal consistency, or their fashionableness, rather than on the rigorously gathered evidence supporting them. In any case, it is premature to judge integrative phonology weak because it knows what it doesn't know; we have yet to hear the confessions of ignorance from the autonomous phonologists.

6. CONCLUSION

In the final analysis, I think it will be found that everyone in the phonetic sciences, including autonomous and integrative phonologists, know very little about the same thing: how speech is structured and how it works. In other words we ask the same questions -- in fact, much the same questions as have been asked throughout history. When the divided parties realize that neither one has all the answers, they can cooperate in trying to resolve their common questions.

7. NOTES

1. Further evidence of this skill comes from comparison or linking of words by their constituent sounds

through the use of such poetic devices as rhyme, alliteration, assonance, the construction of rhyme tables, and establishing a conventional order of the elements of an alphabet or syllabary.

2. In 1871 Brücke published a work on phonetic aspects of verse [8] which included measurements of lip movements obtained with a device of his own invention. These are among the first instrumental phonetic recordings.

3. Purkyně's phonetic work was done in the 1830's but only rediscovered and published in the 1970's [62].

4. An interest in teaching the deaf to speak also motivated in part the research of Wolfgang von Kempelen and Alexander Graham Bell.

5. 'Substance' and 'form' are, of course, elementary notions in Artistotelian metaphysics.

6. Kratzenstein was well known for, among other things, his promotion of electricity for therapeutic purposes and for his advocacy of the two-fluid theory of electricity (in opposition to his contemporary, Benjamin Franklin's, one-fluid theory).

7. Von Kempelen, however, had begun the construction of his speaking machine in 1769.

8. Darwin used slightly different terms; I am 'translating' his terminology into their approximate modern equivalents.

9. By an odd coincidence Willis had an encounter with von Kempelen - though after the latter's death: Willis published an expose of von Kempelen's fake chess-playing automaton which was put on tour throughout Europe after the inventor's death [80].

10. In a widely disseminated directory of electronic mail addresses in North American and Europe, the header indicates that it lists the addresses of " ... linguists and a number of people in *related disciplines like phonetics ...*" [emphasis added].

11. There may have been some phoneticians who advocated a kind of extreme positivism, e.g., Scripture, but this was hardly characteristic of the whole sweep of the phonetic sciences in the early decades of this century and it certainly isn't true of phonetics today.

7. REFERENCES

[1] AMMAN, J. C. (1694), *"The talking deaf man: or, a method proposed whereby he who is born deaf may learn to speak"*, London: Tho. Hawkins.

[2] AMMAN, J. C. (1700), *"Dissertation de Loquela "*, Amsterdam: J. Wolters.

[3] BELL, A. M. (1867), *"Visible Speech"*, London: Simpkin, Marshall & Co.

[4] BINDSEIL, H. E. (1838), *"Abhandlungen zur allgemeinen vergleichenden Sprachlehre"*, Hamburg.

[5] BLADON, R. A. W. & LINDBLOM, B. (1981), "Modeling the judgment of vowel quality differences", *J. Acoust. Soc. Am. 69*, 1414-1422.

[6] BROWMAN, C. P. & GOLDSTEIN, L. M. (1986), "Towards an articulatory phonology", *Phonology Yearbook 3*, 219-252.

[7] BRÜCKE, E. (1856), *"Grundzüge der Physiologie und Systematik der Sprachlaute für Linguisten und Taubstummenlehrer"*, Vienna: Carl Gerold's Sohn.

[8] BRÜCKE, E. (1871), *"Die physiologischen Grundlagen der neuhochdeutschen Verskunst"*, Vienna: Carl Gerold's Sohn.

[9] CASSERIO, J. (1600-1601), *"De vocis auditusque organis historia anatomica"*, Padua.

[10] COLE, F. J. (1949), *"A history of comparative anatomy"*, London:

Macmillan.
[11] CUTLER, A. (1986), "Phonological structure in speech recognition", *Phonology Yearbook 3*, 161-178.
[12] CZERMAK, J. N. (1879), *"Gesammelte Schriften"*, Leipzig: Wilhelm Englemann.
[13] DARWIN, E. (1803), *"The temple of nature"*, London.
[14] DONDERS, F. C. (1870), *"De physiologie der Spraakklanken, in het bijzonder van die der nederlandsche taal"*, Utrecht.
[15] DU BOIS-REYMOND, E. (1862), *"Kadmus, oder allgemeine Alphabetik"*, Berlin.
[16] DUDLEY, H. & TARNOCZY, T. H. (1950), "The speaking machine of Wolfgang von Kempelen", *J. Acoust. Soc. Am. 22*, 151-166.
[17] FABRICIUS AB AQUAPENDENTE. (1600), *"De visione, voce, auditu"*, Venice.
[18] FANT, G. & RISBERG, A. (1963), "Auditory matching of vowels with two formant synthetic sounds", *Quart. Prog. & Status Rep., Speech Transmission Lab, Royal Inst. of Technol.* [Stockholm]. *STL-QPSR 4/1963*, 7-11.
[19] FERREIN, A. (1741), "De la formation de la voix de l'homme", *Mém. d'Acad. Roy. de Sci. 51*, 409-442.
[20] FISCHER-JØRGENSEN, E. (1967), "Perceptual dimensions of vowels", *To Honor Roman Jakobson I*. The Hague: Mouton. 667-671.
[21] FISCHER-JØRGENSEN, E. (1979), "A sketch of the history of phonetics in Denmark until the beginning of the 20th century", *Ann. Rep. Inst. of Phonetics, Univ. of Copenhagen 13*, 135-169.
[22] FLANAGAN, J. L. (1972), "Voices of men and machines", *J. Acoust. Soc. Am. 51*, 1375-1387.
[23] FOWLER, C. A. (1986), "An event approach to the study of speech perception from a direct realist perspective", *J. Phonetics 14*, 3-28.
[24] GOLDSTEIN, L. (1983), "Vowel shifts and articulatory-acoustic relations", In A. Cohen & M. P. R. v. d. Broecke (eds.), *Abstracts of the Tenth Int. Congr. Phonetic Sciences*. Dordrecht: Foris. 267-273.
[25] GRAMMONT, M. (1933), *"Traité de phonétique"*, Paris: Librairie Delagrave.
[26] GRASSMANN, H. (1854), "Leitfaden der Akustik", Programm des Stettiner Gymnasiums.
[27] GRASSMANN, H. (1863), "Ueber die Aspiraten und ihr gleichzeitiges Vorhandensein im An- und Auslaute der Wurzels", *Z. f. v. Sprachforschung auf dem Gebiete des Deutschen, Griechischen und Lateinischen 12*, 2.81-138.
[28] HALLER, A. VON. (1757-1766), *"Elementa physiologiae corporis humani"*, Lausanne.
[29] HELMHOLTZ, H. L. F. VON. (1863), *"Die Lehre von den Tonempfindungen"*, Braunschweig: Friedrich Vieweg u. Sohn.
[30] HOLDER, W. (1669), *"Elements of speech: An essay of inquiry into the natural production of letters"*, London.
[31] JACOBI, W. A. T. (1843), *"Beitrage zur deutschen Grammatik"*, Berlin: T. Trautwein.
[32] JAKOBSON, R., FANT, G., & HALLE, M. (1952), *"Preliminaries to speech analysis. The distinctive features and their correlates"*, (Acoustic Laboratory, MIT, Technical Report No. 13) Cambridge: Acoustic Laboratory, MIT.
[33] JESPERSEN, O. (1933), "Karl Verner", In *Linguistica. Selected papers in English, French, and German*. Copenhagen: Levin & Munksgaard. 12-23.
[34] VON KEMPELEN, W. (1791), *"Mechanismus der menschlichen Sprache"*, Wien. J. B Degen.
[35] KEY, T. H. (1855), "On vowel-assimilation, especially in relation to Professor Willis's experiment on vowel-sounds", *Transactions of the Philological Society* [London]. 5, 191-204.
[36] KOZHEVNIKOV, V. A. & CHISTOVICH, L. A. (1965), *"Speech: articulation and perception"*, JPRS (U.S. Dept. of Commerce) No. 30,543.

[37] KRATZENSTEIN, C. G. (1781), "*Tentamen resolvendi problema ab Acad. Scient. Imp. Petropolitana ad annum 1780 ... Praemio coronatum*", Petropol.
[38] KRATZENSTEIN, C. G. (1782), "Essai sur la naissance & la formation des voyelles", *J. de Physique 21*, Supplement, 358-380.
[39] LADEFOGED, P. (1980), "What are linguistic sounds made of?", *Language 56*, 485-502.
[40] LEHISTE, I. (1972), "The timing of utterances and linguistic boundaries", *J. Acoust. Soc. Am. 51*, 2018-2024.
[41] LINDBLOM, B. (1984), "Can the models of evolutionary biology be applied to phonetic problems?", In M. P. R. van den Broecke & A. Cohen (eds.), *Proceedings of the Tenth Int. Congr. Phonetic Sciences*. Dordrecht: Foris. 67-81.
[42] LINDBLOM, B. (1986), "Phonetic universals in vowel systems", In J. J. Ohala & J. J. Jaeger (eds.), *Experimental Phonology*. Orlando, FL: Academic Press. 13-44.
[43] LINDBLOM, B. (1989), "Phonetic invariance and the adaptive nature of speech", In B. A. G. Elsendoorn & H. Bouma (eds.), *Working models of human perception*. London: Academic Press. 139-173.
[44] LISKER, L. & ABRAMSON, A. (1970), "The voicing dimension: Some experiments in comparative phonetics", In B. Hala, M. Romportl, & P. Janota (eds.), *Proc. 6th Int. Congr. Phonetic Sciences*. Prague: Czechoslovak Acad. of Sci. 563-567.
[45] LISKER, L., COOPER, F. S., & LIBERMAN, A. M. (1962), "The uses of experiment in language description", *Word 18*, 82-106.
[46] MADDIESON, I. (1984), "*Patterns of sounds*", Cambridge: Cambridge University Press.
[47] MASSARO, D. (1987), "*Speech perception by ear and eye: A paradigm for psychological inquiry*", Hillsdale, NJ: Erlbaum.
[48] MENZERATH, P. & DE LACERDA, A. (1933), "*Koartikulation, Steuerung, und Lautabgreuzung: Eine experimentelle Untersuchung*", [Phonetische Studien. No. 1] Berlin & Bonn.
[49] MENZERATH, P. (1954), "*Die Architektonik des deutschen Wortschatzes*", [Phonetische Studien #3] Bonn.
[50] MILLER, J. (1989), "Auditory-perceptual interpretation of the vowel", *J. Acoust. Soc. Am. 85*, 2114-2134.
[51] MÜLLER, J. (1837), "*Handbuch der Physiologie des Menschen*", Coblenz.
[52] NEAREY, T. (1989), "Static, dynamic, and relational properties in vowel perception", *J. Acoust. Soc. Am. 85*, 2088-2113.
[53] NEAREY, T. & ASSMANN, P. (1986), "Modeling the role of inherent spectral change in vowel identification", *J. Acoust. Soc. Am. 80*, 1297-1308.
[54] OHALA, J. J. (1983), "The origin of sound patterns in vocal tract constraints", In: P. F. MacNeilage (ed.), *The production of speech*. New York: Springer-Verlag. 189-216.
[55] OHALA, J. J. (1989), "Sound change is drawn from a pool of synchronic variation", L. E. Breivik & E. H. Jahr (eds.), *Language Change: Contributions to the study of its causes*. Berlin: Mouton de Gruyter. 173-198.
[56] OHALA, J. J. (1990), "There is no interface between phonetics and phonology", *J. Phonetics 18*, 153-171.
[57] OHALA, J. J. (In press *a*), "The costs and benefits of phonological analysis", To appear in P. Downing, S. Lima, & M. Noonan (eds.) *Literacy and Linguistics*. Amsterdam: Benjamins.
[58] OHALA, J. J. (In press *b*), "The segment: Primitive or derived?", In D. R. Ladd and J. Docherty (eds.), *Proc., 2nd Laboratory Phonology Conference, Edinburgh*. Cambridge University Press.
[59] OHALA, J. J. (In press c), "What's cognitive, what's not, in sound change", In M. Morrissey & G. Kellermann (eds.), *Diachrony*

within synchrony. Peter Lang Verlag.
[60] O'NEILL, Y. V. (1980), *"Speech and speech disorders in Western thought before 1600"*, Westport, CN: Greenwood Press.
[61] PISONI, D. R. (1990), "Effects of talker variability on speech perception: Implications for current research and theory", Plenary talk, Int. Congr. on Spoken Language Processing, Kobe, 18-22 Nov 1990.
[62] KRUTA, V. (1975), "Purkyně", *Dictionary of Scientific Biography*, Vol. 11. New York: Charles Scribner's Sons. 213-217.
[63] RAPP, K. M. (1836), *"Versuch einer Physiologie der Sprache nebst historischer entwicklung der abendländischen Idiome nach physiologischen Grundsätzen.]* Stuttg. u. Tübg Cotta.
[64] VON RAUMER, R. (1863), *"Gesammelte sprachwissenschaftliche Schriften"*, Frankfurt: Heyder & Zimmer.
[65] ROSSI, M. (1983), "Niveaux de l'analyse phonétique: Nature et structuration des indices et des traits", *Speech Communication 2*, 91-106.
[66] ROUSSELOT, P. J. (1891), *"Les modifications phonétiques du langages"*, Paris: H. Welter.
[67] ROUSSELOT, P. J. (1903), *"Phonétique expérimentale et surdité"*, Paris.
[68] STETSON, R. H. (1928), *"Motor phonetics. A study of speech movements in action"*, The Hague.
[69] STETSON, R. H. (1946), "An experimentalist's view of Hidatsa phonology", *Int. J. of Am. Ling. 12*, 136-138.
[70] STEVENS, K. N. (1972), "The quantal nature of speech: evidence from articulatory-acoustic data", In: E. E. David, Jr. & P. B. Denes (eds.), *Human communication, a unified view*. New York. 51-66.
[71] STEVENS, K. N. (1989), "On the quantal nature of speech", *J. Phonetics. 17*, 3-45.
[72] STEVENS, K. N. & BLUMSTEIN, S. E. (1978), "Invariant cues for place of articulation in stop consonants", *J. Acous. Soc. Am. 64*, 1358-1368.
[73] STRANGE, W. (1989), "Dynamic specification of coarticulated vowels spoken in sentence context", *J. Acoust. Soc. Am. 85*, 2135-2153.
[74] TECHMER, F. (1880), *"Phonetik. Zur vergleichenden Physiologie der Stimme und Sprache"*, Leipzig: Wilhelm Englemann.
[75] TRUBETZKOY, N. (1933), "La phonologie actuelle", *Psychologie du Langage*. [*Journal de Psychologie*, No. 1-4, 1933], 227-246.
[76] TRUBETZKOY, N. (1969), *"Principles of phonology"*, Berkeley, CA: University of California Press. [Original: 1939. *Grundzüge der Phonologie*. Prag]
[77] VERNER, K. (1875), "Eine Ausnahme der ersten Lautverschiebung", *Z. f. vergleichende Sprachforschung. 23*, 97-130.
[78] VERNER, K. (1913), [Letters to Hugo Pipping]. *Oversigt over det kongelige danske videnskabernes selskabs forhandlinger*. No. 3, 161-211. København.
[79] WALLIS, J. (1653), *"Grammatica linguae Anglicanae"*, [With appendix: *De loquela, sive sonorum formatione, tractus grammatico-physicus*.] Oxford: Leon Lichfield.
[80] WILLIS, R. (1821), *"An attempt to analyze the automaton chess player, of Mr. de Kempelen"*, London: Booth.
[81] WILLIS, R. (1830), "On the vowel sounds, and on reed organpipes', *Transactions of the Cambridge Philosophical Soc. 3*, 229-268.
[82] ZIPF, G. K. (1935), *"The psycho-biology of language"*, Boston. Houghton Mifflin Co.
[83] ZWAARDEMAKER, H. (1909), *"La phonétique expérimentale considerée au point de vue médical"*, Paris: Libraire J.-B. Bailliere et Fils.
[84] ZWAARDEMAKER, H. & EIJKMAN, L. P. H. (1928), *"Leerboek der phonetiek. Inzonderheid met betrekking tot het Standaard-Nederlandsch"*, Haarlem: DE Erven F. Bohn.
[85] ZWIRNER, E. & ZWIRNER, K. (1936), *"Grundfragen der Phonometrie"*, Basel.

FROM SPEAKING MACHINES TO SPEECH SYNTHESIS

Jean-Sylvain Liénard

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur

LIMSI - CNRS
BP 133 - 91403 ORSAY Cedex

ABSTRACT

Speech synthesis has a long history. The first scientific attempts date back to the latter part of the eighteenth century, with Kratzenstein's resonators, Mical's talking heads, and Kempelen's speaking machine. The functioning and results of these devices are presented, as well as those of two other famous machines, Faber's Euphonia and, more recently, Dudley's Voder. The second part of the paper introduces some elements of discussion concerning the magical aspect of speech, the modelling of the human speech production system, and the very concept of a speaking machine.

1 - INTRODUCTION

Kempelen was the eighteenth century inventor of the first speaking machine, the good functioning of which has been testified by numerous contemporaries. This achievement was not due to chance; it was part of the framework of a movement of interest in the topic of speech production, which itself was related to the wide curiosity of this time for the understanding and imitation of natural phenomena. Among the other attempts which appeared afterwards, the most accomplished was undoubtably Faber's machine, unfortunately ignored, probably because its author, unlike Kempelen, had not published a book describing it in detail. The last speaking machine which really deserved this name was that of Dudley, the Voder, which was also famous in its time, but which carried in itself the end of an era and the beginning of a new one. Today the term "speaking machine" and the dreams it denotes are abandoned to the benefit of that, seemingly more technical or rigourous, of "speech synthesis".

We shall try, in the following paper, to draw the main lines of this evolution, to show its ruptures and survivals and, perhaps, to extract from it some lessons for the future.

2 - HISTORICAL SURVEY

Kempelen's idea was not new. It corresponded to an old dream of humanity, which had been formulated as early as antiquity and Middle Ages. It was felt then that to give speech to inanimate objects was to give thought, a divine gift, specific to the human species. Making objects speak had a magical, if not sacrilegious, aspect. Even today, everything related to the mechanisation of speech retains some kind of mystery; the domain is intrinsically spectacular, and the researcher often has to restrain the untimely enthusiasm of his contemporaries.

2.1 - Myths and mystifications [1, 11, 14]

Several antique authors describe the talking statue of Memnon, the son of dawn in the mythology of ancient Egypt. This statue was said to emit intelligible speeches, and even seven-verse long oracles ! This extraordinary gift was due to the intervention of the egyptian priests; the statue emitted only one sound, evoking the breaking of a string. The phenomenon happened at sunrise; it may have been caused by the dilatation of some parts of the monument because, in the country in question, the temperature deviations between night and day may be very great.

Throughout history, there are many testimonies related to talking heads. At Lesbos a speaking head attributed to Orpheus was famous for having predicted the violent death of Cyrus the Great, during his expedition against the Scythians. Odin, the nordic magician, had a talking head which passed as wise Minos' head, and which gave divine answers...The mechanician Gerbert of Aurillac, who became pope at the turn of the first millenary under the name of Sylvester the Second, was supposed to have built a talking head of brass, which said the words "yes" and "no". Monk Albert, who became for posterity Albert the Great (thirteenth century), was reported to have build a head of baked clay, which spoke and moved. This masterpiece had such a sacrilegious character that Thomas Aquinas broke it into pieces...

It is useless to lengthen the list of examples. Most of those heads are a matter for legends. If however some of them seem to have existed, they probably worked with concealed pipes, or through the ventriloquist talents of their authors.

2.2 - The beginning of a scientific approach

Since the seventeenth century it has been possible to observe the maturation of ideas concerning the mechanism of speech production. The preoccupations then were of a philosophical and anatomical order. An alchemist, Van Helmont, imagined in 1668 a theory according to which the letter shapes of the Hebrew alphabet would represent the positions of the tongue in the mouth...The same year, a less fanciful study was performed by Wilkins, who defined for each speech sound a corresponding arrangement of the vocal organs. Debates on the nature of the voice producing organ arrived in the eighteenth century at the notion of vocal cords, due to Ferrein. Finally, in 1779, the Academy of Sciences of St Petersburg proposes as the topic of its annual contest the following questions:

a) What are the nature and character of the sounds of the vowels A E I O U (probably /a e i o u/), so different one from another ?

b) Can an instrument be constructed like the *vox humana* pipes of the organ, which shall accurately express the sounds of the vowels ?

At that time three researchers - Kratzenstein, Mical and Kempelen - had already obtained some results. But they did not know each other and, apparently, only Kratzenstein presented a realization at the Academy. He was the one who won the prize.

Christian Gottlieb Kratzenstein was a professor of physiology at Copenhagen. He presented five resonators, of which the internal volumes and openings corresponded to those of the mouth during the emission of the requested vowels. They were adjusted on a windchest and modified the timbre furnished by a free reed, except for the resonator producing the sound I, which received the airflow directly.

In 1778 Abbot Mical, a Frenchman who had a passion for mechanics, had built a talking head which could articulate a long sentence. Forced into the public eye by a friend's indiscretion, he destroyed his machine. But this inventor, modest as well as quick-tempered, started again and at the French Academy of Sciences in 1783, he presented a machine made of two talking heads, which pronounced the two following sentences in the form of a dialog:

- first head: "Le Roi a donné la Paix à l'Europe" *("The King has given Peace to Europe")*,

- second head: "La Paix couronne le Roi de Gloire") *("Peace crowns the King with Glory")*.

The committee put in charge of examining the machine was composed of distinguished scholars [3]. It worked out a long report in which, unfortunately, was no diagram. Here is a part of it (translated from [3]):

"... These two sentences are not clearly pronounced in all of their parts, especially the last one. This is mainly due to the fact that the basis of the voice produced by this machine greatly differs from a human voice; that, since some syllables result from the combination of several sounds, their joining does not occur with all the possible precision; and also that the pronunciation of several consonants needs to be perfected... One can consider it as made of two different parts:

a) A wind chamber, in which a bellows brings the air and from which this fluid escapes when different valves are raised. The air is then directed by ducts towards the cavities, where it is modified, and where it becomes sonorous.

b) A cylinder which moves levers, and which gives them the necessary impetus, either to raise the valves at the appropriate moment or to give the cavities where the sound modifies itself the shapes required by its diverse modifications."

From the description given in the report it follows that each of the resonators is fitted with its own reed, which probably explains some of the difficulties encountered in the stringing together of successive speech sounds.

Despite the flattering terms of the prerecorded sentences, the King did not express interest in Mical's machine. The inventor became impoverished and died in 1789.

2.3 - Kempelen's machine [5, 10, 11, 12, 13, 14]

Wolfgang von Kempelen, a nobleman living at the court of Austria-Hungary, was born in 1734 in Presburg (now Bratislava) and died in Vienna in 1804. He was an organiser and inventor of great talent. He was the designer of the fountains of Schoenbrunn palace, as well as of the plans of the royal castle of Buda. He was the organiser of a wool factory in south Hungary. But above all he was the author of two renowned machines: the chessplaying automaton, which was immortalized by Edgar Allen Poe in a famous novel, and the Speaking Machine. Kempelen never claimed the chessplayer to be a real automaton, but the trickery was so perfect that nobody was ever able to fault it. As for the speaking machine, it resulted from two decades of scientific investigations.

In his book [13], Kempelen presents at length his theory of speech production. Only the last chapter is devoted to the machine: in his mind the two are intimately related. His first trials date back to 1769. With a variable volume resonator and a bagpipe reed he succeeds in imitating the sound of some vowels. Then he makes several resonators producing the vowels /a/, /o/, /u/, as well as others in two articulated parts, producing the consonants /p/, /l/, /m/. These elements, fixed on an ogan windchest and put into action by a set of keys, constitute his first machine which is presently at the Vienna museum. But it is a failure. The sounds do not connect with each other, and the emission of the vowels is preceded by a sort of explosion which does not resemble a speech sound. Thus he gives up the result of two years of work, which probably exhibit the same kind of defects as those observed in Mical's machine, and resolutely adopts an anthropomorphic design: since Nature has provided us with a single glottis and a single mouth, it must be the same in a speaking machine.

The final machine, as described in the book, is composed of a bellows, a free reed, a windbox, a rubber open resonator and two openings, which play the part of the lungs, the vocal cords, the pharynx, the mouth and the nostrils (Fig 1). The operator's right elbow rests on the bellows and produces the air pressure. The right hand is busy with the different levers and openings on the top of the windbox, while the left hand more or less closes

up the "mouth", the whole constituting an adjustable resonator. The reed length is fixed using a piece of wire, in order to produce a high pitched voice, attributable to a child.

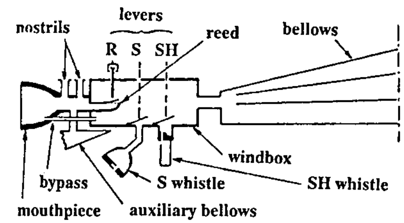


Fig 1 - functional diagram of Kempelen's machine

In order to pronounce /m/ and /n/ the mouth is closed up by the left hand and the sound escapes through one nostril (for /m/) or both (for /n/). This arrangement permits a smooth connexion between those consonants and the following sounds.

The phoneme /r/ is produced by creating some irregularities in the functioning of the reed: by depressing a key a brass needle is brought into contact with the reed during its vibration. The needle rebounds against a wooden stopper, which limits its course. This produces a scraping noise, the duration and intensity of which are adjusted by the operator according to the depression of the key.

The phoneme /l/ is produced by a quick movement of the left hand; two fingers are introduced into the mouth until the reed canal is partly closed up, so as to divide the airflow in two for a short time.

The phoneme /p/, and the other unvoiced plosives, are produced by closing up the mouth and nostrils, then by rapidly removing the left hand, which goes into the position necessary for the next vowel. In order to prevent any vibration of the reed during the occlusion Kempelen found it necessary to balance the pressures using a narrow pipe, which acts as a bypass of the reed canal. In addition, a small bellows located under the glottis increases the efficiency of the compression and contributes to a better restitution of the burst. The voiced plosives are produced in the same manner, with an extra airflow through one slightly uncovered nostril, so as to allow the reed to continue its vibration during the occlusion.

The fricatives /s/ and /ʃ/ are produced in accessory whistles located on both sides of the windbox, put into action by two valves controlled by levers. The phoneme /f/ is produced very simply by the unavoidable airflow losses which subsist when the mouth is closed up and the air pressure increased. The aspirated /h/ is obtained with the mouth open and a pressure low enough not to make the reed vibrate.

At the time of the demonstrations, the machine was covered with a small wooden box with two openings through which the operator could pass his hands. The top of the box was made of fabric. According to Kempelen the purpose of the box was to protect the machine from dust as well as to provide a passage for the sound. It might also have had the purpose of surrounding the machine with mystery, as was usual at that time, unless it was intended to prevent any imitator from copying its mechanism. The inventor describes his results in the following way (translated from [13]):

"... Although imperfect, it at least gives some good principles for designing a more perfect one. Finally I have brought it to the point where I can make it pronounce at the first trial and without any exception all of the Latin, French and Italian words that are proposed to me, some, it is true, better than the others, but at least several hundred words clearly and distinctly. For instance: Papa, Maman, Marianna, Roma, Maladie, Santé, Astronomie, Anatomie, Chapeau, Racine, Soupé, Charmante, Opéra, Comédie, Pantomime, as well as long and difficult words such as Constantinopolis, Monomotapa, Missisipi, Astrakan, Anastasius, etc... As for complete sentences, I can only pro-

nounce a few of them, for instance: Vous êtes mon ami - Je vous aime de tout mon coeur - Leopoldus Secundus - Romanorum Imperator, etc...".

These results were confirmed by numerous contemporaries. Grimm, the writer, who saw the machine in 1783, testified as follows (translated):

"... As it is today, the machine already clearly answers several questions; its voice is pleasant and soft; only the Rs are pronounced in a guttural way, with a tedious snoring noise. When one has not understood its answer, it repeats it again, but with a tone of infantile irritability and impatience... The pronunciation of Mr Abbot Mical's machine is far from being as distinct as that of Mr Kempelen's machine...".

Kempelen's machine was imitated several times. The copy which is exposed at the Deutsches Museum of Munich does not fit exactly with the description given in the book. It has two extra levers on the top of the windbox, one of which seems to control the length of the reed, that is the pitch of the voice. This kind of improvement might have been worked out by Kempelen himself after the publishing of the book. Another reconstitution was attempted by the physicist Sir Charles Wheatstone, some sixty years later [5, 14]. We ourselves made a reconstitution in order to check some points [10, 11]. In particular we could verify that the vowels were restituted only as crude approximations, except for /a/ and /o/, the device permitting only the creation and variation of a single wide formant in the 1000-2000 Hz region. Similar observations were made by Van den Broecke, who made a replica around the same time [12]. The consonants require some manual skills, which necessitate a long training period.

2.4 - Faber's machine [2, 6, 8, 11]

Kempelen claimed to have made his machine *"for the benefit of some master's hand, who would know how to raise it to the highest degree of perfection".* Actually his machine was missing a tongue and teeth, and he had envisioned the great improvements that variable pitch and keyboard control would have brought.

His real successor was Joseph Faber, a professor of mathematics born in Vienna around 1786, who could possibly have known Kempelen directly. He probably read Kempelen's book, because his machine comprised many similarities, as well as some of the improvements mentioned above. This machine, called "Euphonia", finished in 1835, was presented in several European main cities over a period of twenty years by Faber himself, then sold to Barnum, the famous show director. It appeared again in Paris around 1880, and probably lies today as a wreck in the basement of the Paris School of Medicine.

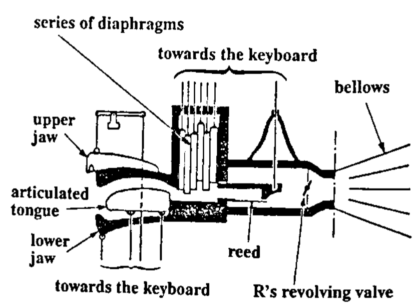


Fig 2 - Faber's machine diagram, after Du Moncel

This machine spoke, with normal or whispered voice, and sang "God save the Queen" ! It comprised a foot-manipulated bellows, a tongue, articulated jaws made of a flexible material, and a set of six diaphragms which modified for each sound the shape and section of the vocal tract (Fig 2). The controls were grouped on a 14-key keyboard; pressing down each in turn, one could obtain the following sounds : /a/, /o/, /u/, /i/, /e/, /l/, /r/, /v/, /f/, /s/, /ʃ/, /b/, /d/, /g/. Each key simultaneously controlled several parts, in fixed proportions. In other

words, the control of Euphonia was practically phonemic in nature, even though some controls remained physiological or acoustical (nasality, voicing). The phoneme /r/, produced by a modulation of the airflow, was probably more plausible or pleasant than Kempelen's. The reed was articulated in such a way that the vibrating length could vary so as to control intonation.

One cannot help being impressed by the accuracy and modernity of the design of this machine. The part corresponding to the vocal tract was about 15 centimeters long. The rear portion of the tract was defined by a set of six sections of variable area and shape. The front part was defined by three parameters controlling the positions of the jaws and tongue, which evoke the now classical parameters of openness, aperture and place of articulation. Considering such a richness it is probable that this machine had the capability of working out realistic formants and transitions from one phoneme to the next.

Unfortunately nothing remains of Euphonia, except some descriptions which do not come from Faber himself and as such are necessarily superficial. It is certain that Euphonia was much more sophisticated than Kempelen's machine.

2.5 - Dudley: the rupture [4, 7]

A century passed before the reapparition of a speaking machine, which marked simultaneously a change of technology and a change of design. In 1937 Homer Dudley and his colleagues Riesz and Watkins, engineers with the Bell Telephone Company, finished out the VODER (VOice DEmonstratoR), which was exhibited to a large public in 1939, at the San Francisco exhibition and at the New York World's Fair. Externally the Voder looks somewhat like Faber's machine (an operator playing speech on a keyboard), but it differs from it in two respects. The first one deals with the physical nature of the vibrating phenomenon: sound is processed through its electrical analog, of which telephony has shown the equivalence as far as transmission is concerned. This analogy remains widely used nowadays, even though the signals are processed digitally. The other aspect concerns the parameters controlled by the operator, which are strictly related to the pitch and spectral envelope, without any reference to the vocal tract physiology or to phonetics.

The Voder is well enough known (Fig 3) for us not to spend too much time on its description. Let us just mention that the operator uses keys to control the signal amplitudes in ten spectral bands. Four extra keys and a pedal are used to control the occlusions and bursts of the plosives, the voiced/unvoiced feature and the pitch.

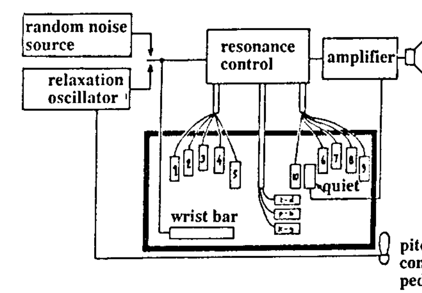


Fig 3 - Principe of the Voder, after Dudley

The Voder was so difficult to operate that the 24 telephone operators selected among 320 to demonstrate the machines had to receive a year of intensive training. The result was conclusive, however, and the operators could effectively play intelligible sentences on their machines, and even make them sing.

The Voder cannot be separated from the VOCODER (VOice CODER) presented by the same authors at the same time. In the Vocoder the control signals were automatically extracted from the analysis of real speech by means of a filterbank completed by a voicing and pitch analyser. The Vocoder is not a speaking machine, but a speech compression system. It is motivated by an economical stake, which will be the main driving force for speech research for forty years. Even nowadays the analysis/synthesis paradigm it illustrates remains prominent, despite a change of technology (from analog to digital) and some new methods of signal processing.

2.6 - Speech Synthesis today

We shall not present here an inventory which can be found elsewhere [7, 9], but only observe that the term of Speech Synthesis has substituted that of Speaking Machines. Speech Synthesis uses different types of knowledge according to the nature of the control parameters. At the lowest level these parameters are acoustic (they come or could come from a signal analysis very similar to the one implemented in the Vocoder) or articulatory (they represent the variables of a simulation of the vocal tract).

Controlling the synthesis process in terms of phonetic and prosodic parameters implies another step, in which the knowledge used, either explicitly (rules) or implicitly (segments), is related to the dynamic functioning of the vocal apparatus, partly guided by phonetic considerations. Let us recall here the large body of work that has been devoted to this aspect since the early fifties, in which Dennis Klatt took a major part [9].

Text-To-Speech Synthesis still requires a third step, mainly of linguistic nature (but not exclusively), which governs the interpretation of a text into oral terms. Finally, one should mention the Concept-To-Speech Synthesis, which could reveal itself to be extremely rich within the next few years in the context of Man-Machine Communication, but on which very little work has been done as yet.

Let us just observe that, after a twenty-year period during which the speech synthesis problem was wrongly considered as practically solved - a side effect of the Vocoder paradigm -, a powerful renewal of interest has now appeared, at all processing levels. This is due to many reasons that exceed the scope of the present paper. In the discussion which follows we shall get back to our initial topic concerning the Speaking Machines.

3 - WHAT COMES OUT OF THE SPEAKING MACHINES SAGA

We shall now try to distance ourselves from this historic evolution, in order to emphasize some of its aspects in relation to the contemporary views on speech processing.

3.1 - The magical aspect of speech

Throughout the early history of speaking machines the divine or magical aspect of speech was prominent. Even the eighteenth century scientific efforts were not definitely cleared of any mystification. Is this aspect really absent from the contemporary speech research ?

The fact of having a machine pronounce only words or sentences known in advance may look like a kind of mystification. However it reveals something which is basic in human understanding. In everyday life it is rare for a message to be totally unpredictable. Even when it is poorly articulated a message can be understood if it is partially predictable in the situation context. Conversely, when a speaker knows that his interlocutor can predict some of the message, he does not have to take the care of a perfect articulation. Kempelen had understood that point and made some use of it: *"... One is particularly misled when one knows in advance the word that the machine has to say, and when it pronounces it one imagines to have heard it..."*. Let us mention that this effect is extremely disturbing when one works out rules or patterns in speech synthesis, and that it necessitates the use objective listening tests, free from any uncontrolled previsibility, be it phonological, lexical, semantic or pragmatic. On the other hand, it indicates that speech synthesis could, in some cases, be thought as a predicting and interacting game between the machine accomplishing a task and the operator who supervises it.

In the same vein, Kempelen knew very well that some phonemes were not correctly pronounced by his machine (he used to replace /d/, /g/, /k/, /t/ by /p/ or /b/), and it was in full awareness that he gave it a child's voice: *"... the childish voice of the machine is always advantageous to it. One willingly forgives a child who sometimes stammers the mistake of using one letter in place of another, and one satisfies oneself of having understood what he meant..."*.

Hiding the active part of the machine in a box, as well as using some of the tricks mentioned above, brings conjuring tricks in mind. But one has to remember that at that time, today's scientific criteria were not strictly defined. Curiosity, ingenuousness, the capacity for amazement, were as essentiel to progress as the scientific method in the rigorous sense. The very idea transmitting speech at a distance or recording it seemed to be a dream (the poetic notion of "frozen speeches" had been formulated by François Rabelais in 1548).

Even today, in the latter part of the twentieth century, it is not certain that our research activities in the field of speech are perceived as being totally free of something magical. People are always surprised to hear a machine speak. When they realize that it is genuine, they tend to ascribe to it intelligence, language, and feelings like a human's; while it only superficially reproduces some of man's linguistic abilities. And it is highly significant that the potential users of TTS synthesis nowadays expect a more "natural" voice, although this is not absolutely necessary in most practical situations. Also, what makes speech such a fascinating domain of investigation to us, the rigorous speech scientists of 1990 ?

3.2 - Is it necessary to imitate human speech production ?

A great debate, opened in the earliest times of the speaking machines, is still going on today. On one side there is the idea that imitating nature as best as we can must improve speech synthesis. The degree of imitation is evidently a function of current knowledge and techniques. In the Middle Ages it was thought that it sufficed to materially imitate a human head for it to spontaneously produce speech, and if it did not one added to it some artifice... Faber illustrates best the success of this anthropomorphic view, which manifests itself today in the articulatory models. On the other side are the functional approaches, according to which it is not the conformity of the model to the original that counts the most, but the very result, the function, obtained by using different materials and techniques. The Voder is a perfect illustration of this view, in several respects: direct modification of the spectrum, ignoring the vocal tract functioning, and use of electronics to simulate acoustical phenomena.

We have no intention here of choosing between these views, each of which has its own merits and limitations. Obviously it is impossible to strictly imitate Nature in all respects; even if one succeeded in reconstituting a system presenting all the physical and physiological properties of the human apparatus, one would only have pushed the problem a little farther, because one would have then to build the equivalent of its nervous control, as well as the proprioceptive and auditory organs which allow it to learn and function. Conversely, any purely functional approach quickly encounters some limitations due to a lack of knowledge concerning the real vocal apparatus, which constitutes one of the possible realizations of the function that is investigated. It must be observed that, generally, the two views complement rather than oppose each other. In this spirit, Kempelen, after the failure of his first machine, succeeded by imitating the human speech production more closely. In a reciprocal spirit some notions which come from the physiology of speech production, such as the notion of a formant and the vocal source/vocal tract duality, are a great help in the functional systems as first approximations or sources of knowledge to be used in another form.

Let us observe that such a debate is extremely general and could concern speech analysis, visual perception, pattern recognition, artificial intelligence as well.

3.3 - What sense does it make to play a speaking machine ?

The notion of a speaking machine as a "speech instrument", in the sense of a "musical instrument" was abandoned after the Voder, to the benefit of systems using delayed controls. Interesting questions can be asked about the causes of this disappearance.

Firstly, the only use of a "speech instrument" is the demonstration that it is possible to play it. Mute people can be divided into two groups. For the first, the difficulty in speaking is due to an auditory deficiency; replacing their vocal apparatus by a manual device does not change anything, inasmuch as they cannot learn how to control it. For the second the problem comes from the poor functioning of the vocal cords; an artificial larynx is sufficient in that case.

Secondly all the realizations in the past have shown that it is extremely difficult to learn how to play such an instrument, which moreover delivers a result greatly inferior in quality to the normal production of any human being.

Does this mean that the notion of "speech instrument" is of no interest today ? Maybe not, because the interest of spontaneity, of real-time interaction, of the individual and expressive aspects of the voice have been forgotten a little too quickly. For a speaking machine to raise some practical interest its control should be simple to learn (i.e. it should be phonemic or syllabic in nature), and some expressive capabilities should be available. It should allow for the generation of several voice timbres, as well as of non-speech sounds, for musical or sound engineering purposes.

4 - CONCLUSION

Several speaking machines have been built in the past. For some of them it is beyond doubt that they worked satisfactorily for the listeners of their time. Their inventors had understood in an empirical way some of the structures of speech. Kempelen had understood the importance of the proper linking of successive phonemes, as well as the principle of the separation between the glottis and the vocal tract, not to mention some of the perceptual phenomena related to speech communication. Faber added to that a much more sophisticated modelling of the vocal tract, as well as the successful realization of a phonemic control using a keyboard. Dudley demonstrated with the Voder the possibility of reconstructing speech using electrical signals, without any reference to the physiology of the human apparatus, and without using a phonemic control.

The synthesizers which came later made it possible to experiment with pre-recorded controls; the concept was there in Mical's talking heads, but the technology and knowledge were too primitive to permit any serious investigation at that time. This evolution is now oriented towards Text-To-Speech synthesis, under the pressure of computer technology. During the progress in this direction some aspects of speech communication have been neglected. Spontaneity, expressivity, flexibility, interruptibility, underlying intelligence and sensitivity are expected by potential users of synthetic speech, because they want it to imitate human speech at a deep level.

Making a computer speak is not the same as playing a speaking

machine. Speaking machines, in the sense of "speech instruments", have disappeared, for they lacked the capability of being easily played and of being useful for something. With them disappeared the idea of a direct and instantaneous control by a human operator, as well as the idea of feedback from the listener to the operator, since everything has been frozen into the algorithms. Another aspect which has also disappeared, due to the success of the electrical analogies, is care for the real physical phenomenon of speech, which refers to fluid mechanics: what synthesizer, what articulatory model, takes into account the physical phenomena created by the airflow in the vocal tract, the role of the saliva, the directivity of the sound in the three dimensions of the space ? On those points as well as on a few others it may prove fruitful to adopt once more the naive attitude of the pioneers.

5 - REFERENCES

1 - BREWSTER D. : "Letters on Natural Magic", London, 1832.

2 - CHAPUIS A. et GELIS E. : "Le monde des automates", Paris 1928.

3 - DE MILLY, LE ROY, LAVOISIER, LAPLACE, FERRIEU, VICQ D'AZIR, Compte-rendu de l'Académie des Sciences, 3 septembre 1783.

4 - DUDLEY H., RIESZ R.R., WATKINS S.S.A. : "A Synthetic Speaker", J. Franklin Inst., **227**, 739-764, 1939.

5 - DUDLEY H. and TARNOCSY T. "The Speaking Machine of Wolfgang von Kempelen", J. Acoust. Soc. Amer. **22**, 151-166, 1950.

6 - DU MONCEL T.: "Le Téléphone", Paris, 1887.

7 - FLANAGAN J.L. : "Voices of Men and Machines", J. Acoust. Soc. Amer. **51**, 1375-1387, 1972.

8 - GARIEL : "La machine parlante de Mr Faber", Journal de Physique Théorique et Appliquée, T 8, 1879.

9 - KLATT D.H. : "Review of Text-To-Speech Conversion for English", J. Acoust. Soc. Amer. **82**, 737-793, 1987.

10 - LIENARD J.S. : "Reconstitution de la machine parlante de Kempelen", 4° Conférence Internationale d'Acoustique, Budapest, 1967.

11 - LIENARD J.S. : "La machine parlante de Kempelen", bulletin n°34 du Groupe d'Acoustique Musicale, Université Paris VI, mars 1968.

12 - VAN DEN BROECKE M. : "Wolfgang von Kempelen's Speaking Machine as a Performer", in Sound Structures, M. Van den Broecke ed., Foris publications, Dordrecht, 1983.

13 - VON KEMPELEN W. : "Le Mécanisme de la Parole, suivi d'une Description de la Machine Parlante", Vienne, 1791.

14 - WHEATSTONE C. : "The Scientific Papers", London, 1879.

THE CONTRIBUTION OF SPEECH SYNTHESIS TO PHONETICS: DENNIS KLATT'S LEGACY

Kenneth N. Stevens

Research Laboratory of Electronics and Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A.

ABSTRACT

Some of Dennis Klatt's contributions to the science and application of speech synthesis are described, and the effects of these contributions on the study of phonetics are discussed. The synthesizer developed by Klatt, with an extended set of control parameters, can be manipulated to simulate different female and male voices and can produce a variety of classes of speech sounds in context, based on principles of human speech sound generation. The problem of controlling the multiple parameters of the synthesizer is considered, in view of the constraints imposed on the parameters by the articulatory and aerodynamic processes in speech.

1. INTRODUCTION

One of Dennis Klatt's contributions to the field of phonetics was to advance the science and application of speech synthesis. He approached the problem of speech synthesis in a systematic way, incorporating and contributing to what is known about the speech production process and collecting empirical data for situations where theoretical models were inadequate. I shall try to summarize the major contributions he made to this field, the relevance of these contributions to the study of phonetics, and some new directions in speech synthesis and indeed in phonetic theory that have been made possible because of the groundwork established by Dennis Klatt.

Some of his innovations in speech synthesis are concerned with the organization of the synthesizer itself, and others involve his development of rules for controlling the synthesizer. We shall examine first the arrangement of components in the most recent version of the Klatt synthesizer, called KLSYN88.

A block diagram of the synthesizer is shown in Fig. 1. There are four main components of this synthesizer: (1) a source simulating the glottal output, (2) a source of frication noise, (3) a transfer function for the glottal source, and (4) a transfer function for the frication source. The arrows pointing to the upper and lower sides of the boxes indicate that certain parameters of the sources and filters can be controlled by the user. Dennis Klatt has contributed significantly to several of these components and their interconnections.

2. LINKING FORMANTS FOR GLOTTAL AND FRICATION SOURCES

One of the problems for the synthesis of speech with a formant or terminal-analog synthesizer is that the nature of the transfer function from the source to the sound output at the mouth or nose is different when the source is at the glottis than when it is a transient or frication source located at one or more points along the length of the vocal tract. When the source is at the glottis, the transfer function is an all-pole function in the case of nonnasal vowels, and there may be additional poles and zeros for nasal vowels and for nasal and liquid consonants. In the case of a frication source, only certain of the natural frequencies of the vocal tract are excited, and it is possible to describe the transfer function as being characterized by free poles, free zeros, and pole-zero pairs [1].

Dennis Klatt observed that, for both types of sources, the poles are the same, being the natural frequencies of the vocal tract independent of the source location. For the frication source, these natural frequencies are excited with very different strengths, with primary excitation of the cavity in front of the constriction and of a possible palatal channel behind the constriction. He designed a synthesizer configuration which had two separate paths, as shown in Fig. 1 -- a cascade arrangement of poles and zeros for the glottal source and a parallel arrangement of resonators, with associated adjustable gains, for the frication source. The frequencies of the parallel resonators and of the poles in the cascade path are linked together, as the figure indicates, thereby incorporating the constraint that the natural frequencies change continuously independent of the source location within the vocal tract [7].

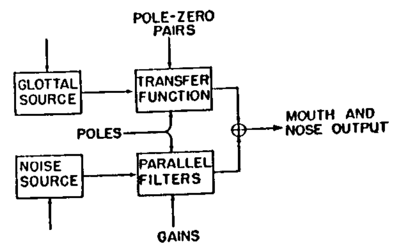


Fig. 1 Block diagram of the main components of a terminal-analog speech synthesizer such as KLSYN88 [9]. The vertical arrows on the sides of the boxes indicate arrays of control parameters.

3. SYNTHESIS OF SOUNDS WITH FRICATION NOISE

The part of the synthesizer that generates sounds with a frication source is shown in the form of a block diagram in Fig. 2. The spectrum of a sound produced with frication noise is shaped by adjusting the gains for each of the parallel formant resonators (A2F, A3F, etc.), together with a gain for a bypass path (AB) for which there is no filtering of the noise source. In order to generate noise bursts and fricative consonants, usually only a subset of the gains are active, since there is appreciable excitation of only some of the natural frequencies of the vocal tract.

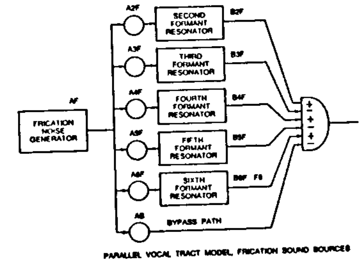


Fig. 2 Organization of the components of the KLSYN88 synthesizer for producing sounds generated with frication noise.

One of the projects that Klatt was working on in 1987 and 1988 was to improve the synthesis of fricative consonants. One component of this project was to synthesize frication noise with the proper spectrum. The task involved measuring the spectra of different fricative consonants in different vowel environments, produced by a male and a female speaker. In order to synthesize the fricative consonants, it was necessary to adjust the various gains for the formant filters so that the synthesized spectrum matched the original spectrum. The match could be achieved quite accurately, as Fig. 3 shows. This figure displays the spectra of two different fricative sounds produced by a male speaker, together with spectra synthesized by proper selection of the array of gains in Fig. 2. The upper two panels with the naturally produced fricatives also show smoothed spectra for a vowel adjacent to the fricative, to indicate the relative spectrum amplitudes of vowel and fricative. Comparison of the fricative spectra in the upper two panels and the synthesized spectra in the lower two panels shows reasonable agreement. In the case of the labiodental fricative, the gain AB of the bypass path was adjusted, with all other gains set to zero, whereas for the alveopalatal fricative, the gains A3F and A4F contributed the salient attributes to the spectrum. These observations are consistent with theories of fricative production [1].

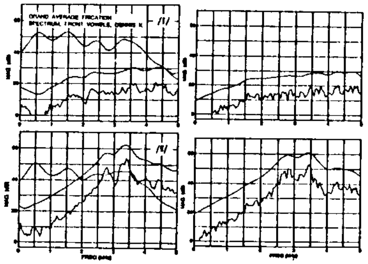


Fig. 3 The left two panels show measured spectra from syllables consisting of voiceless fricatives /f/ and /ʃ/ followed by front vowels. The fricative spectra are averages over a number of utterances, and a smoothed spectrum is displayed above each measured spectrum. The superimposed vowel spectrum is the smoothed spectrum of an adjacent vowel /ɛ/. The right two panels show attempts at matching the two left fricative spectra by setting the gains in the system in Fig. 2.

For some fricatives, particularly /s/ and /ʃ/, there are some differences in the spectrum depending on the vowel environment. In a rule system for speech synthesis in English, Klatt used somewhat different arrays of gains for fricative synthesis depending on whether the adjacent vowel was in one of three classes: front vowels, unrounded back vowels, and rounded back vowels.

4. SYNTHESIS OF NASAL CONSONANTS AND VOWELS

Synthesis of oral vowels is achieved in the usual way with an all-pole system consisting of a cascade of resonators. For the synthesis of nasal vowels and consonants, Klatt introduced a pole-zero pair into the cascade path, in addition to the poles normally involved in the generation of nonnasal vowels. The arrangement of poles and pole-zero pairs in this cascade path is shown in Fig. 4. The figure shows that the bandwidths of the formants as well as their frequencies are controllable — an essential feature if a variety of voices are to be synthesized or if natural vowels and transitions are to be generated.

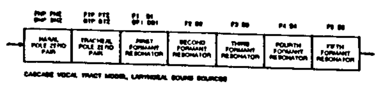


Fig. 4 Organization of the components of the KLSYN88 synthesizer for producing sounds with a source at the glottis.

Theoretical and experimental work has suggested that a single pole-zero pair in the low- to midfrequency range is sufficient to provide natural-sounding nasal murmurs and nasal vowels with spectra that match the spectra of natural utterances [1], [3], [4], [5]. The frequencies and bandwidths of the nasal pole and zero are designated in the figure as FNP, FNZ, BNP, and BNZ. In the case of a nasal murmur, the sound output is from the nose, and the zero in the transfer function arises because of the side branch formed by the mouth cavity. For a nasal vowel, the output is from both the mouth opening and the nose, and the zero for an adult speaker is usually in the range 500 to 1500 Hz, depending on the vowel and the size of the velopharyngeal opening. When a nasal consonant is released into a vowel, the natural frequencies of the combined vocal- and nasal-tract configuration change continuously, although some of the frequencies (such as the lowest one) might change quite rapidly. The zero also changes continuously, but its motion can be extremely rapid, and may traverse 1000 Hz or more in a few milliseconds.

An example showing the trajectories of the poles and the zero for synthesis of a nasal consonant in intervocalic position is given in Fig. 5. In the first part of the initial vowel and after 20-odd ms of the final vowel, there is no nasalization, and the extra pole and zero cancel each other. Creation of a velopharyngeal opening causes a separation of the pole and the zero. The velopharyngeal movement usually begins well in advance of the consonant closure, as shown by the separation of the pole and zero in the first vowel. As the constriction is formed and the output shifts from the mouth to the nose, there is a rather abrupt change in the zero. This rapid movement of the zero, together with the less extensive changes of the poles, causes a rather abrupt change in spectrum amplitude in the mid- and high-frequency range at the implosion and release of the consonant. A spectrogram of the synthesized sound is shown in Fig. 6. This spectrogram is similar to those observed in spoken syllables containing nasal consonants. The optimum trajectories for the additional pole and zero depend to some extent on the vowel, but theoretical principles as well as experimental data can help to guide the selection of these trajectories.

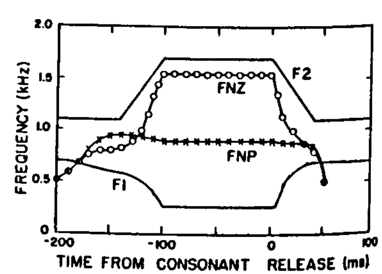


Fig. 5 Time course of several of the parameters used to synthesize the utterance /ɑnɑ/.

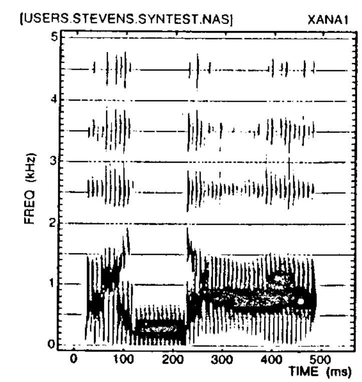


Fig. 6 Spectrogram of the utterance /ɑnɑ/ synthesized using the parameters in Fig. 5.

When nasal consonants are synthesized by this procedure, the amplitude and spectrum of the glottal source remains fixed as the boundary between the consonant and the vowel is traversed. This is, of course, just what happens for human production of syllables of this kind. The abrupt change in spectrum amplitude at mid and high frequencies across a transition from a nasal consonant to a vowel occurs as a consequence of rapid changes in the characteristics of the transfer function and not as a consequence of a change in the source.

5. SYNTHESIS OF LIQUID CONSONANTS

The configuration with an additional pole-zero pair in the cascade branch makes it possible to synthesize liquid consonants in a natural way, as well as nasal consonants. As Fant showed in his early work [1], the transfer function of the vocal tract for a lateral consonant configuration is characterized by a zero or antiresonance, usually in the frequency range 2-3.5 kHz, as a consequence of the formation of a side branch in the acoustic path by the tongue blade. As the lateral consonant is released into a following vowel, this zero is annihilated or canceled by a pole. The elimination of the zero occurs rapidly, and results in an abrupt increase in spectrum amplitude in this higher frequency range.

Typical trajectories for the resonances and the antiresonance needed to synthesize a lateral consonant in a consonant-vowel syllable are given in Fig. 7. A retroflex consonant of the type that occurs in American English is also characterized by an additional pole-zero pair in the vocal-tract transfer function. The time course of the frequencies of this pole and zero, as well as of the other resonances, for a syllable like /rɑ/ is also given in Fig. 7. Not shown in the figure are bandwidth changes for some of the formants as the consonant is released from the

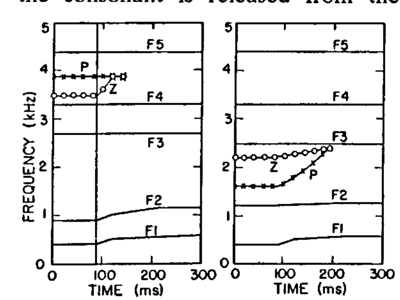


Fig. 7 Time course of the formant parameters and the additional pole-zero pair used to synthesize the syllables /lɑ/ (left) and /rɑ/ (right).

constricted position. Acoustic analysis of synthesized syllables produced in this way shows spectral changes very similar to those in natural utterances. In particular, the discontinuity in spectrum amplitude at high frequencies at the release of the lateral consonant is achieved readily by the rapid movement of the zero in the transfer function.

6. GLOTTAL SOURCE AND TRACHEAL COUPLING

Perhaps the most significant improvement that Klatt made to the synthesizer configuration is related to the glottal source and the effect of the trachea. The details of the design of this source were given in a paper by Klatt and his daughter which appeared in the Journal of the Acoustical Society of America in February of 1990 [9]. The new synthesizer includes a controllable glottal source that is a modification of the source proposed by Fant, Liljencrants and Lin [2], and by Rosenberg [11]. The voicing source and the aspiration source are generated in the manner shown in Fig. 8. Not shown in the figure is the fact that the effects of the radiation characteristic have been folded into the source models, in effect yielding a waveform that is the time derivative of the output shown in this figure. The source controls are arranged so that adjustment of the open quotient OQ only affects the spectrum of the source at low frequencies, and has little influence on the high-frequency spectrum amplitude. The high-frequency amplitude is varied by manipulating the TL parameter. Numerically TL is the reduction (in decibels) in the spectrum amplitude of the source at 3 kHz, relative to the spectrum amplitude of a source with a simple 12 dB per octave slope at high frequencies.

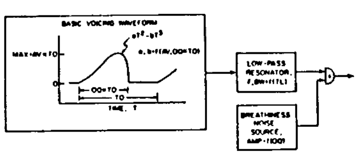


Fig. 8 Block diagram of the voicing source for the KLSYN88 formant synthesizer. The effects of the radiation characteristic have also been folded into the source models, resulting in a voicing source spectral output that falls off at about 6 dB/oct at high frequencies and an aspiration source spectrum that is essentially flat over the frequency range of interest.

Addition of an aspiration noise source to simulate turbulence noise in the vicinity of the glottis takes into account the spectrum of the turbulence noise and, when noise and voicing occur together, reflects the fact that the airflow, and hence the noise, are modulated. Furthermore, the turbulence noise source is, for the most part, a sound-pressure source, whereas the periodic glottal source is essentially a volume-velocity source. The spectral and temporal characteristics of the aspiration source in KLSYN88 are adjusted to take these factors into account, so that no further adjustment of the spectrum of the noise is necessary when the combined source forms the input to the cascade branch of the synthesizer.

An example of the smoothed spectrum of the synthesized vowel /ɑ/, superimposed on the smooth spectrum for /h/ with the same formant frequencies and synthesized with the laryngeal source of Fig. 8 is shown in Fig. 9. The spectrum of /h/ in relation to that of the vowel shows substantial differences in amplitude at low frequencies but similar spectrum amplitudes in the F4 and F5 region. This comparison is consistent with acoustic data from natural speech as well as theoretical predictions [10], [12].

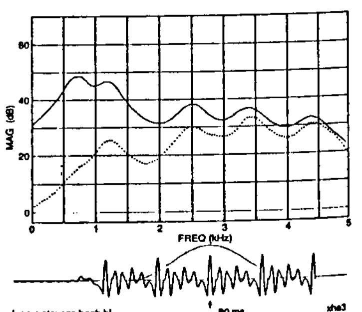


Fig. 9 Smoothed spectra of the sounds produced by the synthesizer for the vowel /ɑ/ (solid line) and when the periodic glottal source is replaced by the aspiration source, with a widened first formant (dotted line).

Proper adjustment of the parameters OQ (open quotient), TL (high-frequency tilt), and AH (amplitude of aspiration noise) permits the generation of a glottal source with a spectrum that is a good approximation to the spectrum of the glottal source for almost any female or male speaker. Furthermore, within an utterance by a given speaker, these parameters can be varied as active laryngeal adjustments are made to produce voiceless obstruent consonants or prosodic changes within phrases and sentences. The parameters can also be modified as the laryngeal state reacts passively in response to manipulation of constrictions in the airway, for example during voiced obstruents and for sonorant consonants produced with a narrow constriction.

As we have seen in the cascade branch of the synthesizer in Fig. 4, two pole-zero pairs are included in addition to the series of poles for conventional synthesis of vowels. In addition to their use in the synthesis of nasals and laterals, one or both of these pole-zero pairs can be employed to simulate acoustic coupling to the trachea when the glottal opening is sufficiently large. Proper positioning of a pole-zero pair introduces an additional peak in the spectrum at a relatively fixed frequency. The most prominent peak is usually the second tracheal resonance, which is in the frequency range 1400-1800 Hz for an adult.

To illustrate how the parameters of the glottal source can be manipulated to match the voices of different male and female speakers, we have attempted to match the spectra of selected vowels produced by several speakers, by manipulating the parameters of the synthesizer. In particular, we adjusted the frequency and amplitude of the glottal source and the formant frequencies and bandwidths to match the corresponding measured characteristics in the spoken vowels. We then manipulated the glottal parameters OQ and TL to provide a best match to the spectrum. Examples of the spectra of the spoken vowels and of the best matching synthesized vowels are given in Fig. 10. These spectra illustrate quite diverse characteristics of the glottal source. For the female voice at the top, the parameter OQ was 70 percent, resulting in a prominent first harmonic. The first-formant bandwidth was wide (about 300 Hz), as might be expected with an increased average glottal width. On the other hand, the OQ value needed to match the male spectrum was 30 percent, with a first-formant bandwidth of 100 Hz. A slight high-frequency tilt (TL=2 dB) was necessary for this speaker.

For a number of voices it was possible to obtain a match to within 3-5 dB up to about 4 kHz by selecting values of the parameters OQ and TL. The values of OQ for different voices ranged from 30 to 70 percent, whereas TL was in the range 0 to 10 dB. For some voices, it was necessary to add a pole-zero pair to account for a minor spectral peak arising from acoustic coupling to the trachea. Some aspiration noise is routinely added to the glottal source during voiced intervals, and the amount of aspiration noise varies from speaker to speaker, presumably.

7. SYNTHESIS OF VOICED AND VOICELESS CONSONANTS: GLOTTAL SOURCE ADJUSTMENTS

As has been noted above, a speaker makes significant adjustments in the waveform of the glottal source during the production of various types of consonants, as well as in the syllabic nuclei over longer time intervals within a phrase or sentence. Dennis Klatt illustrated several of these types of adjustments in his paper published in 1990. Figure 11 shows a typical pattern of change of some of the glottal and other related parameters that are manipulated when a voiceless aspirated stop consonant in intervocalic position is synthesized. These parameters can best be interpreted in terms of the effects of the glottal spreading maneuver that occurs in conjunction with the supraglottal closing movement. The open glottis that assists the termination of voicing at the end of the first vowel

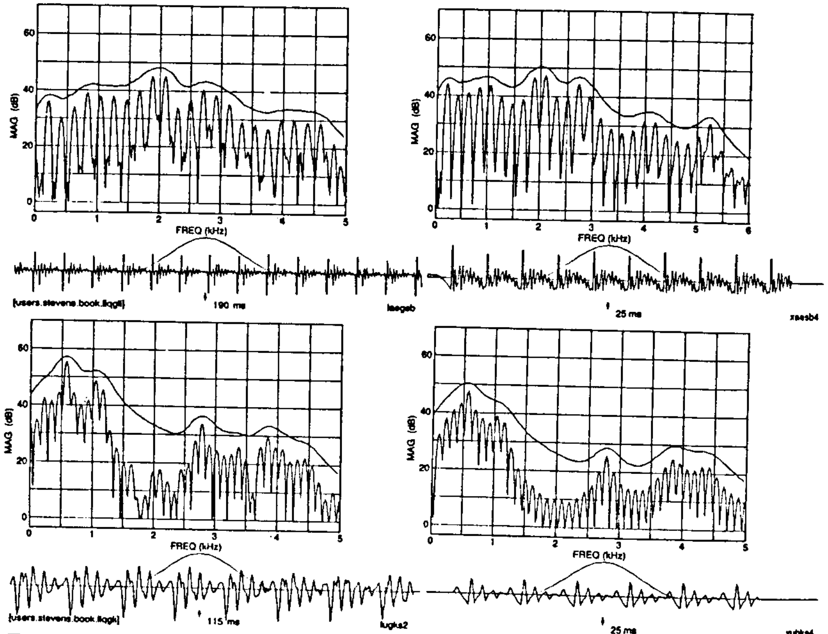


Fig. 10 The left panels show the spectra of two vowels produced by a female speaker (top) and a male speaker (bottom). The right panels are the spectra that are produced by the synthesizer when the glottal and bandwidth parameters are adjusted to give a best match. See text.

is reflected in an increased open quotient OQ, an increased high frequency spectral tilt TL, an increased amplitude of aspiration noise AH, and a widened bandwidth B1. The reverse occurs following the consonantal release preceding the onset of the second vowel. The transitions in formants F1 and F2 reflect the supraglottal movements toward and away from closure, and the burst of frication noise (identified by AF) also indicates the consonantal release. Thus a large number of synthesizer parameters need to be manipulated in order to provide an accurate acoustic representation of the glottal and supraglottal movements needed to produce the stop consonant.

8. TOWARDS REVISED RULES FOR SYNTHESIS FROM A PHONETIC INPUT

The few examples of synthesis of vowel-consonant and consonant-vowel sequences given here have indicated that if rules are to be formulated for synthesizing utterances from a phonetic input, these rules must specify the time variation of an extensive set of control parameters. We turn now to consider what these rules are trying to capture and how they might be organized.

In the case of vowels, the rules are relatively simple, and they specify the time course of a small number of formant frequencies --- usually just three formants. The glottal configuration specified for the vowels may change slowly depending on prosodic considerations. With little additional complexity, synthesis of the glides /w/ and /j/ can be specified in terms of formant movements. Since these glides involve a more constricted vocal-tract configuration, adjustment of the bandwidths of some formants may be necessary to account for the additional acoustic losses in the vocal tract, and some modification of the glottal vaveform (increased OQ and TL) may occur be-

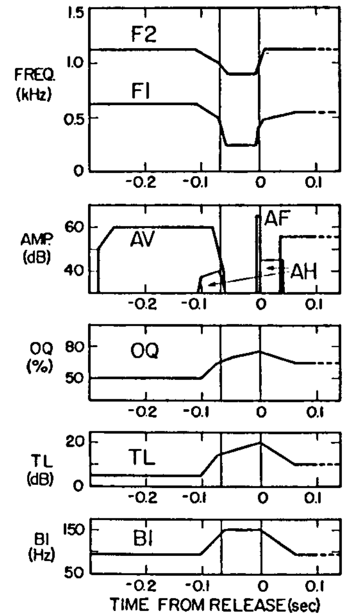


Fig. 11 Time course of several of the parameters used to synthesize the utterance /ɑpə/. The vertical lines indicate the times of voicing offset and consonantal release.

cause of interaction of the glottal source and the vocal-tract acoustics. In the case of /h/, the spread glottal configuration is marked by appropriate adjustment of the parameters OQ, TL, AH, and B1.

For the consonantal segments the synthesis rules become substantially more complex. At least eight parameters need to be manipulated, for example, to synthesize the sound that occurs when a voiceless aspirated stop consonant is released into a vowel, as shown in Fig. 11. A similar large number of parameters is needed for nasal consonants and fricatives. Some of the parameters are related directly to the movement of the articulator that forms the consonantal constriction --- usually the lips, the tongue blade, or the tongue body. Other parameters are influenced primarily by adjustments of the laryngeal configuration, the velopharyngeal opening, or other articulators that are not directly involved in forming the constriction. There are, however, some interactions between these groups of parameters primarily through aerodynamic and acoustic processes. Thus, for example, the amplitude of the burst at the release of a stop consonant is determined in part by airflow that may be limited by the laryngeal configuration. Or the degree of abduction or adduction of the vocal folds can influence the frequencies and bandwidths of the lower formants. Or, the velopharyngeal opening can cause shifts in the natural frequencies of the vocal tract whose movements signal the place of articulation of a nasal consonant.

One approach to synthesis with the large array of parameters is to design a set of higher-level control parameters (HL parameters) that are related more closely to articulatory parameters than are the acoustically oriented parameters currently used to control KLSYN88 (KL parameters) [13]. The arrangement is shown schematically in Fig. 12. The HL parameters specify articulatory dimensions such as size of velopharyngeal opening, area of a consonantal constriction, and glottal adduction-abduction. The lower-level KL parameters that control the synthesizer itself are derived from the HL parameters through a set of mapping relations. The mapping relations automatically incorporate the constraints that exist between the various KL parameters because of aerodynamic and acoustic interactions.

The generation of speech with the KLSYN88 synthesizer requires that quantitative data and explicit models be developed in two areas of phonetics. One area is concerned with the constraints that the articulatory and aerodynamic systems impose on the sound. In terms of the diagram in Fig. 12, these are the mapping relations between the HL parameters and the KL parameters. The other area involves the temporal control of the articulatory processes, as these processes are manifested in the HL parameters.

Developing the mapping relations requires that theories and models of glottal vibration, aerodynamic noise generation and vocal-tract acoustics be re-

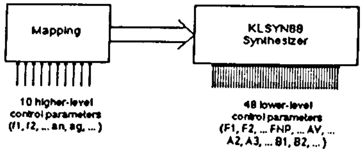


Fig. 12 Block diagram of a system for controlling the parameters of the KL-SYN88 synthesizer by a reduced array of HL parameters. The KL parameters are derived from the HL parameters by a set of mapping relations.

fined. Examples of problems that must be addressed are: estimating the distribution of turbulence noise sources when there is a constriction in the vocal tract, determining the time course of onsets and offsets of vocal-fold vibration for voiceless consonants, and modelling the acoustic losses when there are consonantal constrictions.

Refinement of our understanding of articulatory control processes highlights the need for several types of data and models. Quantitative data must be obtained on rates of release and closure of articulators that form the primary consonantal constrictions for stops, fricatives and affricates. It is necessary to determine how articulatory parameters that are not directly involved in forming the consonantal constriction are timed in relation to the primary articulators. For example, is the control of these secondary articulators adjusted so that the acoustic evidence for the movements of these articulators is optimally represented in the sound?

Beyond these problems of controlling the synthesizer parameters to produce a representation of speech sounds in syllables, there are a variety of questions relating to timing and prosody of larger units. Klatt made important contributions here, with his extensive data and rules concerned with segmental durations [6] and his implementation of rules for generating fundamental frequency contours in phrases and sentences [8].

The existence of a speech synthesizer such as KLSYN88, with its ability to generate speech of high quality, focusses attention on these problems that are central to the study of phonetics --- problems relating to how individual sounds are produced, how different articulatory structures are coordinated, and how larger production units are organized. In this sense, the synthesis work of Klatt has not only contributed a body of knowledge to phonetics but has also provided a focus and a stimulus for future research.

9. ACKNOWLEDGEMENT

The preparation of this paper was supported in part by grant DC-00075 from the National Institutes of Health.

10. REFERENCES

[1] FANT, G. (1960), "Acoustic theory of speech production", The Hague: Mouton & Co.

[2] FANT, G., LILJENCRANTS, J., & LIN, Q.G. (1985), "A four-parameter model of glottal flow", Speech Transmission Labs. QPSR 4, Royal Institute of Technology, Stockholm, 1-13.

[3] FUJIMURA, O. (1960), "Spectra of nasalized vowels", Research Laboratory of Electronics QPR 62, Mass. Inst. of Technology, Cambridge, MA, 214-218.

[4] FUJIMURA, O. (1962), "Analysis of nasal consonants", J. Acoust. Soc. Am. 34, 1865-1875.

[5] HAWKINS, S.S. & STEVENS, K.N. (1985), "Acoustic and perceptual correlates of the non-nasal/nasal distinction for vowels", J. Acoust. Soc. Am. 77, 1560-1575.

[6] KLATT, D.H. (1979), "Synthesis by rule of segmental durations in English sentences", in Frontiers of speech communication research, edited by B. Lindblom and S. Ohman (Academic Press, New York), 287-300.

[7] KLATT, D.H. (1980), "Software for a cascade/parallel formant synthesizer", J. Acoust. Soc. Am. 67, 971-995.

[8] KLATT, D.H. (1987), "Review of text-to-speech conversion for English", J. Acoust. Soc. Am. 82, 737-793.

[9] KLATT, D.H. & KLATT, L.C. (1990), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am. 87, 820-857.

[10] MANUEL, S.Y. & STEVENS, K.N. (1989), "Acoustic properties of /h/", J. Acoust. Soc. Am. 86, Suppl. 1, S49.

[11] ROSENBERG, A. (1971),"Effect of glottal pulse shape on the quality of natural vowels", J. Acoust. Soc. Am. 49, 583-590.

[12] STEVENS, K.N. (1990), "Noise at the glottis during speech production", J. Acoust. Soc. Am. 87, Suppl. 1, S121.

[13] STEVENS, K.N. & BICKLEY, C.A. (1991), "Constraints among parameters simplify control of Klatt formant synthesizer", J. Phonetics 19.

PERCEPTION: AUTOMATIC AND COGNITIVE PROCESSES

Terrance M. Nearey

University of Alberta, Edmonton AB, Canada, T6G 2E7

ABSTRACT

Speech perception is that process by which humans map acoustic waveforms onto strings of linguistic symbols. While accepting many of the premises about the complexity of signal-to-symbol mapping that have been so influential in Liberman and Mattingly's [9] motor theory of speech perception, it is argued that there exists an upper limit on that complexity imposed by perceptual mechanisms that map acoustic properties directly onto phonological units [15]. Evidence for this claim is presented together with its implications for the nature of cognitive processes in speech perception and their relative automaticity.

1. INTRODUCTION

The terms "automatic" versus "cognitive" may be seen to relate to a continuum of computational complexity. *Automatic* processes are likely to be associated with simple architectures with few free parameters, with a constrained, largely bottom-up data flow and with an overall "reflex-like" character. *Cognitive* processes may have a complex architecture with many free parameters, may possess a less constrained, strongly "top-down" data flow and may exhibit relatively "intelligent" behavior. These concepts are closely related to aspects of Fodor's modularity hypothesis [5].

Fodor proposes that there exists a highly flexible central processor representing the pinnacle of cognition. The central processor serves as a kind of "executive Sherlock", brilliantly integrating and evaluating information from a variety of sources. However, Sherlock's data doesn't come directly from the world at large; rather, it is bureaucratically passed upstream by a set of clever but narrow minded "forensic specialists", the input modules, who preprocess raw input in highly stylized ("work-to-rule") ways.

Fodor postulates the following sobering, if tongue-in-cheek, first law of cognitive science: "The more global (i.e., isotropic) a cognitive process is, the less anybody understands it." Isotropic is a term borrowed from the philosophy of science, meaning "facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established truths." In Fodor's scheme, isotropism is a property only of the central processor, while input modules are much more constrained in their operation. Furthermore, he contends, it is precisely because they are of limited cognitive capacitiy that we are able to understand them at all. Input modules are viewed as computationally complex but specialized "cognitive reflexes" that constitute, in part, "the means whereby stupid processing systems manage to behave as though they were smart ones (p. 81)."

1.1 Motor (Gestural) Theories

Liberman and Mattingly (= "LM" [9]) adopt an overtly modularist perspective and they martial a wide variety of arguments in support of a special phonetic decoder as an input system in the Fodorian sense. Although their other arguments are important, I will be concerned only with the problem of the signal-to-symbol mapping, that is, to listeners' categorization of speech.

LM state their main premise as follows: "The first claim [of the motor theory] is that the objects of speech perception are the intended phonetic gestures of the speaker, represented in the brain as invariant motor commands, that call for movements of the articulators through certain linguistically significant configurations (p. 2)." They continue: "But the relationship between the gesture and the signal is not straightforward. The reason is that the timing of articulatory movements --the peripheral realizations of the gestures-- is not simply related to the ordering of gestures that is implied by the strings of symbols in phonetic transcription (p. 3)." Thus, in this framework, an articulatory space is essential in understanding the signal-to-symbol mapping.

To fix ideas, consider an example from Cooper et al. [3] involving classification of voiceless stop+vowel stimuli. An approximation of the decision space for this experiment is given in Fig. 1. It shows dominant response regions for the three voiceless stops /p/, /t/ and /k/ for a stimulus space consisting of a narrow band noise burst (characteristic of plosive release) followed by two-formant synthesized vowels. The authors emphasize that there are no absolutely invariant properties associated with perception of the stops, but only fairly complex relational properties. Such complexity, they contend, "requires the consonant-vowel combination as a minimal acoustic unit (p. 598)." The motor theory claims such intricate acoustic-to-phonetic patterns can be understood only by reference to the Rosetta stone of the underlying gestures. i.e., in this case, the coarticulation of a stop with its following vowel.

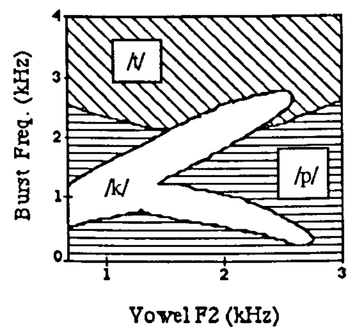


Figure. 1. Categorization of stop consonants (after [3]).

Fowler [6] and her colleagues also argue for the close relation between perception and articulation, but propose Gibsonian "direct perception" of the *actual* articulatory gestures of the moving vocal tract. Liberman and Mattingly are highly skeptical of this possibility, noting that, e.g., "given the many-to-one relation between vocal-tract configurations and acoustic signal, a purely analytic solution to the problem of recovering movements from the signal seems to be impossible." LM believe that it is "phonetic intentions", rather than actual peripheral events that count and they advocate analysis-by-synthesis (=ABS) decoding, whereby the phonetic module "has merely to determine which (if any) of a small number of gestures that might have been initiated at a particular instant could, in combination with gestures already in progress, account for the signal (p. 27)."

Klatt [8] agrees with LM's pessimistic assessment of the possibility of the recoverability of gesture from the acoustic waveform. But, having actually explored ABS for automatic speech recognition, he seems to conclude that it too is computationally intractable. Klatt nonetheless believes that "[p]roduction and perception are clearly closely tied in the sense that perceptual strategies must know a great deal about production options and their acoustic manifestations (p.178)." Though he voices hope for a more scientifically appealing approach in the future, he suggests that the most promising way to deal with context dependency is to pre-compile acoustic patterns of words in large networks that make no use of traditional phone-sized units.

The strong motor theory position of LM is best motivated in the context of a "full speed ahead and damn the torpedoes" model of speech production, as stereotyped in Hockett's "soft-boiled Easter egg plus wringer model." As MacNeilage [12] points out, this charicature bears a striking resemblence to actual models of speech production of the late 60's. MacNeilage's (and much subsequent) work has shown this view to be very mistaken. Rather, despite the persistence of residual peripheral variability, the motor system is capable of performing rather remarkable feats in achieving relatively invariant peripheral manifestations of articulatory targets.

1.2 Auditory Theories

As the peripheral configurations associated with phonetic elements approach invariant targets, so to do their acoustic consequences. Many researchers believe that acoustic/auditory properties have a direct role in defining goals for speech production and perception. Blumstein and Stevens [1] and their colleagues argue for relatively invariant signal properties that actually motivate target articulations for stop consonants. Nearey [18] claims that acoustic properties of vowels exhibit demonstrably greater invariance across speakers than do articulatory manifestations. Diehl and Kleunder [4] compile arguments for the primacy of auditory rather than gestural considerations in speech perception. Finally, from diverse perspectives, researchers including Martinet, Lieberman, Lindblom and Ohala have insisted on emphasizing simultaneously articulatory and acoustic properties in understanding the long-term (diachronic) properties and even the evolution of language capacity.

2. SEGMENTS AS SYMBOLS

Nearey [15] argues that speech production and perception represent a compromise in complexity between articulatory and acoustic patterns. I elaborate here on that framework from a neo-Sapirian perspective. There are (at least) three domains involved in speech, two physical and one symbolic. The symbolic part consists of the sequence of language-specific phonological elements. Without loss of generality (we can change our minds later), assume those symbolic elements are "phoneme-size" units called segments. The two physical domains are the articulatory (gestural) and the acoustic (auditory). Speech production is the mapping from segments to signals and speech perception is the opposite (not to say inverse) mapping.

In *strong motor* theories (e.g., LM) it is assumed that a natural invariant relationship exists between gestures and segments, while the mapping from acoustics to segments can be arbitrarily complex. In *strong auditory* theories, the roles of acoustics and articulation are reversed. In *double-strong* theories (e.g. Blumstein and Stevens taken to the extreme), the relationship of both physical domains to segments is assumed to be natural and invariant.

I have long been impressed by the sophistication of arguments of the strong frameworks and of the scholarliness and sincerity of their proponents. In fact, they have each convinced me that the others are wrong. To resolve this, I have adopted a "symbolic segment model" that is *double-weak* (weak motor, weak auditory). Segments are symbols and are neither articulatory nor acoustic in character. There is no fully natural, simple invariant relationship between either gesture or signal and symbol. Left as it stands, this amounts to a retreat into radical structuralist arbitrariness, wherein almost anything can happen in phonetics. Yet, if the gesture- and/or sound-to-symbol relationship is tightened up in the extreme, we arrive back at one of the three strong systems of the preceding paragraph.

Instead, I assume that the relationships between gesture segment and signal approach an "equilibrium of complexity" [15], a compromise between efficiency in production and rapid decoding in perception. Both auditory and articulatory properties will have a profound long-term influence on phonological systems. However, for perception only "the weakest form of a motor theory [8] (p. 204)" holds, which involves merely "what has been learned about relations between speech-production capabilities and the resulting acoustic output." Conversely, a weak form of an "auditory theory of speech production" also holds: articulatory targets and permissable coarticulation are constrained by what limited perceptual structures can readily decode. The kinds of constraint I have in mind might be formulated as follows: 1) A relatively simple, but not fully transparent, family of articulatory patterns is associated with each symbol. 2) A relatively simple, but not fully transparent, family of acoustic patterns is also associated with each symbol. 3) The relevant families of patterns exhibit moderate within-category variation relatable to contextual factors its own physical domain (articulatory or auditory).

In Fodorian terms, this is a proposal about the limits of computational complexity of separate sub-modules for speech perception and production. Long term pressures force each to respect the other's limitations, while allowing each to exploit the other's flexibility. However, the real-time operation of each is independent of the other. Each is an encapsulated, relatively "stupid" processor that only "looks" like it knows about the internal workings of the other (cf. the discussion of lexical priming in [5]).

Though limitations of arbitririness are imposed by this "symbiosis of encapsulated modules", there is room for much variety in how different languages approach their equilibrium of complexity. This allows for a language-specific Sapirian "warping" of the possible phonetic space. Explicit modeling of how the putative perceptual sub-module might implement such a warping can shed light on the question of its computational complexity.

2.1 Segmental Filter Models

The modeling framework proposed below is a generalization of the pattern recognition system proposed by Nearey and Hogan [16] to account for language-specific warping of the cue-space for simple experimental situations. In this "segmental filter" model, speech perception is assumed to involve an essentially bottom up, "reflex-like" mapping between properties of acoustic waveform and phonological segments. It is assumed further that the following limit exists on this mapping: all the knowledge that the perceptual system has about the consequences of patterns of production can be embodied in a set of Gaussian "filters", one for each phonological segment, tuned by acoustic/auditory properties. (These models are formally related to Massaro's FLMP, see [15].)

2.1.1 The NAPP Model

Fig. 2 illustrates the general properties of a simplified network of Sapirian segmental units. This example is based on the Thai data of [11, 15, 16]. In panel (a), the output of three VOT-tuned stop filters (for /d/, /t/ and /t^h/) are shown. These filters produce output, reflecting the "typicalness" of a given stimulus considered as a member of each category.

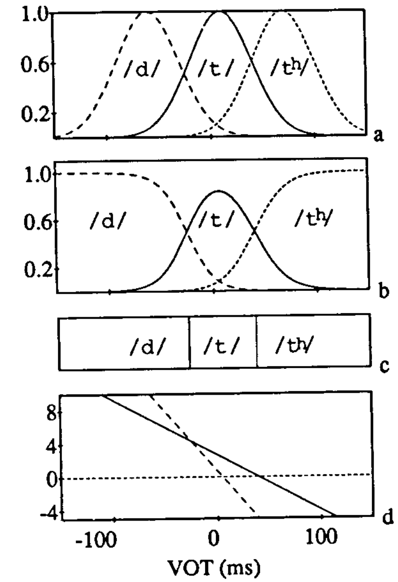


Figure 2. Segmental filters

In categorization, the probability of chosing a particular response, *r*, for a particular stimulus value, *s*, is a function of the relative distance from the "most typical value" for that category compared to the sum of the analogous distances from all three categories. Formally, this process can be separated into an *evaluation* function that determines the "fuzzy typicalness values" of Fig. 2(a) and a *choice* function that converts them into response probabilities. For a version of the "normal *a posteriori* probability" or NAPP model [15, 16], the evaluation function may be specified as follows:

$$f(r,s) = \frac{-.5\,[x(s)-m(r)]^2 + k(r)}{D(r)}, \qquad (1)$$

where *x(s)* is the stimulus value (here, VOT) for stimulus *s*, *m(r)* is the mean VOT value for response category *r* (ranging over /d/, /t/ and /t^h/), *D(r)* is the standard deviation of VOT values in category *r* and *k(r)* is a normalizing parameter parameter that can accommodate a Bayesian *a priori probability* and can absorb a response bias term in a perceptual model. The models considered below involve homogeneous (homoschedastistic) Gaussian distributions, where the standard deviations (and correlations, in the multivariate case) are equal across all choice categories.. The choice function for the NAPP model is:

$$p(s,r) = \frac{\exp[f(r,s)]}{\sum_{r'} \exp[f(r',s)]}, \qquad (2)$$

where f(s,r) represents an evaluation function defined in (1) and the summation in the denominator is over all response categories.The result of the application of Equations (1) and (2) is shown in panel (b) of Fig. 2. which constitutes the *response surface* of the model and contains all the information about its stimulus-response mapping. In panel (c), the stimulus space is divided into three regions, each labeled with the dominant response in its range. The boundaries of such a *territorial map* or *decision space* are determined by the crossover points in the response surface of neighboring categories, which are in turn determined by the crossover points in the evaluation functions.

2.1.2 Logistic Models

For homogeneous Gaussians, the same response surface, (and decision space) can be formed by a set of three *linear logistic functions* as illustrated in panel (d) of Fig. 1. The equations for these lines are specified by:

$$f(r,s) = b(r) + a(r)\, x(s), \qquad (3)$$

where *b(r)* represents a bias term for the category *r*, which is *independent of the stimulus* value; while *a(r)* is a "stimulus-tuned effect." Such linear logistic models are choice-equivalent to Gaussian filter models. As discussed in detail in [15], logistics are readily estimable and can be generalized to characterize very complex decision spaces and response surfaces in a way that can be given interesting phonetic interpretations.

3.0 SEGMENTS VS. DIPHONES

A key property of segmental filter models is that they assume that stimulus properties are mediated in a fundamental way by phonological units of segment size. It has been demonstrated recently by Whalen [21] that response patterns from several experiments show that pure segmental models are not adequate to account for perceptual results. Nearey [15] shows that while Whalen's claim is true in a strict sense, only a minor modification of a "pure segmental" assumptions are motivated by available data.

Whalen sets out to test a claim by Mermelstein [14] concerning the independence of categorization of adjacent segments. Mermelstein's experiment involved simultaneous identification of both vowel and consonant in synthetic VC syllables when F1 and vowel duration were varied. The response categories ranged over the English words "bed, bet, bad, bat." Mermelstein's reults indicated that although vowel duration affected both vowel responses and consonant responses, *vowel and consonant judgments were made independently.*

In a series of analyses involving experiments with ambiguous VC and CV sequences, Whalen finds evidence counter to Mermelstein's claim, showing instead that the judgment of adjacent segments shows interdependencies consistent with a more complex decoding of production effects, in accord with a motor theoretic interpretation. Whalen's Experiment 3 involves categorization of fricative plus vowel sequences, spanning the choice set /si,su,ʃi, ʃu/. The kind of variation involved is typically described as coarticulatory, the most noticeable effect in production data being that frication noise for both fricatives has a lower low-frequency cutoff before /u/ than before /i/, presumably due to anticipatory coarticulation of lip rounding.

Based on previous experiments, Whalen notes that changes in vowel quality from /i/ to /u/ lead to fewer /ʃ/ and more /s/ responses for a given fricative noise. Conversely, changing a fricative context from /s/ to /ʃ/ causes more /i/ and

fewer /u/ responses for vowels in an /i/-/u/ F2 continuum. In order to evaluate the contributions of physical versus phonological context, Whalen's Experiment 3 uses a two-parameter continuum spanning the four diphone choices. The parameters in question are: 1) F2 of a steady state vowel, ranging from 1386 to 1773 Hz in four steps; and 2) the frequency of a fricative pole (with a correlated zero located 1000 Hz below the pole) ranging from 2900 to 3100 Hz.

Roughly speaking, the fricative pole frequency, Pf, can be considered a "primary cue" for the /ʃ-s/ contrast, while F2 is the primary cue for /i-u/. However, Whalen's experiment shows that the /ʃi-si/ boundary along a Pf continuum differs from that of /ʃu-su/ in manner broadly in accord with production norms.

Three general varieties of effects of "vocoid" on "contoid"can be distinguished 1)The *physical value* of *F2* also directly affects or acts as a "secondary cue" for /ʃ-s/; 2). The /i-u/ *judgment* affects the fricative response independently of the stimulus or 3) the acoustic properties directly affect the consonant and vowel choice in a manner that cannot be decomposed into effects like (1) and (2), so that a diphone is the smallest phonological unit that can be thought of as being directly tuned by acoustic properties. From a motor theory perspective, the last alternative represents a model that precompiles contextual variation into larger, more nearly invariant syllabic units.

3.1 Modeling

Nearey [15] presents a series of logistic analyses of Whalen's data which allow for the modeling of increasingly complex response surfaces using ANOVA-like factoring of terms. Specifically, it allows a decomposition of stimulus-response relationships in terms of 1) "stimulus-tuned" effects which cause changes in response probabilities as a function of changes in stimulus properties and 2) bias effects that are independent of stimulus properties. A further breakdown is possible in terms of the "size" of the phonological entity being considered, segments versus diphones. The factorization is represented by the terms in Table 1 (see [15]).

Table 1. Terms in logistic model.

Abbrev.	Term	Unit
Bias effects (stimulus-independent):		
V	bV(v)]	Vowel
C	bC(c)	Consonant
CV	bCV(c,v)	Diphone
F2-Tuned effects:		
VF	a1V(v) F2	Vowel
CF	a1C(c) F2	Consonant
CVF	a1CV(c,v) F2	Diphone
Fricative Pole-tuned effects:		
VP	a2V(v) Pf	Vowel
CP	a2C(c) Pf	Consonant
CVP	a2CV(c,v) Pf	Diphone

Various models can be constructed from these elements, producing decision spaces of varying complexity. For technical reasons, all models include the segmental bias terms V and C. A primary cue model would include the terms VF (vowel tuned by F2) and CP (consonant tuned by Pf). Its decision space is shown in Fig. 3. Note that the vowel boundary is independent of the fricative pole (parallel to the Pf axis) and the consonant boundary is similarly independent of F2.

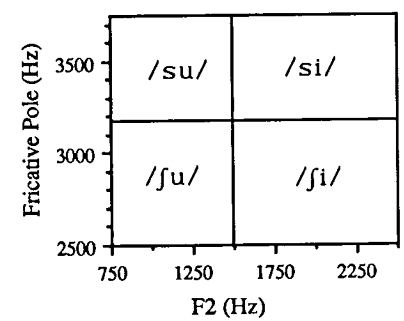


Figure 3. Primary cue model.

A secondary cue model could include the additional stimulus-tuned terms VP and CF and could lead to the decision space of Fig. 4. It remains a *pure segmental* model because *none* of the diphone terms of Table 1 are included; that is, although their *cues* overlap, the *symbolic* remain segmental. Note that, while in some sense, this builds in context sensitivity that seems to have articulatory motivation, it does so "unintelligently" in that a generalized *acoustic* context effect can be directly incorporated as a (secondary) cue in the individual *consonant* filters, independent of phonological context, as in Mermelstein's original suggestion[14].

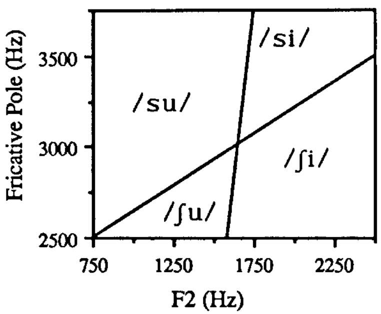


Figure 4. Pure segmental secondary cue model.

However, Nearey's analysis indicates, in accord with Whalen's claim, that *no* pure segmental model can adequately account for Whalen's empirical results.

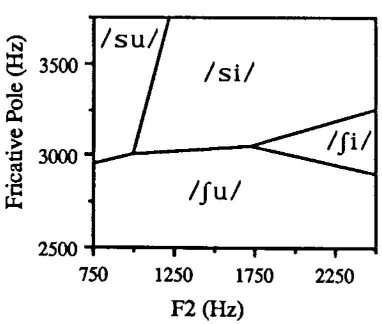


Figure 5. Hypothetical true diphone model.

On the other hand a true diphone model including *all* the terms in Table 1 can approximate more "intelligent" phonological context dependencies, achieving a decision space as complex as that of Fig. 5. But Nearey finds that such true diphone models are too powerful and, instead, an intermediate class of models, refered to as "transsegmentally biased segmental models" is completely adequate to account for Whalen's data.Such models can include all the terms of the secondary cue model plus the diphone bias terms (CV). However, the stimulus-tuned diphone terms (CVF, CVP) are *not* included. The decision space for the "best" model in Nearey's analysis is shown in Fig. 6.Formal properties of the model require that line segments separating syllables that share one segment must be parallel to each other, a restriction not shared by true diphone models. In fact, the model finally selected by Nearey

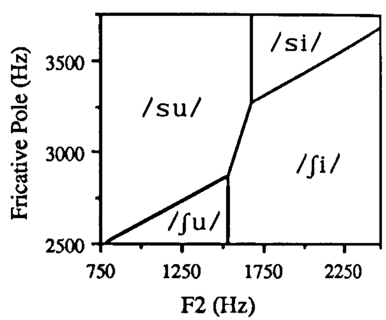


Figure 6. Restricted diphone-biased secondary cue model.

as best supported by the data is slightly simpler than the most complex possible biased segmental model, since the fricative-pole tuned vowel term (VP) is not included in this model. This is reflected by the fact that the /si–ʃi/ and /su–ʃu/ boundaries are parallel to the Pf axis in Fig. 6. Because of formal properties of the biased segmental models, these restrictions on parallelism of lines also would extend to *all other vowel conexts*, so that in a larger experiment, with more vowel responses, the same slope of the /s–ʃ/ boundary would be predicted within *all* vowel categories. In other words, the *relative efficacy* of the two cues (F2, Pf)

in *changing* /s/ to /ʃ/ would be the same, independent of the following vowel.

3.2 Correction for coarticulation and allophonics

The above results have an interpretation in terms of a Fodorian "pseudo-smart" (one that is stupid, but looks smart) processor for coarticulation effects. Consider the following: Anticipatory lip-rounding makes /s/ more /ß/- like before /u/, while anticipatory spreading makes /ʃ/ more /s/ -like before /i/. That is, the vowel environments tend to produce "weaker cues" in those environments. But the diphone biases have the net effect of favoring the combinations with weakened contrasts, thus increasing their response areas. However, although useful, this is not a truly "intelligent correction", since it is *not cue sensitive*, but rather is a global bias on category pairs. This has implications for new experiments with more stimulus dimensions: namely that the response areas of the favored syllables would be in-creased along all stimulus axes vis-à-vis less favored ones, even those not affected directly by the main coarticulation effect in production. (So, /ʃi/ might "encroach" on /ʃɪ/ along the F1 axis, even though F1 was not involved in fricative vowel coarticulation).

In addition to its possible role as a "coarse correction" for coarticulation, Nearey notes [15] that many other experiments reported in the literature seem to be compatible with the restrictions of the biased segmental model and that there is as yet no clear experimental evidence to indicate that models as complex as true diphone models are ever required. Biased segment models can be viewed as a multilayer system. The first layer comprises a set of segmental filters wherein all stimulus tuning takes place, while higher-level units implement additive, stimulus-independent corrections for (passive) coarticulation, (preplanned) extrinsic allophony and phonotactic constraints. Such models also appear adequate to accommodate the kinds of "cognitive context effects" suggested by Ohala [19]. It also appears that the Ganong effect ([7]; see [15]) and the role lexical effects play in Lindblom's hypospeech [10] could be handled by lexical bias effects that do not interfere with the internal operation of the segmental filters.

4.LEXICAL ACCESS

Could informationally encapsulated segmental filters of the type described above really serve as the basis for lexical access? Marslen-Wilson ([13] has divided up the problem of "projecting sound into meaning" into two largely autonomous components: access and integration. Lexical access is viewed as *form-based* processing, whereby bottom-up phonetic information interacts with the lexicon to select a unique lexical item. This lexical candidate is then presented to a higher level "content-based" process of *integration*, wherein the newest lexical item is incorporated into the syntactic and semantic processing of the sentence. From the point of view of the existence of a sub-modular language processing system, the key conclusions are: "First, that sentential context does not function to override perceptual hypotheses based on the sensory input system. (p 19)." Second, that top-down effects (e.g., sentential context) "do not affect the basic perceptual processing of the sensory input." Some of the evidence for these conclusions is considered below.

Important work by Samuel indicates that there are very strong constraints on how syntactico-semantic information influences lower level processing. Samuel's work involves the use of a classical signal detection paradigm to investigate the decomposition of effects into what he refers to as *perceptual* and *post-perceptual* components. Subjects try to detect the difference between two kinds of distorted natural speech: one in which a phoneme has been replaced by noise and one in which noise has been added to the original phoneme. In a series of carefully designed experiments, Samuel varies a number of characteristics, including the phonetic nature of the segments distorted, lexical status (word versus pseudo-word) and sentential context.

Sentential context is shown to only affect listener's bias toward saying "added" (i.e. the phoneme is restored) in appropriate contexts, but the discriminability measure *d'* is not affected. That is, differences between "added noise" and "replaced by noise" stimuli were equally salient to listeners, regardless of sentential context. They were simply globally more likely (biased) to say "restored" to stimuli in appropriate semantic contexts. However, in contrast to syntactico-semantic effects, Samuel's workindicates that *lexical status* (being a real word) may affect lower-level (phono-logical) processing, since *discriminability* for real words was less than for non-words. However, the work of Samuel and Ressler [20] confirms the finding (by Nusbaum and colleagues) that the lowered discriminability for words is strongly affected by attentional factors and may result mainly from subjects' inability to focus on segments within words. While more research is clearly needed, this result, coupled with the tractability of Ganong-effect in synthetic experiments, leaves open the possibility that an encapsulated set of segmental filters operating prior to lexical access.

5 EXTENSIONS AND PROBLEMS

While the biased segmental models seem to be compatible (so far) with a variety of results from the literature, there are at least a few cases that their simple linear boundaries cannot handle. The facts surrounding the famous case of place of articulation of stops appear to require something somewhat more complex. To the best of my knowledge, the Cooper et al. experiment represented by Fig. 1 manifests the most complex decision space ever found in phonetic research. The main pattern as characterized by the authors is roughly as follows: /t/ dominates when burst frequency is high; /k/ when its frequency is slightly above F2; and /p/ otherwise. While beyond the reach of homogeneous dispersion Gaussians, this general pattern can be achieved a Gaussian model which allows separate covariance matrices for each group (corresponding to a quadratic logistic model.; cf. Nearey and Shammass [17] for application of the related quadratic discriminant analysis to *transitions* in stop consonants). Finally, the minor mode of the /k/ region that occurs when the burst is just above the F1 in front vowels may require an additional wrinkle. Nonetheless, the general pattern of this decision space can be generated by segmental filters as described below: A single bivariate Gaussian in F2 and burst frequency is used to characterize each of the /p/ and /t/ distributions. However, /k/ requires a mixture of two bivariate Gaussians: one to characterize the F2 burst relationship and the other for F1 and burst. In fact, Fig. 1 was generated analytically in just this way. Although this pattern is more complex than those of the simple logistics of the previous example, it still represents a relatively elementary problem in pattern recognition. Note that this does not deny that aspects of the pattern are motivated in the long run by articulatory factors, only that real-time perceptual behavior does not need to compute articulatory or gestural properties to decode them. Since this may be as complex it ever gets, there seems good reason to continue to explore segmental filter approaches to speech perception.

There is, however, one very large problem that must be faced squarely in any such exploration: while the above model makes inroads on the traditional problem of invariance, it has ignored the problem of segmentation. First note that although the models are segmental at the symbolic level, they are manifestly not so at the acoustic level, since pervasive temporal overlap is allowed in the cue domains of neighboring segments. Though these segments are not necessarily phonemes ("major allophones" would do), I propose that the constraints of the acoustic-to-segment mapping be modified forms of Chomsky's conditions on the relation "systematic phones" to taxonomic phonemes[2]. (i) weak linearity: the centers of the window of relevance of the acoustic cues preserves the left-right order of the strings of segments. (ii) local determinacy: such windows are not arbitrarily wide; (iii) strict bottom-up mapping (replacing biuniqueness); (iv) higher-order invariance.

With respect to (iv), given temporal alignment of the window of relevance (and cue-extraction!), the claim is that the patterns are relatively invariant, usually mapping to simple linear decision spaces. This, however, is a very large "given." The plausibility of the scheme presented above, no matter how successful it may be for "toy" problems in the phonetics lab, is ultimately dependent on the ability to supply cognitively plausible models of signal alignment. In this regard, I think we have much to learn from the computational methods of time alignment being developed in the speech recognition community.

ACKNOWLEDGEMENTS

Work supported by SSHRC. Thanks to T. Welz for technical support.

REFERENCES

[1] BLUMSTEIN, S. & STEVENS, K. Acoustic invariance in speech production : Evidence from measurements of the spectral characteristics of stop consonants. JASA **66**: 1001-1017, 1979.

[2] CHOMSKY, N. "Current issues in linguistic theory." 1964 Mouton. The Hague.

[3] COOPER, F.S., DELATTRE, P., BORST, J. & GERSTMAN, L. Some experiments on the perception of synthetic speech sounds. JASA **24**: 597-606, 1952.

[4] DIEHL, R. & KLUENDER, K. On the objects of speech perception. Ecological Psychology. **1**: 121-144, 1989.

[5] FODOR, J. "The modularity of mind." 1983 MIT Press. Cambridge, MA.

[6] FOWLER, C. An event approach to the study of speech perception from a direct-realist perspective. J. Phonetics. **14**: 3-28, 1986.

[7] GANONG, W. Phonetic categorization in auditory word perception. Journal of Experimental Psychology: Human Perception and Performance. **6**: 110-125, 1980.

[8] KLATT, D. "Review of selected models of speech perception." Lexical Representation and Process. Marslen-Wilson ed. 1989 MIT Press. Cambridge MA.

[9] LIBERMAN, A. & MATTIGLY, I. The motor theory of speech perception revised. Cognition. **21**: 1-36, 1985.

[10] LINDBLOM, B. "Explaining phonetic variation: a sketch of the H&H theory." Speech production and speech modeling. Hardcastle and Marchal ed. 1990 Kluwer Academic Publishers.

[11] LISKER, L. & ABRAMSON, A. "The voicing dimension: some experiments in comparative phonetics." Proceedings of the 6th International Congress of Phonetic Sciences, Prague. Hala, Romportl & Janota ed. 1970 Academia. Prague.

[12] MACNEILAGE, P. Motor control of the serial ordering of speech. Psychological Review. **77** (182-196): 182-196, 1970.

[13] MARSLEN-WILSON, W. "Access and integration: Projecting sound onto meaning." Lexical Representation and Process. Marslen-Wilson ed. 1989 MIT Press. Cambridge MA.

[14] MERMELSTEIN, P. On the relationship between vowel and consonant identification when cued by the same acoustic information. Percep. and Psychophys. **23**: 331-335, 1978.

[15] NEAREY, T. The segment as a unit of speech perception. J. Phonetics. **18**: 347-373, 1990.

[16] NEAREY, T. & HOGAN, J. "Phonological contrast in experimental phonetics: relating distributions of measurements production data to perceptual categorization curves." Experimental Phonology. Ohala & Jaeger ed. 1986 Academic Press. New York.

[17] NEAREY T. &. SHAMMASS, S. Formant transitions as partly distinctive invariant properties in the identification of voiced stops. Canadian Acoust. **15**: 17-24, 1987.

[18] NEAREY, T. M. On the physical interpretation of vowel quality: cinefluorographic and acoustic evidence. J. Phonetics. **8**:213-241, 1980.

[19] OHALA, J. "What's cognitive, what's not." Morrissey ed. in press Peter Lang Verlag. In press.

[20] SAMUEL, A. & RESSLER, W. Attention within auditory word perception: insights from the phonemic restoration illusion. Journal of Experimental Psychology: Human Perception and Performance. **12**(1): 70-79, 1986.

[21]WHALEN, D. Vowel and consonant judgments are not independent when cued by the same information. Percep. and Psychophys. **46**(3): 284-292, 1989.

Perceptual Processing and Ecological Validity

Willy Serniclaes

Brussels' Free University

ABSTRACT

According to Nearey [1], direct effects of stimulus properties would be limited to the lowest level stage of speech perception. This raises the question to know whether the experimental stimuli display the relevant properties for being processed at higher stages of perceptual processing. The interest of ecological validity for explaining syllabic and sentence context effects is illustrated.

INTRODUCTION

In the keynote address by T.M. Nearey [1], two main topics emerge. The first is the proposal to leave aside the controversy between motor and auditory theories of speech perception in favour of a better understanding of the relationship between the two domains, via language-specific regularities in the speech wave. The second topic concerns the broad version of the "segmental filter" model, which states that the direct effects of stimulus properties are limited to the low-level stage of phone-sized segments identification and do not interfere with lexical factors in the course of top-down processing. Although I am in complete agreement with Nearey's paradigm, the strong version of the segmental filter model is, in my opinion, quite premature.

Much of the work in the field of speech communication is devoted either to the evaluation of perceptual models or to the collection of acoustic data for testing hypotheses on speech production, whereas little has been done for relating the distribution of acoustic cues in the speech wave to their perceptual processing. Yet, distributional characteristics have important perceptual implications, not only for explaining cross-linguistic differences, but also for understanding the specificity of the perceptual use of speech cues. In the ecological view developed by Brunswick [2], the description of semiregularities of the environment, or "ecological validities", is a prerequisite for testing the adaptative response of organism. Before concluding that acoustic

properties cannot affect some stage of processing, we should have the insurance that the experimental stimuli display the relevant properties for being processed at that stage. In this framework, the hypothesis which states that stimulus properties do not directly affect the syllabic stage of processing can only be accepted insofar the cues under study indeed provide relevant information for syllable-sized units. In the same way, it is only when the relevant information for being processed at the lexical level is present in the stimuli that the absence of interference between sentence context and sensory input can provide an argument in favor of a modular approach of speech comprehension [3].

SYLLABIC CONTEXT EFFECTS

One should be very cautious in interpreting the failure of the stimulus-tuned diphone model to improve the fit of Whalen's data [1, 4]. While log-linear analyses clearly suggest that the acoustic cues under study do not directly affect the syllabic stage of processing of isolated CV stimuli, an important question is to know whether the cues really provide relevant information to feed the syllabic decoder. Let us first look at the acoustic cues in Wahlen's first experiment. The fact that duration and F1 frequency each depend on the identity of the vowel and of the following consonant provides a clear motivation for their use as perceptual cues both for vowel and consonant distinctions. But this does not necessarily mean that a further analysis of these cues for the classification into syllabic units is also motivated. This depends on whether the cues also convey some specific information on the syllable, irreductible to the one they give on vowels and consonants. In quantitative terms, a stimulus-based syllabic decoding is required only insofar vowel and consonant have interactive effects on the production of the acoustic cues.
In the measurements of vocalic durations in CVC syllables presented by Peterson and Lehiste ([5]; Tab.II, p.702), 28 values corresponding to 7 vowel duration and voicing contrasts (such as bead-beat-bit-bid) have been taken here for testing the additivity of vowel and consonant effects. For the logarithm of the durations, which is more appropriate for perceptual modelling, the vowel and consonant effects are largely significant (vowel: $F_{1,24} = 46.86$; $P < .001$; consonant: $F_{1,24} = 50.25$; $P < .001$) whereas the interaction is by far below the theshold of significance ($F_{1,24} = .962$; $P = .336$). The additivity of vowel and consonant effects is also apparent in the vocalic duration measurements in connected speech reported by Crystal and House ([6], Tab.X, p.1560). The magnitude of the effect of consonantal voicing on vocalic duration is reported here in Tab.1 for vowels preceding prepausal word-final stops, the only

not the case. The ratio between F2 frequencies , taken at 10 different intervals during the vocalic segment, can be derived from Soli's data (Figs. 3, 4, 7 8) and are presented here in Tab.2.

Table 2. Ratio between F2 frequencies (adapted from [9])

Voicing included (n=60*10)	
si/su	1.48
chi/chu	1.55
zi/zu	1.37
ji/ju	1.31
Voicing excluded (n=120*10)	
si, zi/ su, zu	1.43
chi, ji/ chu, ju	1.43

As can be seen, the magnitude of the vowel effect (i versus u) does not change systematically as a function of consonant place of articulation (s versus ch or z versus j). Although variances are not available, this suggests that vowel and consonant do not have interactive effects on F2 frequency and hence that the magnitude of the vowel effect on F2 does not depend on the consonant, and vice-versa. Just like the three other cues used by Nearey for testing the relevance of stimulus information at the syllabic stage of processing, F2 frequency, as a joint cue for vowel and fricative identity, does not convey adequate information for being processed at this stage. Taking account of the absence of ecological validation, the fact that these cues do not interfere with syllabic identification in the course of perceptual processing [1] does not allow to conclude that stimulus properties in general cannot reach the syllabic level. Other cues, if any, might exhibit an interactive relationship with two or more segments and could then provide the adequate stimuli for testing the segmental filter model.

SENTENCE CONTEXT EFFECTS

The phonetic structure of the stimulus also has implications for the general debate on the modularity of speech processing stages. Non-acoustic top-down processes are known to aid and bias speech perception. The phoneme restauration effect [10] shows that context can control the perception of phonemes. Many researchers have examined the influence of some kind of non-acoustic information on phonetic categorisation. Sentence context can bias phonetic categorisation but only when the available phonetic information is ambiguous [11, 12, 13]. The effect of lexical information on speech perception has also been investigated. It has been demonstrated that ambiguous stop consonants tend to be perceived so that the whole stimulus is a meaningful word than a non-word [14, 15]. The problem, however, is whether lexical information either biases phonetic categorisation or directly affects the mechanisms of cue integration before feature categorisation. Conine and Clifton [16] showed that prestored lexical information may be used directly in perception, contrary to semantic

position for which the effect is clearly present.

Table 1. Ratio between mean vocalic durations before voiced or voiceless consonants for long and short vowel categories (adapted from Crystal & House [6] Tab.X p.1560). N indicates the number of tokens for voiced and voiceless categories.

Vowels:	Long	Short
Obstruents	1.23	1.16
N	23 & 42	33 & 48
Stops	1.21	1.22
N	18 & 33	27 & 24
Fricatives	1.24	0.73
N	5 & 9	3 & 2

The increment of vocalic duration before voiced consonants is fairly stable, around 20 %, at the exception of short vowels before fricatives, for which the effect is reversed, probably as a consequence of the reduced sample size.

Given the absence of interaction between the vowel and consonant effects on the production of vocalic duration, the information conveyed by this cue can be entirely extracted at the low-segmental level and does therefore not require a further analysis at the syllabic level. Vocalic duration is thus not a good candidate for testing the relevance of acoustic information at a syllabic stage of processing. The other cue under study in Whalen's first experiment is the rate of F1 transition which covaries with the F1 stable frequency. Given the lack of acoustic data, we do not know whether the effects of vowel and consonant on these cues are additive or not. Notice however that vowel and consonant identification do not depend on the same aspect of F1 contour. As far as I know, voicing identification only depends on F1 transition rate (or at least on F1 initial frequency: [7, 8]) whereas vowel identification of course depends on F1 characteristics but not specifically on transition rate. F1 frequency does thus not provide the same cue for vowel and consonant and is therefore also not a good candidate for testing the relevance of acoustic information at the syllabic stage of processing. Finally, in the second example taken by Nearey and which deals with the effects of fricative pole and F2 frequency on the si-su-chi-chu distinctions (Whalen's [4] third experiment), only the latter cue clearly has a significant effect on both vowel and consonant identification. The fact that F2 frequency does not affect the syllabic stage of perceptual processing could therefore again suggest that stimulus information cannot reach this level. The question, again, is to know whether this cue conveys relevant information for being processed in a syllabic frame. The acoustic measurements of F2 in English initial fricatives presented by Soli [9] seem to indicate that this is context which is used postperceptually. No firm conclusion can however be drawn in view of the versatility of the lexical effect on phonetic categorisation. Among the factors which contribute to the effect, stimulus structure has a special interest for the present discussion. The relevance of the phonetic structure of the stimulus is evidenced by the the fact that the lexical effect disappears when stimulus variations more nearly approximate the multiple acoustic differences between phonetic categories in natural speech [17]. This does not mean that the lexical effect is a laboratory phenomenon. Strong sentential effects have indeed be obtained for voicing perception in excerpts from spontaneous French speech [18]. This requires some further explanations. Natural voiceless French stops usually exhibit a silent interval at the vicinity of closure release whereas voiced stops almost always display continuous periodic vibrations. As other voicing cues are by far less reliable, the presence vs. absence of silence provides a major perceptual cue [19]. However, in spontaneous speech, voiceless stops may exhibit continuous periodic vibrations and are then identified as voiced when excerpted from the sentence, although they are identified as voiceless in the sentence frame. This effect, as evidenced by other aspects of the data, is clearly due to semantic factors and shows that the influence of top-down processing is not restricted to ambiguous stimuli. Even when the major acoustic cue is completely non-ambiguous, sentential context can completely modify the phonetic decision. The question which is raised is whether such a large top-down effect arises from a bias in phonetic categorisation or, more conceivably, from a direct effect of lexical information into the process of acoustic cue integration.

More importantly for my purpose here, is the fact that the magnitude of the lexical effect is seemingly due to the presence of conflicting cues in the stimulus, the major cue supporting a voiced percept whereas secondary cues support a voiceless percept. Such an internal conflict within the acoustic correlate could generate a strong appeal for extraneous evidence from lexical processes, which would in turn explain the exceptional magnitude of the lexical effect. Ecological validity, which in this case depends on the presence of contradictions within the acoustic correlate, would then again have decisive importance for testing perceptual models.

REFERENCES

[1] NEAREY, T.M. (1991) "Perception: Automatic and Cognitive Processes" Keynote address in this volume.
[2] BRUNSWICK,E. (1957) Scope and aspects of the cognitive problem. In Contemporary Approaches to

Cognition. Cambridge, MA: Harvard U. Press, 5-40.
[3] FODOR, J.A. (1983) "The modularity of mind" Cambridge, MA: MIT Press.
[4] WHALEN, D. (1989) "Vowel and consonant judgments are not independent when cued by the same information" Percep. and Psychophys., 46, 284-292.
[5] PETERSON,G.E. and LEHISTE,I. (1960) "Duration of syllable nuclei in English" J.Acoust.Soc.Am. 32, 693-703.
[6] CRYSTAL,T.H. and HOUSE,A.S. (1988) "The duration of American English vowels: an overview" J. of Phonetics, 16, 263-284.
[7] LISKER,L. (1975) "Is it VOT or a first formant transition detector?" J.Acoust.Soc.Am. 57, 1547-1551.
[8] SUMMERFIELD,Q., and HAGGARD,M. (1977) "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants" J.Acoust.Soc.Am. 62, 435-448.
[9] SOLI,S.D. (1981) "Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation" J.Acoust.Soc.Am., 70, 976-984.
[10] WARREN,R.M. (1970) "Perceptual restauration of missing speech sounds" Science 167, 392-393.
[11] GARNES,S. and BOND,Z.S. (1976) "The relationship between semantic expectation and acoustic information" Phonologica, 3, 285-293.
[12] MILLER,J.L., GREEN,K., and SCHERMER,T. (1984) "A distinction between effects of sentential speaking rate and semantic congruity on word identification" Percept. and Psychophys., 36, 329-337.
[13] CONNINE,C. (1987) "Constraints on interactive processes in auditory word recognition: The role of sentence context" J. of Memory and Language, 26, 527-538.
[16] GANONG, W.F. (1980) "Phonetic categorization in auditory word perception" J. of Exp. Psychology: Human Perc. & Perform., 6, 110-125.
[15] FOX,R.A. (1984) "Effect of lexical status on phonetic categorization" J. of Exp. Psychology: Human Perc. & Perform., 10, 526-540.
[16] CONNINE,C., and CLIFTON,C.(1987) "Interactive use of lexical information in speech perception" J. of Exp. Psychology: Human Perc. & Perform., 13, 291-299.
[17] BURTON,M.W., BAUM,S.R., and BLUMSTEIN,S.E. (1989) "Lexical effects on the phonetic categorization of speech: The role of acoustic structure" J. of Exp. Psychology: Human Perc. & Perform., 15, 567-575.
[18] SAERENS, M., SERNICLAES, W., and BEECKMANS, R. (1989). "Acoustic versus contextual factors in stop voicing perception in spontaneous speech" Language and Speech 32, 1989, 291-314.
[19] SERNICLAES, W. (1987) Etude expérimentale de la perception du trait de voisement des occlusives du français. Thèse de doctorat en Sciences psychologiques, ULB, Octobre 1987.

A MORE INTIMATE COCKTAIL PARTY PHENOMENON: THE PERCEPTUAL ORGANIZATION OF TONAL ANALOGS OF SPEECH

Robert E. Remez

Barnard College, New York, New York U.S.A.

Perceptual organization of auditory patterns is often explained by appeal to Gestalt grouping principles. Despite the evidence for such principles in the grouping of simple acoustic displays, the time-varying nature of speech spectra eludes a similar account. Studies with sinewave analogs of speech show that auditory grouping is not required for phonetic perception, nor are principles of the Gestalt variety sufficient to explain the perceptual integration of oral, nasal, and fricative formants. We have identified a grouping principle keyed to speechlike spectral change, in tests using dichotic sinusoidal components which are grouped both in phonetic and auditory modes. Our findings warrant extending the simple characterizations available within the framework of the Gestalt laws.

1. PRINCIPLES OF PERCEPTUAL ORGANIZATION

In 1923, Max Wertheimer [16] reported the results of an inquiry into perceptual organization. His quest was to show that perception of an ambiguous display was organized, and not a simple summary of the elements of stimulation. Our recent concern has been the validity of these Gestalt principles in accounting for the perceptual organization of speech, and is occasioned by the fact that Wertheimer's principles are still very much with us in auditory form. They are often cited as a preliminary step in auditory recognition of objects and events. Supporting evidence comes from many studies of arbitrary acoustic displays, though we have recently submitted the principles to test using speech signals, or replicas presenting spectro-temporal attributes common to speech. These new studies are not encouraging about the descriptive or theoretical adequacy of the Gestalt account when it is applied to the case of speech.

Wertheimer exposed the principles using ambiguous plane shapes and brief tone sequences, deriving a collection of perceptual devices for grouping the figural elements of stimulation: proximity, similarity, common fate, set, continuity, symmetry, closure, and habit. Explicitly auditory instances of many organizational principles were again offered by Julesz & Hirsch [8] who sought common principles for perception in visual and auditory modalities. Although they concluded that the dissimilarities of vision and hearing outweighed the shared attributes, their review brought an influential information-processing rationale to subsequent studies. Julesz & Hirsch themselves contended that the Gestalt organizational principles alone might prove inadequate to explain the perceptual integrity of the speech signal, due both to its acoustic complexity, and to the contribution to perception of the listener's extensive knowledge of speech and language. In the 25 years since this article appeared, a large body of evidence has been gathered about its numerous hypotheses. These studies support the detailed claims that an auditory scene is organized or analyzed perceptually according to principles of proximity [2], similarity [3, 15, 6], common fate [1, 4] and closure [10], operating in the domains of frequency, amplitude, and spectrum. The application of the grouping principles is held to promote the formation of separate auditory streams, which, once sorted, are passed along for more detailed perceptual analysis about the objects and events which gave rise to the stimulation.

2. PERCEPTUAL ORGANIZATION OF SPEECH SIGNALS

Our question is simple: Do the diverse components of a single speech signal cohere perceptually through the application of Gestalt grouping principles? Pertaining to speech, this question is typically framed about the isolation of a single voice against an acoustic background of other talkers, clinking glasses, popping champagne corks, and whirring air conditioning systems; in short, the familiar "cocktail party phenomenon" [5]. But our present focus on perceptual organization requires a more intimate setting. After all, the listener must perceive that a single speech signal produced in a quiet room is the product of a single vocal source of sound. Does streaming account for that?

Current formulations of auditory perceptual organization warrant the fracture of a speech signal into perceptually incoherent streams, rather than accounting for the fusion of the diverse acoustic components into the single ongoing perceptual event which the listener hears. This outcome is due to the reliance of the grouping principles on durable similarities or coordinate changes occurring among the elements of the incident acoustic pattern. This is the clear and unavoidable consequence of grouping acoustic elements by physical similarity, physical continuity and common, coordinate transient characteristics. The fracture of speech into incoherent streams then follows from the acoustic nature of the speech signal, which is ordinarily replete with failures of similarity, continuity and common fate. These familiar acoustic attributes are observed when the frequency changes of one formant center do not match frequency changes of another formant in direction, degree, or duration; when onsets and offsets fail to occur in synchrony; and when episodes of nasal and fricative formants occur, lacking frequency continuity and spectral similarity with the oral resonances. None of these types of lapse is particularly exotic, as any spectrogram will reveal [7].

Despite this acoustically diverse collection of elements, the listener's perception is typically of a single stream of consonants and vowels, and not of disjoint simultaneous streams, each comprising a single kind of auditory element. Disintegrated impressions can occur when a brief snippet of a speech signal is presented in a rapidly repeating train [9]. But, the specific conditions required to elicit such impressions serve to underscore the difference between the perception of speech and the segregation of auditory components through streaming.

Were we to suspect that the common vocal excitation in the formants ordinarily holds them together perceptually—a kind of common fate—we would nonetheless have a hard time explaining phonetic perception of tonal analogs of speech. Here, the co-modulation of formant centers is eliminated by the use of digital synthesis to compose a collection of linear emitters which convey the momentary acoustic maxima. The familiar timbres of consonants and vowels are not evoked by such resonance-free and grossly unnatural short-term spectra [14], yet phonetic perception occurs nonetheless. The listener's ability to transcribe these odd replicas of speech depends on the perceptual disposition to treat three- and four- tone analogs as coherent despite their violation of grouping principles and unfamiliar timbre [11, 12, 13].

3. TESTS WITH SINEWAVE REPLICAS OF SPEECH

In tests to determine the kind of organization occurring in tone analogs of speech we have tried to distinguish perceptual organization, in which simultaneous dissimilar components are actually integrated, from a low-level peripheral fusion of sinusoids, due, perhaps, to auditory coupling of the first and second formant tones. The latter possibility is a likely mechanical consequence of some models of basilar function [17], and if true eliminates much of the interest in the case of sinewave analogs.

To determine the likelihood that sinusoidal components in a tonal analog are organized due to peripheral interactions at transduction, our tests used dichotic presentation requiring the listener to integrate a single tonal component presented to one ear—corresponding to the second formant—with the remainder of the replica presented to the other ear. Were transcription to deteriorate in this dichotic presentation, relative to the binaural case, we would conclude that (i) perceptual organization is a trivial consequence of auditory transmission of tonal components; and, (ii) disjunctive azimuth precludes active perceptual organization. Were transmission to survive dichotic presentation of essential acoustic ingredients, we would conclude that (iii) organization is not attributable to passive conduction of tonal components; and, (iv) failures in similarity, continuity, common fate—not to forget azimuth—are insufficient to prevent phonetic organization from occurring.

Our first test compared binaural and dichotic presentation of sinewave replicas. The second test assessed the selective power of phonetic organization by requiring the listener to combine the appropriate dichotically presented components despite the presence of a competing speechlike tone.

The acoustic materials which we used in these tests were sinusoidal replicas of utterances of sentences. In the basic dichotic test, one ear received the tones corresponding to the first, third and fourth formant centers, while the other ear received only the tone corresponding to the second formant center. In one control, the dichotic condition was compared with the binaural presentation of the full tone set. In another, the intelligibility of the patterns presented to each ear in the dichotic case was assessed in binaural tests; one test examined the information available from Tones 1, 3, and 4, lacking the second formant tone; the other assessed the phonetic effects of of Tone 2 alone.

Second, we ran a test of organization with interfering acoustic material, to identify the acoustic criteria of the organizational principles. In this test, the phonetically coherent tones were presented dichotically, along with a foil in the ear opposite Tone 2. This distractor tone either exhibited speechlike variation or constant frequency, though neither distractor satisfied the acoustic criteria for grouping with other concurrent tones. Although neither distractor was phonetically coherent, we expected only the tone with speechlike properties to compete organizationally with the true second formant tone.

4. THE FINDINGS

Can listeners integrate tonal components presented dichotically? The first test compared transcription accuracy for ten sentences in four conditions, and Figure 1 portrays the outcomes. Integration occurred despite violations of Gestalt grouping principles, for the dichotic performance surpassed the combination of each ear's contribution from a partial signal. Note, also that there is a clear performance decrement with dichotic organization relative to the binaural case, perhaps reflecting attentional load differences.

Our question in the second test series derived from the first: Is phonetic perception driven by an organizational principle that works by acoustic similarity? If so, then listeners should be indifferent to the presence of a tone

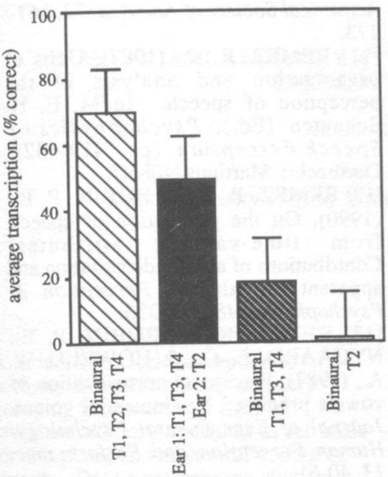


Figure 1. Group results for the perceptual organization test.

that is both dissimilar to the acoustic components of a speech signal and incoherent, in the sense that it could not have issued from a the same vocal source as the other tones in the presentation. We tested this with a condition requiring perceivers to reject a temporally flipped second formant tone in the pattern at one ear and to integrate the dichotically available veridical second formant tone. Neither second formant tone was similar in the Gestalt sense to the acoustic ensemble of Tone 1, Tone 3, and Tone 4, but only one was coherent in that it belonged to the tonal replica. This task of rejecting a second formant tone that had appropriate azimuth and speechlike time-varying frequencies—but which nevertheless was inappropriate for the rest of the tonal ensemble—and integrating the appropriate second formant tone presented in the other ear proved to be quite difficult for our listeners. However, when the dichotic competing tone lacked the spectro-temporal attributes of speech, exhibiting constant frequency, listeners easily rejected it, as if it were not competitive at all. Figure 2 shows performance in this test. The combined results suggest that listeners are vigilant in listening for plausible speechlike components, and are therefore misled by natural frequency variation and azimuth of the phonetically incompatible tone.

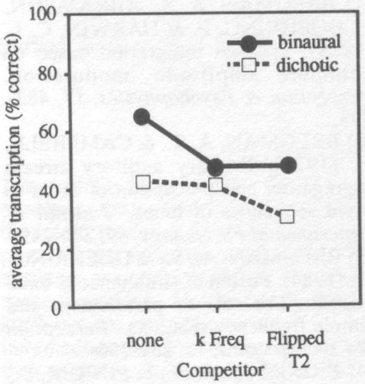


Figure 2. Group results for the test of competitive organization. Listeners heard either no distractor, or one of two possible distractors: a constant frequency tone, or a temporally reflected formant tone.

5. CONCLUSION

To summarize the outcome of our tests, the perceptual organization of speech does not rely necessarily on the acoustic properties featured in contemporary—or historical—accounts of grouping. Listeners were quite able to integrate a pattern of tones lacking similarity, continuity and common fate in the acoustic spectrum. No resort to auditory familiarity is available for explaining this finding, given the highly unnatural timbre of tonal analogs of speech. Moreover, the integrating mechanism seems keyed to speechlike signal variation in the simultaneous component tones, for listeners were less able to reject a speechlike tone that shared azimuth with the remainder of the tone ensemble in favor of a phonetically appropriate tone presented with inappropriate azimuth. Altogether, then, it seems that the complex spectrum of speech places it beyond the reach of the simple, elegant Gestalt rules. Our search leads now toward the underlying physical and linguistic principles of perceptual organization of speech, and the perceiver's ability to exploit them in understanding an utterance.

6. REFERENCES

[1] BREGMAN, A. S., ABRAMSON, J., DOEHRING, P. & DARWIN, C. J., (1985), Spectral integration based on common amplitude modulation. *Perception & Psychophysics*, *37*, 483-493.

[2] BREGMAN, A. S., & CAMPBELL, J., (1971), Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*, 244-249.

[3] BREGMAN, A., S., & DOEHRING, P., (1984), Fusion of simultaneous tonal glides: The role of parallelness and simple frequency relations. *Perception & Psychophysics*, *36*, 251-256.

[4] BREGMAN, A. S., & PINKER, S., (1978), Auditory streaming and the building of timbre. *Canadian Journal of Psychology*, *32*, 19-31.

[5] CHERRY, E. C., (1953), Some experiments on the recognition of speech with one and two ears. *Journal of the Acoustical Society of America*, *25*, 975-979.

[6] DANNENBRING, G. L., & BREGMAN, A. S., (1978), Streaming vs. fusion of sinusoidal components of complex tones. *Perception & Psychophysics*, *24*, 369-376.

[7] FANT, C. G. M., (1962), Descriptive analysis of the acoustic aspects of speech. *Logos*, *5*, 3-17.

[8] JULESZ, B., & HIRSCH, I. J., (1972), Visual and auditory perception: An essay of comparison. In E. E. David and P. B. Denes (Eds.), *Human communication: A unified view* (pp. 283-340), New York: McGraw-Hill.

[9] LACKNER, J. R., & GOLDSTEIN, L. M., (1974), Primary auditory stream segregation of repeated consonant-vowel sequences. *Journal of the Acoustical Society of America*, *56*, 1651-1652.

[10] MILLER, G. A. & LICKLIDER, J. C. R., (1950), The intelligibility of interrupted speech. *Journal of the Acoustical Society of America*, *22*, 167-173.

[11] REMEZ, R. E., (1987), Units of organization and analysis in the perception of speech. In M. E. H. Schouten (Ed.), *Psychophysics of Speech Perception* (pp. 419-432). Dordrecht: Martinus Nijhoff.

[12] REMEZ, R. E., & RUBIN, P. E., (1990), On the perception of speech from time-varying attributes: Contributions of amplitude variation and apparent naturalness. *Perception & Psychophysics*, *48*, 313-325.

[13] REMEZ, R. E., RUBIN, P. E., NYGAARD, L. C., & HOWELL, W. A., (1987), Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 40-61.

[14] REMEZ, R. E., RUBIN, P. E., PISONI, D. B., & CARRELL T. D., (1981), Speech perception without traditional speech cues. *Science*, *212*, 947-950.

[15] STEIGER, H., & BREGMAN, A. S., (1982), Competition among auditory streaming, dichotic fusion, and diotic fusion. *Perception & Psychophysics*, *32*, 153-162.

[16] WERTHEIMER, M., (1923), Untersuchungen zur Lehre von der Gestalt, II. *Psycholigische Forschung*, *41*, 301-350.

[17] WILSON, J. P., & JOHNSTONE, J. R., (1975), Basilar membrane and middle-ear vibration in guinea pig measured by capacitative probe. *Journal of the Acoustical Society of America*, *57*, 705-723.

The author gratefully acknowledges the assistance of S. M. Berns, J. M. Lang, J. S. Nutter and P. E. Rubin. This research was supported by a grant from the NIDCD (DC00308).

INVARIANT AUDITORY ATTRIBUTES AND A MODEL OF SPEECH PERCEPTION

James R. Sawusch

State University of New York at Buffalo, Buffalo, New York

ABSTRACT

A model of speech perception is outlined that incorporates auditory grouping processes and a set of invariant auditory attributes as the basis for the phonetic and lexical coding of speech. Data from priming studies indicates that the phonetic code for words includes the positional specification of each phoneme. Based on this and some considerations of how invariant attributes might support auditory event and speech perception, a set of invariant auditory attributes for perception is described. This set is supported by perceptual data and, in conjunction with the model, accounts for a number of phenomena (such as trading relations) in the speech literature.

1. INTRODUCTION

In any model of speech perception, there are three key elements that need to be addressed. The model must detail the nature of the auditory attributes or invariants that support perception. The nature of perceptual grouping or organization processes that bind cues from the acoustic signal together as a single entity and separate the acoustic components of one event from other events needs to be specified. The mapping process that converts auditory cues to words must also be detailed. In the sections that follow, a model designed to address these issues will be outlined. Once the model is outlined, some details of the auditory invariants that underlie speech and other auditory perception will be considered.

The model is set in an information processing framework [cf., 12]. Like the LAFS model of Klatt [6], it proposes that invariant auditory attributes are captured from a series of spectral sections. Unlike LAFS, these attributes are mapped onto an intermediate phonetic representation and then to words. Like the proposals of Fowler [3], the ultimate goal of perception is recognition of the object or event that produced the sound. The set of auditory attributes is designed to support the recognition of speech and nonspeech. Unlike Fowler's proposal, in this model perception is mediated by stages of processing. The model does contain a "speech mode" [cf., 7]. However, this mode is layered over a set of auditory coding processes [see 15].

2. MODEL OUTLINE

The model consists of a sequence of representations and transformations. After the transformation of sound into an internal spectral/temporal representation by the peripheral auditory system, a basic set of auditory features is extracted. These features include the amplitude envelope, periodic or aperiodic nature of the waveform, fundamental frequency, and the frequencies and amplitudes of peaks in the spectrum. As these features are extracted over time, they are grouped together in sets that represent a common source or event [cf., 1]. The grouping process at this stage is driven by local spectral and temporal information. All further processing is then performed within a set of features.

The next step in processing represents the extraction of a set of invariant auditory attributes from the local cues. This set of attributes constitutes the basic information that drives acoustic object or event recognition. That is, regardless of whether the sound source is music, bird calls, door slams or speech, the attributes capture the information that preserves object or event identity. One example of such invariant attributes is the set of frequency differences between adjacent spectral peaks at each point in time. The nature of this set of invariants will be elaborated below.

The third stage in the model maps the auditory attributes, over time, onto objects or events. In speech these events are phonetic features, phonemes, syllables and words. Since all of these units seem to have a role in perception [see 11, 17], the model should explain the role of each abstract speech unit. The model should also provide an account of adult perception that is compatible with data and theories of infant speech perception and the acquisition, by children, of the words of their language [5].

The model assumes that in an adult the process of transforming auditory attributes into words is a two step process. The first step is a mapping of attributes onto phonetic features while the second step maps features, over time, onto words. The mapping of attributes onto phonetic features might be accomplished within a connectionist architecture. In this case, a time delay net could be used to map sequences of invariant attributes onto phonetic features. Interactions within the net would reproduce many of the phonetic trading relations described in the literature [13]. In addition, information about the rhythm of speech as cued by acoustic/auditory features such as the rise and fall of the amplitude envelope would govern the scope of cue interactions in the net. That is, a syllable-like integration window or context would result from the use of the amplitude envelope as a perceptual grouping factor at this stage of processing.

Finally, a sequence of features, over time, is mapped onto the words of the listener's lexicon. Since the phonetic features contain positional information and were grouped according to the rhythmic structure of speech, potential word boundaries are marked in the phonetic feature information for use by the word recognition process [cf., 9].

3. PHONETIC CODING

At this point, some further elaboration of the phonetic feature representation is in order since it will influence the nature of the set of invariant attributes as well as how they are mapped onto phonetic units. The phonetic features here are not the abstract entities used in linguistic theories. Rather, these features are position specific and contain some allophonic detail. Based upon data from auditory priming experiments using lexical decision and naming tasks, Gagnon and Sawusch [4] proposed that the phonetic representation used in word recognition includes information about the syllable position of each phonetic element. Thus, a syllable initial /b/ and a syllable final /b/ would be coded as two separate entities. In the present model, each phonetic feature would include a positional specification within the syllable. For example, the stop manner feature would be coded as either an onset stop or an offset stop.

The implication of this for the extraction of auditory attributes is that no single invariant or set of invariants is used by humans in the perception of all variations of a phoneme or feature. Rather, the invariant attributes themselves are part of the cue to the position of the feature in the syllable. Thus, the focus in our search for invariant attributes has been to examine the acoustics of stops before the vowel in a syllable and stops following the vowel separately. No common invariant across position has yet been found which is consistent with perceptual data [see 4].

4. INVARIANT ATTRIBUTES

The set of invariant auditory attributes must meet a number of constraints. First, they must represent a set of acoustic attributes that are not specific to speech. Rather, this set should be capable of supporting all auditory object or event recognition, including speech and music. This does not imply that speech perception, music recognition and the classification of a sound as a door slamming shut are all variants of the same perceptual processes. The process of mapping invariant attributes onto phonetic features described previously is a speech specific coding process. If implemented in a connectionist network, the weights on the connections would be the result of learning and represent a "speech mode" of processing. A similar learning process would be involved in the perception and recognition of other sounds.

A second requirement for this set is that they meet the requirements described by Sawusch and Dutton [16] for a formal, computational model or metric for perception. The attributes must represent a robust set in which the information supporting phonetic coding is preserved in spite of variation in talker, talking rate and the speech context. The attributes should also support perception even when degraded and should not lead to a sudden failure of perception in a noisy environment. The attributes should support graceful degradation. Finally, the attributes must be formally specified or computable in a manner that does not require intelligent guidance.

To illustrate the nature of invariants for sound recognition, consider the phonetic dimension of place of articulation for consonants and vowels. Miller [8] and Syrdal and Gopal [18] have proposed that the frequency differences between adjacent peaks in the spectrum capture the essential properties necessary for vowel recognition. Forrest, Weismer, Milenkovic, and Dougall [2] proposed that the statistical moments of the spectrum (mean, variance, skewness and kurtosis) capture a sufficient set of qualities for perception of the voiceless fricatives and stops. Sawusch and Dutton [in press] examined both of these alternatives for voiced stops and vowels. They found that the peak difference metric did not degrade gracefully while the statistical moments metric was not as robust as desired. However, these failures were largely complimentary so that a hybrid of both proposals might be sufficient to capture human perceptual capabilities.

In the present model, both the statistical moments of the short term spectrum and the frequency differences between adjacent peaks would be computed. In addition, the amplitude differences between adjacent peaks in the spectrum are also computed. The rate of zero crossings, the rate of change of overall amplitude and a set of source attributes such as the degree of periodicity in the spectrum and the fundamental frequency are also a part of the information represented here. These properties would be computed for each temporal section of the waveform on a continuous or running basis. At this point the representation of sound is still continuous and has not yet been segmented.

The statistical moments is a generalized description of the spectrum that subsumes the spectral tilt cue that has been proposed as an invariant to stop place. In a study of the efficacy of changes in spectral tilt as a cue to stop place, Richardson and Sawusch [14] found that changes in spectral tilt did not predict human listener classification of synthetic syllables. In a subsequent analysis of these syllables, it was found that both changes in the frequency differences between spectral peaks and changes in the statistical moments do predict listener's classification responses. Further tests of the efficacy of the statistical moments and peak differences as cues to stop consonant place are now in progress.

Certain properties are not represented at this level. The duration of an acoustic segment or the duration of a change in one of the attributes listed above, such as the rise-time, are not directly represented in the set of attributes. Instead, these are recovered by the process of mapping the static attributes, over time, onto phonetic features. There is also no single invariant for voice onset time (VOT). Rather, VOT is a composite that represents the mapping of source attributes, the amplitude differences between adjacent peaks and other attributes onto a position specific voicing feature. The processing of a nonspeech analog to VOT such as tone onset time [10] produces classification and discrimination results similar to speech because the nonspeech stimuli contain attributes in common with speech which are mapped (in the context of an experiment) onto categories correlated with speech categories.

Conversely, nonspeech experiments and speech experiments with the same stimuli sometimes produce differences in category boundary placements. This reflects a speech mode with connections from attributes to phonetic features that include learned contextual dependencies. In the context of a single experiment, no such detailed learning would take place for a nonspeech distinction. Consequently, comparisons between speech and nonspeech perception reveal both commonalities in processing based on a single set of auditory attributes and subtle differences due to the different learning histories between speech and nonspeech groups.

5. CONCLUSION

The outline of a model and elaboration of some aspects of the auditory coding of speech here is the beginning of a process of building a theory of speech perception. Among the issues that remain to be addressed are how phonotactic constraints and phonological knowledge should be incorporated into the model [see 19]. A further elaboration of the acoustic cues that guide perceptual grouping at each stage of the model is also needed. Simulations using a connectionist network to model the mapping of auditory attributes to phonetic features proposed here are also under way to test the sufficiency of the set of attributes. Finally, we are running a series of experiments in which the acoustic properties of synthetic stimuli are varied to examine how the attributes described here are extracted in perception and assess their relative roles in speech recognition. These data, together with theoretical elaborations, should lead to a more complete specification of the auditory to phonetic coding of speech.

6. REFERENCES

[1] BREGMAN, A. S. (1990), "*Auditory scene analysis*", Cambridge, MA: MIT Press.

[2] FORREST, K., WEISMER, G., MILENKOVIC, P., & DOUGALL, R. N. (1988), "Statistical analysis of word-initial voiceless obstruents. Preliminary data", *Journal of the Acoustical Society of America, 84*, 115-123.

[3] FOWLER, C. A. (1990), "Sound-producing sources as objects of perception: Rate normalization and nonspeech perception", *Journal of the Acoustical Society of America, 88*(3), 1236-1249.

[4] GAGNON, D. A., & SAWUSCH, J. R. (1990), "Priming and perceptual units in auditory word recognition", Paper presented at the 31st meeting of the Psychonomic Society, New Orleans, LA.

[5] JUSCZYK, P. W. (in press), "Infant speech perception and the development of the mental lexicon", In H. C. Nusbaum, & J. C. Goodman (Eds.), *The transition from speech sounds to spoken words: The development of speech perception.* Cambridge, MA: MIT Press.

[6] KLATT, D. H. (1989), "Review of selected models of speech perception", In W. Marslen-Wilson (Ed.), *Lexical representation and process*. Cambridge, MA: MIT Press.

[7] LIBERMAN, A. M., & MATTINGLY, I. G. (1985), "The motor theory of speech perception revised", *Cognition, 21*(1), 1-36.

[8] MILLER, J. D. (1989), "Auditory-perceptual interpretation of the vowel", *Journal of the Acoustical Society of America, 85*(5), 2114-2134.

[9] NAKATANI, L. H. & DUKES, K. D. (1977), "Locus of segmental cues for word juncture", *Journal of the Acoustical Society of America, 62*(3), 714-719.
[10] PISONI, D. B. (1977), "Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops", *Journal of the Acoustical Society of America, 61*, 1352-1361.
[11] PISONI, D. B. (1982), "In defense of segmental representations in speech processing", In A. S. House (Ed.), *Project SCAMP 1981: Acoustic phonetics and speech modeling*. Princeton, NJ: Institute for Defense Analysis, Communications Research Division.
[12] PISONI, D. B., & SAWUSCH, J. R. (1975), "Some stages of processing in speech perception", In A. Cohen, & S. G. Nooteboom (Eds.), *Structure and process in speech perception*. New York: Springer-Verlag.
[13] REPP, B. H. (1982), "Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception", *Psychological Bulletin, 92*(1), 81-110.
[14] RICHARDSON, K. H. & SAWUSCH, J. R. (1989), "The perception of stop consonants: Spectral tilt revisited", Paper presented at the 117th meeting of the Acoustical Society of America, Syracuse, NY.
[15] SAWUSCH, J. R. (1986), "Auditory and phonetic coding of speech", In E. C. Schwab, & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines, Vol. 1*. New York: Academic Press.
[16] SAWUSCH, J. R., & DUTTON, D. L. (in press), "Computational metrics for place of articulation information in consonants and vowels", In J. Charles-Luce, P. Luce, & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production and development*. Norwood, NJ: Ablex.
[17] STUDDERT-KENNEDY, M. (1976), "Speech perception", In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics*. New York: Academic Press.
[18] SYRDAL, A. K., & GOPAL, H. S. (1986), "A perceptual model of vowel recognition based on the auditory representation of American English vowels", *Journal of the Acoustical Society of America, 79*(4), 1086-1100.
[19] WANNAMACHER, C. A. & SAWUSCH, J. R. (1989), "The role of phonological permissibility in the phonetic coding of speech", Paper presented at the 118th meeting of the Acoustical Society of America, St. Louis, MO.

7. ACKNOWLEDGMENTS

I would like to thank Dawn Dutton, Deborah Gagnon, Laurie Garrison, Kevin Richardson and Carol Wannamacher for their helpful discussion and critique of the ideas presented here. This research was supported by NIH grant R01DC00219 to the State University of New York at Buffalo.

INTEGRATION FOR EXTRACTION
What speech perception researchers can learn from Gibson and Marr

Jean-Luc Schwartz, Pierre Escudier

Institut de la Communication Parlée
CNRS UA 368 - INPG / ENSERG - Université Stendhal
INPG, 46 Av. Félix-Viallet, 38031 Grenoble Cedex, France / e-mail : schwartz@saphir.imag.fr

ABSTRACT

We try to make the case that integration mechanisms play a key part in information processing in the auditory system, because of the poor coding abilities of single channels, and the complexitity of the auditory image, but that integration mechanisms must be "clever", able to group in a complex way all the information relevant to a given ecologically relevant source. Gibson's intuitions about the "resonance of neural systems" on the "invariant of the physical environment", later reinforced by Marr's representational framework based on parallelism and specialization of intermediary processes should provide a very strong insight into the study of integration.

1. INTRODUCTION

Auditory processing basically begins by cochlear analysis which results in projecting the incident acoustic signal in one of the 50000 primary Type-I neurons of the auditory nerve. This first neural representation of the acoustic signal – the spatio-temporal pattern of discharges in the auditory nerve – exhibits properties which must strongly determine some of the main processing characteristics of further auditory centers :

(i) there is a great deal of redundancy in the neural outputs,
(ii) which means, in some sense, that there is too much information in the auditory nerve representation ;
(iii) however, a first decomposition of the acoustic signal is achieved, each unit looking more closely at a specific characteristic (the energy in the signal within a given frequency band),
(iv) but it is carried out by noisy channels with a poor ability to "represent" what they are looking for (because of classical limitations of neural cells).

Points (ii) and (iii) are related to one of the key concepts developed by Marr in his theory of vision [19], namely that the input – in his case, the image ; for us, the sound – carries too much information, and, more importantly, types of information that are specified at a number of different scales ; hence this input must be analysed in parallel by a number of different processing systems, which produce different representations, from which specific feature extraction can be achieved (property detectors). The pool of detectors grouped in this level of "intermediary processing" converges, according to Marr, towards what he calls a 2–1/2 D level, a level of "pure perception", which "provides the cornerstone for an overall formulation of the entire vision problem" (pp.269-272), and which corresponds, for the specialist of Cognitive Sciences Petitot, to the "morphological level" of the "phenophysics" (phenomenologically significant physics providing the ecological objective information about the real world) [23].

Points (i) and (iv) largely constraint what should be the key information processing mechanism for the elaboration of intermediary representations, namely *integration mechanisms*, able to build up the necessary *statistics* of the noisy poor-coder channels (iv), taking as best profit as possible of the *redundancy* of the auditory nerve representation (i).

Obviously, however, integration cannot be a kind of "smoothing" mechanism that would just give a gross insight into the content of the auditory nerve – it cannot compute a simple first-order statistics – but it must on the contrary be able to group in a complex way all what is relevant to a given feature, and filter out what is not relevant : it must be able to *extract* and *reveal*. This is highly reminiscent of Gibson's view about the "resonance on the ecological invariant" across the variability of the stimulus [8]. As Marr clearly shows, Gibson's brilliant intuition needs a computational framework ; integration could provide one of the cornerstones of this framework, considering that resonance can happen only if the resonant system has found time or space enough in order to filter out transients and build up its resonant behaviour.

We shall discuss here two possible examples of *clever* integration processes, which could help to recover important articulatory manoeuvers from auditory representations, i.e. (i) enable detection of acoustic events controlled in the timing of speech production, or (ii) extract a stable articulatory target "hidden" inside a continuously varying acoustic trajectory.

2. INTEGRATION ACROSS PLACE FOR EVENT DETECTION

Auditory perception needs a good instrument for estimations of temporal relationships between acoustic events. This is true for any ecological situation where timing provides key information on the structure of the objects that produced the sound, and specially for speech perception where the temporal organization of the glottal and supra-glottal gestures is finely controlled by the speaker in order to produce such phonological contrasts as voiced vs non-voiced plosives, simple vs double consonants (gemination), or tense vs lax vowels.

Timing estimations obviously begin with a system able to *detect* acoustic events. The cochlear nucleus seems able to provide a biological equipment for such event detection, with its various kinds of "on" cells [30]. In our laboratory, Wu [28] proposed a simulation of "on" cells based on a model of neural adaptation that he had shown to produce good results at the level of primary neurons in the auditory nerve. At the output of each individual cell of a model of the auditory nerve, a strongly reinforced adaptation mechanism produces the "on" behaviour, with a very large component of temporal derivation, followed by a rather long (100 ms) forward masking effect. We showed that such a model can indeed cope with the acoustic consequences of glottal or supra-glottal articulatory events such as beginning of voicing, beginning of a vocalic state of the vocal tract, beginning of friction [29].

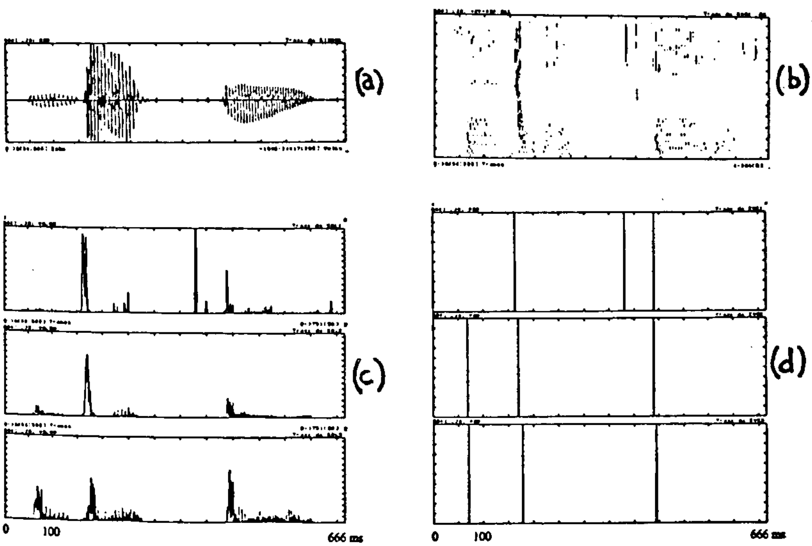


FIGURE 1 - "On" cells modelling and detection of acoustic events
(a) French logatome [baki] (b) response of 64 "on" cells (Characteristic Frequency CF increase from bottom to top) (c) from bottom to top : integrated responses in the low (100-300 Hz), middle (300-900 Hz) and high (900-4000 Hz) CF-regions (d) detected events in the three corresponding CF-regions (see [28] for a precise description of the detection algorithm).

On Fig.1 we show the modelled response of an array of "on"-cells to a French logatome [baki]. The events appear as the *synchronized* occurence of a strong maximum of discharge in a sub-array of cells covering a broad frequency range.

Event detection is based on a *summed response* in one of three sub-arrays, respectively in the low-, middle- or high-frequency region (Fig.1c, 1d). Hence, the functioning of the model relies on (i) strong responses to acoustic events in single frequency channels, due to reinforced adaptation, and (ii) summation of responses based on synchrony of behaviour in a number of channels. Summation of synchronized responses seems to be a very efficient way of signalling events localized in time, with a rather efficient behaviour in noise [1] : indeed in this case the response driven by an event in one frequency channel can be confounded with an amplitude fluctuation inherent to the noise, but, since the acoustic event cannot have a good temporal localization (small time spreading) without covering several frequency bands, it results in *synchronized fluctuations in a number of neighbour channels* : this helps disambiguate the "signal" (the acoustic event) from the noise.

The proposal of mechanisms for detection of coherence of response in a number of frequency channels is in line with an increasing number of psychoacoustic studies about signal detection in noise which show that a signal is easier to detect if inter-channel synchronies help separate the signal from the inherent noise fluctuations, and more generally if inter-channel coherences of the signal and the noise are as different as possible. Thus :

(i) if a noise band masking a pure tone signal is flanked by noise bands in temporal coherence with it, these adjacent bands improve the knowledge of the inherent noise fluctuations and increase the pure tone detectability (Comodulation Masking Release [11]);

(ii) coherent noise bands flanking a signal made of another noise band create less masking on the signal if their fluctuations are uncorrelated with those of the signal than if masker and signal band noises are coherent (Comodulation Detection Differences [18]);

(iii) a temporal gap applied in synchrony on several band noises is easier to detect if the band noises are not coherent, since the gap induces a coherent amplitude fluctuation (a local amplitude minimum) which emerges more when superimposed on uncoherent noise fluctuation patterns [10] ;

(iv) temporal gaps applied on several band noises are easier to detect if they are synchronized than if they are not [9].

All these data show, as Grose & Hall [10] notice, that "the auditory system is able to combine information across critical bands in order to improve temporal acuity" (pp.312), which "might act to facilitate a segregation between a target event (signal or gap) and the background pattern" (pp.313). *Hence, in the case of event detection – a key task for "ecological acoustics" – Gibson's "resonance on the invariant" should involve integration mechanisms finely tuned in time – probably after a stage of reinforcment of the response to temporal variations – and widely spread in place, which is exactly what is realized in Fig.1.* Or, put in Marr's framework, detection of temporal discontinuities – probably crucial for accessing to a possible morphological level equivalent to the vision 2-1/2 D level – should involve neurons owning specific "receptive fields" (see [12] for an extensive use of this concept in audition), which implement a derivation in time and an integration in frequency.

Notice finally that the concept of coincidence of neural discharges could be very general and seems to provide a powerful principle for neural integration in auditory processing [5].

3. RECOVERY OF NON-REACHED TARGETS BY INTEGRATION ALONG ACOUSTIC TRAJECTORIES

After the famous paper by Strange et al. [26] showing that vowels in consonantal context were better identified than isolated ones, a number of contradictory results about static vs dynamic specification of vowel identity were published in the next years : results of several studies from the Haskins Laboratories were in favor of the role of dynamic information, while a number of authors failed to replicate these data and obtained high identification performances for isolated vowels, sometimes better than with vowels in context, and never much lower [27].

Whatever the experimental details or linguistic conditions that could account for the divergence, it remains that formant variability of vocalic targets in function of consonantal or vocalic context, rate and speech conditions is a well-known and classical fact [13, 16, 20] which must be accounted for by human or machine vowel identification systems. Thus we obtained in our laboratory as large as 400 Hz F1- and 1000 Hz F2-variations for [a] and respectively 300 Hz and 600 Hz for F1 and F2 variations for [ɛ] in contexts [iVi] for a French speaker, depending on rate or focus conditions. This speaker displayed values going from (F1 = 800 Hz, F2 = 1250 Hz) for [a] and (F1 = 650 Hz, F2 = 1800 Hz) for [ɛ] in the best conditions (reached target) to (F1 = 400 Hz, F2 = 2250 Hz) for [a] and (F1 = 350 Hz, F2 = 2400 Hz) for [ɛ] in the worst ones (quick rate, no focus on the central vowel), while perceptual tests showed that listeners could easily recognize [a] in all cases, though [ɛ] was more difficult to recognize in the worst conditions.

Listeners' ability to integrate these variations was assessed in several studies [7, 17, 25] and a number of recognition systems incorporate such contextual effects by implicit rules, as in HMMs or MLPs, or explicit modelizations [2, 14].

In this framework it is important to remember that, in echo to Lindblom's initial considerations about modelling vowel reduction by a first-order attractor [16], all present models of speech production introduce at some level the crucial concept of *dynamic systems driven towards an attractor* for the specification of the gesture from one target to the other. Thus the pilot work from the Haskins Laboratories is based on second order dynamics in a target space, from which the articulator dynamics are "backwards" specified by classical principles of inverse dynamics [24]. The work in progress in our laboratory involves a low level of second-order articulatory gestures (mass-spring stiffness driven model [21]) controlled at a higher level by complex optimization principles implemented in a non-linear dynamic system [3]. The important point is that *as soon as a gesture is defined as a dynamic system with the target as an attractor, the equation that specifies the dynamic system provides a constant link between cinematic articulatory variables (such as position x and speed $\dot{x}$) and a limited set of control parameters with values fixed all along the gesture, that determine the gesture and hence the attractor (the "target", reached or not).*

Consider for example a linear second-order system with critical damping, completely described by 2 parameters, namely stiffness and equilibrium point. The determination of 3 close values of x and $\dot{x}$ allows the computation of 2 values of $\ddot{x}$, which then gives 2 linear equations involving the 2 control parameters. Resolution of this system provides us with an estimation of the 2 control parameters and, hence, of the target. Then this estimation can be iterated : each new value of (x, $\dot{x}$), in relation with the last two values of the previous set allows a new target estimation.

There are two major difficulties in this approach, one being the "inverse problem" of estimating articulator positions and speeds from formants, the other concerning the existence and exact nature of dynamic systems for speech gestures. However, these problems are not without solutions [5, 15, 22], though they cannot be discussed in detail here.

Our point is the following. As clearly stated by Elman & Mac Clelland [7], *variability in speech signals is lawful.* However, instead of the complex connectionist system they proposed, with connections variable in time as the "predictors" of the acoustic trajectories, what we perhaps need is *a true system of trajectory analysis, performing all along the trajectory an estimation of its target from estimations of formant positions and speeds, and finally, after integration of these estimations for stabilization of the final result and noise filtering, able to make explicit what was the constant target "hidden" in this trajectory, being it reached or not.* This research program, ambitious but not unrealistic and for which powerful tools are now available, could once more provide us with a "clever" integration mechanism perfectly fulfilling Gibson's requirement about the resonance (of the auditory system) on the invariant (of the speech gesture, namely its attractor specified by its dynamic equation).

CONCLUSION

Grouping (or "linking") of neural excitation is becoming one of the key problems for the neurophysiology of perception. It is our deep conviction that progress in the understanding of integration mechanism for speech perception will be possible only in a general framework, where it is acknowledged that (i) speech sounds are *produced*, and must be perceived as *acoustic consequences of articulatory gestures that obey some of the general* *laws of human gestures*, (ii) which requires *specialized processings, focussed on certain types of information*, at specific time-frequency scales, processings based on *elements of the biological information processing toolbox available in the auditory system.* James Gibson and David Marr seem to concentrate in a remarkable way some of the most profound advances in these two major directions.

ACKNOWLEDGMENTS

The framework presented in this paper was largely determined by highly interesting discussions with C. Abry, F. Berthommier, J. Holdsworth and J. Petitot, and based on works from and exchanges with Y. Arrouas, D. Beautemps, R. Laboissière and Z. L. Wu. Many thanks to all of them.

REFERENCES

[1] ARROUAS, Y., et al. (1991), "Représentations auditives de signaux acoustiques", Workshop on "Traitements et représentations du signal de parole", SFA, Le Mans.

[2] ASSMAN, P.F. (1982), "Vowel identification : orthographic, perceptual and acoustic aspects", *J. Acoust. Soc. Am., 71*, 975-989.

[3] BAILLY, G., et al. (1991), "Formant trajectories as audible gestures - An alternative for speech synthesis", to appear in *Journal of Phonetics*, special issue on Speech Synthesis.

[4] BERTHOMMIER, F. (1991), "Auditory processing based on temporal correlation between adjacent spikes trains", to appear in B. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing (Vol. 3)*, U.K. : JAI Press.

[5] BOE, L.J., et al. (1991), "What geometric variables of the vocal tract are controlled for vowel production ? Some proposals for inversion", *J. of Phonetics*, to appear.

[6] DI BENEDETTO, M.G. (1989), "Frequency and time variations of the first formant: properties relevant to the perception of vowel height", *J. Acoust. Soc. Am., 86*, 67-77.

[7] ELMAN, J.L., Mc CLELLAND, J.L. (1987), "Exploiting lawful variability in the speech wave", in J. S. Perkell and D. H. Klatt (eds.) *Invariance and Variability in Speech Processes* (pp.360-385), Lawrence Erlbaum Asociates.

[8] GIBSON, J. J. (1966), *"The senses considered as perceptual systems"*, New-York, Boston : Houghton-Mifflin.

[9] GREEN, D. M., FORREST, T. G. (1989), "Temporal gaps in noise and sinusoids", *J. Acoust. Soc. Am., 86*, 961-970.

[10] GROSE J.H., HALL, J.W. (1988), "Across-frequency processing in temporal gap detection", in H. Duifhuis et al. (eds.) *Basic Issues in Hearing* (pp.223-231), London : Academic.

[11] HALL, J.W., et al. (1984), "Detection in noise by spectro-temporal pattern analysis", *J. Acoust. Soc. Am.*, 76, 50-56.

[12] HOLDSWORTH, J., et al. (1991), "A multi-representation model for auditory processing of sounds", to appear in 9th Int. Symp. on Hearing.

[13] KOOPMANS - van BEINUM, F. J. (1980), *"Vowel contrast reduction. An acoustic and perceptual study of Dutch vowels in various speech conditions"*, Doct. Thesis, Univ. of Amsterdam.

[14] KUWABARA, H. (1985), "An approach to normalization of coarticulation effects for vowels in connected speech", *J. Acoust. Soc. Am., 77*, 686-694.

[15] LABOISSIERE, R., et al. (1991), "Motor control for speech skills : a connectionist approach", in D. Touretzky et al. (eds.) *Proceedings of the 1990 Connectionist Models Summer School.*, San Mateo, CA : Morgan Kaufmann.

[16] LINDBLOM, B. (1963), "Spectrographic study of vowel reduction", *J. Acoust. Soc. Am., 35*, 1773-1781.

[17] LINDBLOM, B., STUDDERT-KENNEDY, M. (1967), "On the role of formant transitions in vowel recognition", *J. Acoust. Soc. Am., 42*, 830-843.

[18] MAC FADDEN, D. (1987), "Comodulation detection differences using noise-band signals", *J. Acoust. Soc. Am., 81*, 1519-1527.

[19] MARR, D. (1982), "*Vision*", San Francisco : W. H. Freeman and Company.

[20] NORD, L. (1986), "Acoustic study of vowel reduction in Swedish", *STL-QPSR, 4*, 19-36.

[21] PERRIER, P., et al. (1989), "Vers une modélisation des mouvements du dos de la langue", *J. Acoustique, 2*, 69-77.

[22] PERRIER, P., et al. (1991), "Modelling of speech motor control and articulatory trajectories", this Congress.

[23] PETITOT, J. (1990), "Le physique, le morphologique, le symbolique - remarques sur la vision", *Revue de Synthèse, IV/1-2*, 139-183.

[24] SALTZMAN, E. L., MUNHALL, K. G. (1989), "A dynamical approach to gestural patterning in speech production", *Hask. Lab. SR, 99/100*, 38-68.

[25] STEVENS, K. et al. (1966), "Acoustic description of syllabic nuclei : an interpretation in terms of a dynamic model of articulation", *J. Acoust. Soc. Am., 40*, 123-132.

[26] STRANGE, W., et al. (1976), "Consonant environment specifies vowel identity", *J. Acoust. Soc. Am., 60*, 213-222.

[27] WOLF, J. J., KLATT D. H. eds. (1979), *"Speech Communication Papers presented at the 97th Meeting of the Acoust. Soc. of Am."* (pp.15-32), New-York : Acoust. Soc. Am.

[28] WU, Z. L. (1990), *"Peut-on "entendre" des événements articulatoires ? Traitement temporel de la parole dans un modèle du système auditif"*, Doct. Thesis, INP Grenoble.

[29] WU, Z. L., et al. (1991), "Physiologically-plausible modules and detection of articulatory-based acoustic events", to appear in B. Ainsworth (ed.) *Advances in Speech, Hearing and Language Processing (Vol. 3)*, U.K. : JAI Press.

[30] YOUNG, E. D. (1984), "Response characteristics of neurons of the cochlear nuclei", in C. I. Berlin (ed.) *Hearing Science, Recent Advances* (pp. 423-460), San Diego : College Hill Press.

FROM SIGNALS TO SYMBOLS TO MEANING: ON MACHINE UNDERSTANDING OF SPOKEN LANGUAGE[1]

Victor W. Zue

Laboratory for Computer Science, Massachusetts Institute of Technology
Cambridge, Massachusetts 02139, U.S.A.

ABSTRACT

This paper starts with a brief overview of advances in the development of speech recognition systems, with particular emphasis on the past decade. It then moves on to make two points. First, successful development of speech recognition systems will depend on our ability to understand human communication through spoken language, to capture the essential features of the process in appropriate models, and to develop the necessary computational framework to make use of these models for machine understanding. Second, just as human communication using spoken language is an active process of understanding, we must begin to investigate methods that will combine speech recognition and natural language processing technology to achieve speech understanding. Examples to support these arguments will be provided.

INTRODUCTION

Spoken language is the most natural, flexible, efficient, and economical means of communication among humans. As computers continue to play an increasing role in our lives, it is important that we seriously address the issue of providing a graceful human-machine interface through spoken language. Research in speech coding and synthesis has matured over the past decade to the extent that speech can now be transmitted efficiently and generated with high intelligibility. Spoken input to computers, however, has yet to cross the threshold of practicality. To be sure, the last decade has witnessed dramatic improvement in speech recognition technology. Nevertheless, current speech recognition systems still fall far short of human capabilities of continuous speech recognition with essentially unrestricted vocabulary and speakers, under difficult acoustic conditions.

Why is it so hard to develop computer systems to recognize speech? One of the primary reasons is the variabilities that one finds in spoken language communication. Speech can be produced by many speakers with diverse vocal tract anatomies and sociolinguistic backgrounds. Even for a particular speaker, the characteristics of the signal can vary over a wide range, depending on his or her physiological and psychological states. Many external factors, such as the acoustic environment and the types of microphone can also significantly alter the resulting signal. One may be tempted to dismiss these variabilities as undesirable *noise* imposed on the otherwise invariant signal. In reality, however, the process of encoding linguistic information in spoken language is highly stochastic in nature. Speakers of a language can convey the same underlying message with many choices of words and linguistic constructs. Furthermore, even though the inventory of phonemes for a language is quite small, their acoustic-phonetic realizations depend critically on the context in which they appear, as illustrated in Figure 1. For example, while the initial /t/ in the words "two" and "ten" share some acoustic similarities, there are also significant differences that one can readily observe. The burst release for the first /t/ is lower in frequency than the second, a direct consequence of anticipatory coarticulation caused by the following rounded vowel /u/. By the same token, the acoustic similarities of the three /ɛ/'s in the words "seven," "less," and "ten" are overshadowed by the apparent differences. The second /ɛ/ shows articulatory undershoot due to lateralization, as evidenced by the lowering of its second formant, whereas the last /ɛ/ is heavily nasalized, indicated by the smearing of the first formant. Figure 1 also contains more subtle examples of contextual variations. For example, the spectra for the alveolar strident fricatives in the words "is" and "less" both tilt upwards near the end, but for apparently different reasons. In the first case, the upward tilt is due to the following lateral consonant, and is often accompanied by a brief period of epenthetic silence, followed by the relatively sudden lateral release. In the second case, the upward tilt is due to the following dental fricative, which has a more anterior place of articulation. To be sure, remarkable advances have been made in various disciplines of phonetic science, so that we now have a far better

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

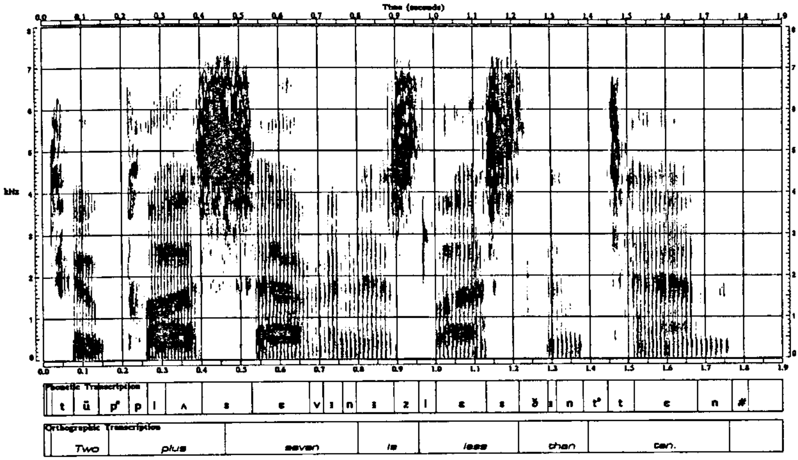


Figure 1: Digital spectrogram of the sentence "Two plus seven is less than ten," spoken by a male talker. Also included are phonetic and orthographic transcriptions that are aligned with important acoustic landmarks in the signal. The spectrogram illustrates some of the acoustic-phonetic variations often found in continuous speech.

understanding of many aspects of this variability than we did a few short decades ago. Nevertheless, researchers in automatic speech recognition have not been able to capitalize on the vast amount of knowledge, primarily because of the lack of a unifying computational framework to make use of it.

I will start this paper with a brief review of the state of the art in speech recognition by machine, with particular emphasis on the past decade. This will be followed by my assessment of the factors contributing to the improvement in systems' performance. I will use the remainder of the paper to make two points. First, successful development of speech recognition systems will depend on our ability to understand human communication through spoken language, to capture the essential features of the process in appropriate models, and to develop the necessary computational framework to make use of these models for machine understanding. Second, just as human communication using spoken language is an active process of understanding, we must begin to investigate methods that will combine speech recognition and natural language processing technology to achieve speech understanding. Indeed, many of the applications of human/machine interface through spoken language require systems possessing the capability of solving a problem interactively with a user. To illustrate my points, I will draw liberally from our own experience in developing speech recognition and speech understanding systems. This is done primarily for the sake of familiarity, and not ethnocentricity.

STATE OF THE ART IN SPEECH RECOGNITION

Defining the Parameters

Speech recognition systems can be characterized by many parameters. An isolated-word speech recognition system requires that the speaker pause briefly between words, whereas a continuous speech recognition system does not. Some systems require speaker enrollment; a user must provide samples of his or her speech before using them. Other systems are said to be speaker-independent in that no enrollment is necessary. Some of the other parameters depend on the specific task. Recognition is generally more difficult when vocabularies are large or have many similar sounding words. The language model is the artificial grammar that restricts the combination of words. The simplest language model can be specified as a finite-state network, where the permissble words following each word are given explicitly. More general language models approximating natural language are specified in terms of a context-sensitive grammar. One popular measure of the difficulty of the task, combining the vocabulary size and the language model, is *perplexity*, P, defined as:

$$P = 2^{-\frac{1}{N}\sum_{i=1}^{N} log_2 P(w_i|w_{i-1},...w_1)}$$

where the w_i are the sequence of all words in all sentences, N is the total number of words, and $P(w_i|w_{i-1},...w_1)$ is the probability of the ith word given all preceding words. Perplexity is related to the average number of words allowed at each node in the language model.[2] Finally, there are some external parameters that can affect speech recognition system performance, including the characteristics of the environmental noise, the type and the placement of the microphone, speaker's level of physiological and psychological stress, and variations in speaking rate.

Performance Review

What follows is a snapshot of current performance of some typical systems on a variety of tasks. It is intended to be illustrative, rather than exhaustive. Interested readers are referred to an extensive review by Mariani for additional information [32].

Performance of speech recognition systems is typically described in terms of word error rate, E, defined as:

$$E = (1 - \frac{S + I + D}{N})100\%$$

where N is the total number of words in the test set, S, I, and D are the total number of substitutions, insertions, and deletions, respectively. Note that insertion and deletion are meaningful measures only for *continuous* speech recognition systems.

Low Perplexity Tasks One of the most popular, and potentially most useful task with low perplexity ($P = 11$) is the recognition of digits. For American English, speaker-independent recognition of digit strings spoken continuously and restricted to telephone bandwidth can achieve an error rate of 0.8% when the string length is known. When the string length is unknown, the error rate increases to 1.4% [48]. This represents a significant improvement over the best systems only a decade ago, which had an error rate of 2%, for digits spoken in isolation by known talkers, recorded under high quality conditions [13]. The French isloated digit recognition system developed at CNET performed robust enough to be deployed over public telephone network [14].

Another potentially useful task is the recognition of English alphabets ($P = 26$). Despite the low perplexity, English alphabet recognition is a very challenging task, since many of the letters are acoustically similar. In 1983, Cole reported an error rate of 10.5% on speaker-independent recognition of isolated digits with a system that makes use of acoustic features known to be important for fine phonetic contrast [9]. Staying with the same philosophy but using an artificial neural net classifier, Cole recently achieved a speaker-independent error rate of 4% for isolated letters of the alphabet [11].

[2]Perplexity for the English language has been estimated from text to be between 150 and 200.

Moderate Perplexity Tasks In the eighties, a number of researchers have pursued speech recognition tasks with a vocabulary of a few hundred words and moderate perplexity. One of the best known is the 1,000-word Resource Management (RM) task, in which inquiries can be made on various naval vessels in the Pacific ocean. This task was made popular by the fact that it is the designated task for common evaluation among contractors of the U.S. Defense Advanced Research Projects Agency's Strategic Computing Program. As a result, speech data for system training and testing, as well as evaluation procedures, have been developed and are readily available [37].

The best speaker-dependent results on the RM task were achieved by BBN and MIT Lincoln Laboratory. Using a word-pair language model that constrains the possible words following a given word ($P = 60$), these systems achieved a word error rate of less than 2% on continuously spoken sentences [38]. The BBN BYBLOS system can also operate in a speaker-adaptive mode, in which the system adapts its models and parameters using only 40 sentences from the new speaker. A 45% reduction in word error rate for the new speaker can be realized with rapid system adaptation [24]. The ARM system developed at RSRE in the United Kingdom achieved an error rate of 13.2% on a 497-word task with no language model ($P = 497$) [42]. For comparison, researchers at IBM reported a word error rate of 9% on the 1,000-word Laser Patent task ($P = 24$) only a few short years ago, and it was the best result at that time [2].

Over the past few years, good performance on speaker-independent recognition for moderate perplexity is beginning to emerge, the best known being the SPHINX system developed at Carnegie Mellon University [27]. On the RM task ($P = 60$), SPHINX achieved a speaker-independent word error rate of 4.5% [38].

High Perplexity Tasks High perplexity tasks with a vocabulary of thousands of words are intended primarily for the dictation application. To make the task manageable and performance reasonable, however, the systems are typically speaker-dependent, and require that the speaker pause between words. Researchers at IBM's T. J. Watson Research Center are among the most active and successful in this area. For example, the TANGORA system achieved word error rate of 2.9% and 5.4% on a 5,000-word and a 20,000 word office dictation task [1]. Similar efforts can also be found in Canada and France [28,34]. The INRS 86,000-word system achieved an error rate of 7.2%, whereas researchers at IBM-France reported an erorr rate of 12.7% on their 200,000-word system.

Discussion

The improvement in speech recognition technology over the last decade was brought on by several factors. First and foremost, there is the coming of age of the utilization of stochastic modelling techniques. The AT&T digit recognition system, the BYBLOS and SPHINX continuous speech recognition systems, as well as all the high perplexity systems mentioned earlier are all based on some form of hidden Markov modelling (HMM). HMM is a doubly stochastic model, in which the generation of the underlying phoneme string and their surface acoustic realizations are *both* represented probabilistically as Markov processes [41, 40]. HMM is powerful in that, with the availability of training data, the parameters of the model can be trained automatically to give optimal performance. While the application of HMM to speech recognition started nearly twenty years ago [21,3], it was not until the past few years that it gained wide acceptance by the research community.

Second, much work has gone into the development of large speech corpora for system development, training, and testing [6,25,37,53,19,5]. Some of these corpora are designed for acoustic phonetic research, while others are highly task specific. These corpora permit researchers to quantify the acoustic cues important for phonetic contrasts and to determine parameters of the recognizers in a statistically meaningful way. The importance of their availability cannot be overstated.

Third, progress has been brought about by the establishment of standards for performance evaluation. Less than a decade ago, researchers trained and tested their systems using locally collected data, and had not been very careful in delineating training and testing sets. As a result, it is very difficult to compare performance across systems, and the system's performance typically degrades when presented with previously unseen data. The recent availability of a large body of data in the public domain, coupled with the specification of evaluation standards [37], has resulted in uniform documentation of test results, thus contributing to greater reliability in monitoring progress.

Finally, advances in computer technology also indirectly influenced our progress. The availability of fast computers with inexpensive mass storage capabilities has enabled many researchers to run many large scale experiments in a short amount of time. This means that the elapsed time between an idea and its implementation and evaluation is greatly reduced.

INCORPORATING SPEECH KNOWLEDGE

Successful systems developed over the last decade are very different from their predecessors. Instead of relying on heuristic rules and intense knowledge engineering, these system derive their power from well formulated mathematical formalisms and automatic training procedures. Nevertheless, it is noteworthy that researchers have generally found that performance of these HMM-based systems can be improved when speech knowledge is incorporated, even if only crudely. For example, the use of triphone models conditioned on the left and right neighbors for a given phoneme implicitly models coarticulation, resulting in approximately 50% reduction in word error rate [27].

While it is hard to speculate on what future speech recognition systems would be like, I believe there are many ways current systems can be made more powerful by the proper utilization of speech knowledge. In this section, I will provide two examples in the area of signal representation and feature extraction.

Signal Representation

Current speech recognition systems perform significantly worse than humans on the same task, even under ideal circumstances [10]. When the operating conditions deteriorate, the difference between human and machine performance becomes even more dramatic. There is clearly much to be learned from studying the process by which human listeners decode the speech signal. While little is known about the decoding process beyond the eighth cranial nerve, advances in auditory physiology and psychophysics [15,22,43] have begun to shed some light on the nature of representations of the speech signal in the human peripheral auditory system. As a result of this pioneering work, many researchers have begun to propose speech signal representations that take into account these known properties of the auditory system [31,23,12, 18,44].

In the recognition system under development in our group, the speech signal is first transformed into a representation based on Seneff's auditory model [44]. The model has three stages. The first stage is a bank of linear filters, equally spaced on a critical-band scale. This is followed by a nonlinear stage that models the transduction process of the hair cells and the nerve synapses. The output of the second stage bifurcates, one branch corresponding to the mean firing rate of an auditory nerve fiber, and the other measuring the synchrony of the signal to the fiber's characteristic frequency.

We believe that outputs from various stages of this model are appropriate for different operations in our system. The nonlinearities of the second stage produce sharper onsets and offsets than are achieved through simple linear filtering. In addition, irrelevant acoustic information is often masked or suppressed. These properties make such a representation well-suited for the detection of acoustic landmarks. The synchrony response, on the other hand, provides enhanced spectral peaks. Since these peaks often correspond to formant frequencies in vowel and sonorant consonant regions, we surmise that the synchrony representation may be particularly useful for performing fine phonetic distinctions.

There has been some evidence suggesting that a representation based on auditory modelling can offer performance advantage, especially when the signal is degraded by noise [16,20,8]. Recently, we conducted a set of formal evaluations that compares several different signal representations [33]. To limit the scope of our investigation, we selected the task of classifying up to 16 vowels in American English, using a multi-layer perceptron (MLP) classifier with a single hidden layer [29]. Vowel tokens were extracted from the TIMIT corpus [26]. Training and test sets consist of more than 20,000 tokens (from 500 speakers) and about 2,000 tokens (from 50 speakers), respectively. Three different types of spectral representations were compared, one based on Seneff's auditory model, one based on mel-frequency cepstral coefficients [35], which are very popular among the HMM-based systems, and one based on a cepstrally-smoothed discrete Fourier transform. To strive towards a fair and meaningful comparison, the mel-frequency filters were carefully designed to resemble the critical-band filters of the auditory model. In addition, the dimensionality of the feature vectors was constrained to be equal. Specifically, a 40-dimensional vector, covering a frequency range of 6 kHz, was computed once every 5 msec. The test tokens were either presented to the classifier unchanged, or were corrupted by additive white noise at an averaged signal-to-noise ratio of approximately 10 dB.

Classification performance is summarized in Table 1. For clean testing tokens, the auditory based representations hold a small but consistent advantage over the other representations. When the test tokens are corrupted by noise, this advantage becomes more substantial. These results suggest that the outputs of the auditory model are more immune to noise degradation, and thus will provide better and more robust performance for phonetic classification.

Signal Representation	Classification Accuracy (%)	
	Clean Speech	Noisy Speech
SAM	66.1	54.0
MFCC	61.6	45.0
DFT	61.2	36.6

Table 1: Comparisons of vowel classification accuracy (in %) for Seneff's auditory model (SAM), mel-frequency cepstral coefficients (MFCC), and cepstrally smoothed discrete Fourier transform (DFT).

Feature Extraction

Most of the current speech recognition systems do not attempt to extract acoustic attributes that are known to signify phonetic contrasts, but instead use the spectral vectors directly for phoneme and word classification. This choice is partly due to the fact that it is difficult to implement reliable algorithms to automatically extract the acoustic attributes, even if we know qualitatively what they are. For example, there does not yet exist a formant tracker that can determine formant frequencies reliably, especially in regions where the direction and the extent of formant transitions provide important information about the place of articulation for consonants. These algorithms also tend to perform poorly near retroflexed and/or nasalized vowels, making incorrect formant assignment that will lead to catastrophic classification errors.

We have recently experimented with a novel procedure for the extraction of acoustic attributes for phonetic classification. We approach this problem by first defining a set of general property detectors based on our knowledge of acoustic phonetics. We then determine the optimal settings of the parameters by a search procedure, using a large body of training data [39,49]. This procedure is illustrated in Figure 2. In this example, we explore the use of the spectral center of gravity as a general property detector for distinguishing front from back vowels. It has two free parameters, the lower and upper frequency edges. An example of this measurement for a vowel token is superimposed on the spectral slice below the spectrogram, with the horizontal line indicating the frequency range. To determine the optimal settings for the free parameters, we first compute the classification performance on a large set of training data for all combinations of the parameter settings. The results are displayed in the middle-right panel in this figure as a performance landscape, where higher values correspond to better performance. We then search for the maximum on the surface defined by the classification performance. The parameter settings that correspond to the maximum are chosen to be the optimal settings. For this example, the classification performance of this attribute, using the automatically selected parameter settings, is shown at the top right corner. Note that an attribute can also be used in conjunction with other attributes, or to derive other attributes.

We believe that the procedure described above is an example of successful knowledge engineering in which a speech scientist provides the knowledge and intuition, and the machine provides the computational power. Frequently, the settings result in a parameter that agrees with our phonetic intuitions. In this example, the optimal settings for this property detector result in an attribute that closely follows the second formant, which is known to be important for the front/back distinction. Our experience with this procedure suggests that it is able to *discover* important acoustic parameters that signify phonetic contrasts, without

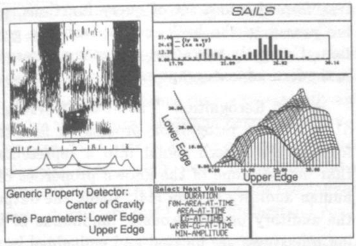


Figure 2: An example of interactive discovery of acoustic attributes for phonetic classification.

resorting to the use of heuristic rules.

Do these attributes offer performance advantage over the direct use of spectral information? We recently performed an experiment on a task of classifying 38 phoneme labels using 55,000 and 9,000 training and testing tokens, respectively, from 350 speakers [30]. The input to an ANN classifier is either the spectral vectors from the auditory model plus segment duration (a 241-dimentional vector), or a set of 80 automatically determined acoustic attributes. The performance for the spectral and attribute representations were 72% and 74%, respectively. This result suggest that the use of acoustic attributes can improve classification performance by a small amount, but at potentially considerable computational savings, since the input vector has been reduced by a factor of three.

FROM RECOGNITION TO UNDERSTANDING

Speech Understanding: The Issues

Speech communication among humans is an active process that utilizes many different sources of knowledge, some of them deeply embedded in the linguistic competence of the talker and the listener. For example, utterances such as "let us pray" and "lettuce spray" can presumably be disambiguated based on acoustic-phonetic knowledge alone, which can be determined from the signal. However, distinguishing others, such as "meet her at the end of the street" and "meter at the end of the street" will require syntactic knowledge. Still others, such as "it is not easy to recognize speech" and "it is not easy to wreck a nice beach," cannot be disambiguated without knowledge of discourse context. On the one hand, higher level linguistic knowledge can serve to constrain the permissible word sequences. Thus, for example, the phoneme sequence /wɛrɪzɪt/ is more likely to be "where is it" than "wear is it," simply because the first one makes more sense. On the other hand, such knowledge helps us understand the meaning of an utterance, which is essential in spoken language communication. The dual role of *filtering* and *understanding* played by syntactic, semantic, and discourse knowledge enables us to converse freely, and to solve problems jointly using spoken language.

All of the systems reviewed earlier have as their goal the production of an orthographic transcription of what was actually spoken. As long as the proper word sequence is produced by the system, it matters little what the underlying linguistic message is. As a result, linguistic knowledge is utilized only to constrain the search space. The constraints are typically implemented as a statistical grammar that specifies the probability of a word given its predecessors. While these simple language models have been effective in reducing search space and improving performance, they do not begin to address the issue of speech understanding. Indeed, many applications suitable for human/machine interaction using spoken language require a system possessing the capability of solving a problem interactively with a user. In addition to converting the speech signal to text, the computer must also understand the user's request, so as to generate an appropriate response.

Speech understanding systems offer a new set of challenges to researchers, and raise several important research issues. Perhaps the most important one is the integration of speech recognition and natural language processing technology to achieve speech understanding. Researchers in each discipline need to investigate how to exchange and utilize information so as to maximize system performance. In some cases, one may have to make fundamental changes in the way systems are designed. For example, most natural language systems are developed with text input in mind; it is assumed that the entire word string is known with certainty. This assumption is clearly false for speech input, whereby many words are competing for the same time span, and some words may be more reliable than others because of varying signal robustness. Therefore, one many not be able to use existing natural language systems without making some modifications.

Another issue related to spoken language system development is that the system must operate in a realistic application domain, where domain-specific information can be utilized to translate spoken input into appropriate actions. For example, the verb "serve" conveys two entirely different meanings, depending on whether one is discussing a restaurant or a tennis match. Realistic application is critical to collecting data on how people would like to use machines to access information and solve problems. The use of a constrained task also makes possible rigorous evaluations of system performance. Finally, the system must begin to deal with interactive speech, where the computer is an active conversational participant, and where people produce spontaneous speech, including false starts, hesitations, and the use of words and linguistic constructs unknown to the system [46].

An Example: The MIT VOYAGER System

Over the past two years, we have begun a research program on speech understanding system development. In this section, we will briefly describe one such system [50,52].

Task Description For our first attempt at exploring issues related to a fully-interactive spoken-language system, we selected a task in which the system knows about the physical environment of a specific geographical area, and can provide assistance on how to get from one location to another within this area. The system, which we call VOYAGER, can also provide information concerning certain objects located inside this area. The current version of VOYAGER focuses on the geographic area of the city of Cambridge, Massachusetts between MIT and Harvard University, and has a vocabulary of only 350 words. Nevertheless, it can carry on a reasonable conversation with the user and answer a variety of queries, as illustrated in Figure 3.

User:	where can i get some chinese food
VOYAGER:	I know of two Chinese places. (VOYAGER shows them on the map.)
User:	which one is closest to me
VOYAGER:	Where are you?
User:	mit
VOYAGER:	The nearest Chinese place that I know of is Royal East.
User:	how far is it
VOYAGER:	Royal East is about one half miles from MIT.
User:	how do i get there
VOYAGER:	I will try to find the best way to walk there. If you are on the same side of the street as MIT, follow the traffic .....

Figure 3: Example of a conversation between a user and VOYAGER that partly illustrates the system's ability to handle discourse information. All words in the user's queries are represented in lower case, with no punctuation, to underscore the fact that such information is not available to the system.

VOYAGER is made up of three components. The first component, SUMMIT, converts the speech signal into a set of word hypotheses. The natural language component, TINA, then provides a linguistic interpretation of the set of words. The parse generated by the natural language component is then transformed into a set of query functions, which are passed to the back-end for response generation. Currently VOYAGER can generate responses in the form of text, graphics, and synthetic speech.

Speech Recognition The SUMMIT system [49, 51] starts the recognition process by first transforming the speech signal into a representation that models some of the known properties of the human auditory system [44]. Using the output of the auditory model, acoustic landmarks of varying robustness are located and embedded in a hierarchical structure called a dendrogram [17]. The acoustic segments represented in the dendrogram are then mapped to phoneme hypotheses, using a set of automatically determined acoustic parameters in conjunction with conventional pattern recognition algorithms. The result is a phoneme network, in which each arc is characterized by a vector of probabilities for all the possible candidates.

Words in the lexicon are represented as pronunciation networks, which are generated automatically by a set of phonological rules. Probabilities derived from training data are assigned to each arc, using a corrective training procedure, to reflect the likelihood of a particular pronunciation. Presently, lexical decoding is accomplished by using the A^* algorithm [4] to find the best path that matches the acoustic-phonetic network with the lexical network.

Natural Language Our spoken language interfaces make use of a natural language component called TINA [45], which is specifically designed to accommodate the integration of speech recognition with natural language processing. TINA is designed so that its grammar rules and associated probabilities can be automatically trained from a set of correctly parsed sentences. This approach has many advantages, including ease of development, portability, and, most important for use with a speech recognition system, low perplexity. We have, in fact, shown experimentally that grammar probabilities can substantially reduce the perplexity of the resulting language model [45].

The grammar is entered by the developer as a set of simple context-free rewrite rules, which are augmented with parameters to enforce syntactic and semantic constraints. The rule set is transformed automatically to a network form. The parser uses a best-first search strategy. Control includes both top-down and bottom-up cycles, and key parameters are passed among nodes to deal with long-distance movement and agreement constraints. The probabilities provide a natural mechanism for exploring more common grammatical constructions first. TINA also includes a new strategy for dealing with movement, which can handle efficiently nested and chained gaps, and rejects crossed gaps.

Control Strategy The integration of the speech recognition and natural language component is currently achieved using an N-best algorithm [7,47, 52], in which the recognizer can propose its best N complete sentence hypotheses one by one, stopping with the first sentence that is successfully analyzed by the natural language component TINA. In this case, TINA acts as a filter on *whole sentence* hypotheses. If all top N word string candidates fail to parse, then the system provides the canned response, "I'm sorry but I didn't understand you."

Application Back End Once an utterance has been processed by the language understanding system, it is passed to an interface component which constructs a command function in order to generate the appropriate response. Figure 4 gives an example of how a query is transformed into a command function. Note that the functions can be nested to construct more complicated functions. The back-end also has some rudimentary but nevertheless effective discourse capability, so that it can deal with simple anaphora, as well as ambiguous queries, as illustrated in Figure 3.

Query:	Where is the nearest bank to MIT?
Function:	`(LOCATE (NEAREST (BANK nil) (SCHOOL "MIT")))`

Figure 4: Example of the translation of a query into a command function for accessing the necessary information from the database.

Performance Evaluation In order to evaluate VOYAGER's performance, we collected a corpus of some 5,000 spontaneously spoken sentences from 100 speakers [46]. The system was trained on approximately 70% of the data and tested on 10%. Errors in the system can occur in several ways; the recognizer can mis-recognize a word, the natural language system can fail to generate a parse, an unknown word can appear, or a query can be outside of VOYAGER's domain. All in all, the system could correctly execute approximately 52% of the queries from unknown users [52]. Only 12% of the queries resulted in incorrect response from the system, which can be viewed as catastrophic error. The remaining 36% of the queries prompted the "I'm sorry but I didn't understand you" message. Currently, VOYAGER is implemented on a SUN-4 workstation, using four commercially available signal processing boards, and runs in 3-5 times real-time.

CONCLUDING REMARKS

Why have advances in speech science in general and phonetic science in particular contributed so little to speech recognition research? I believe that one of the primary reasons must be the fact that our knowledge in this area is very spotty. In many areas, we know quite a bit more than we did a few decades ago. However, every shred of knowledge we possess is more than offset by the vast amount that still eludes us. Locally, the jigsaw puzzle is beginning to fit together, but the overall picture is far from clear. For example, we know that phoneme durations can be very important in signifying phonetic contrast. Despite great gains in our knowledge about segment duration, however, we still do not have an adequate durational model that can simultaneously account for variables such as local phonetic context, higher level linguistic constraints, and speaking rate [36]. Without the complete picture, linguistic use of segment duration for speech recognition is likely to meet with only limited success.

I don't mean to sound pessimistic. Quite the contrary, I think there are ways speech knowledge can help to improve recognition performance, and in this paper I have only given a few examples. Over the past decade, there appears to be a gradual polarization in the positions taken by researchers on speech recognition. Some researchers, mostly engineers enchanted by the elegance of mathematics and the power of computing, believe that the problem will be solved if only we can have enough training data. In their view, speech science has very little to contribute to the solution of the problem. Others, mostly speech scientists who had devoted decades to trying to understand human speech communication, scorn the use of statistical modelling. For them, the solution will not emerge until we truly discover the key that unlocks the speech code. Neither of these extremes can possibly be right. While decades may pass before we can develop systems capable of understanding unconstrained spoken language, we are fast approaching a time when real systems with restricted capabilities will begin to emerge. These systems will in all likelihood operate only in limited domains, but will nevertheless help us interact with computers with greater ease and efficiency, thereby making them more accessible to more people. Success in developing these systems will most likely belong to those who can incorporate the acquired knowledge, however incomplete, into a proper computational model, whose parameters can be determined using a large body of data and the vast amount of computing power that is at our disposal.

ACKNOWLEDGEMENT

Much of the work described in this paper was done in collaboration with Jim Glass, Hong Leung, Mike Phillips, and Stephanie Seneff over the past five years. Their contribution is gratefully acknowledged.

REFERENCES

[1] Averbuch, A., L. Bahl, R. Bakis, P. Brown, G. Daggett, S. Das, K. Davies, S. De Gennaro, P.

de Souza, E. Epstein, D. Fraleigh, F. Jelinek, B. Lewis, R. Mercer, J. Moorhead, A. Nadas, D. Nahamoo, M. Picheny, G. Shichman, P. Spinelli, D. Van Compernolle, and H. Wilkens, "Experiments with the Tangora 20,000 Word Speech Recognizer," *Proc. ICASSP-87*, pp. 701-704, Dallas, TX, April 1987.

[2] Bahl, B., A. Cole, F. Jelinek, R. Mercer, A. Nadas, D. Nahamoo, and M. Picheny, " Recognition of Isolated-Word Sentences From a 5000 Word Vocabulary Office Correspondence Task," *Proc. ICASSP-83*, pp. 1065-1067, Boston, MA, April 1983.

[3] Baker, J. M.,"The Dragon System - An Overview," *Proc. Trans. on Acoustics, Speech and Signal Processing*, vol. ASSP-23, No. 1, pp. 24-29, February 1975.

[4] Barr, A., E. Feigenbaum, and P. Cohen, *The Handbook of Artificial Intelligence*, 3 vols., William Kaufman Publishers, Los Altos, CA, 1981.

[5] Carlson, R., B. Granström, and L. Nord, "The KTH Speech Database," *Speech Communication*, vol. 9, no. 4, pp. 375-380, August 1990.

[6] Carré, R., R Descout, M. Eskenazi, and M. Rossi, "The French Language Database: Defining, Planning, and Recording a Large Database," *Proc. ICASSP-84*, pp. 42.10.1-4, San Diego, CA, 1984.

[7] Chow, Y, and R. Schwartz, "The N-Best Algorithm: An Efficient Procedure for Finding Top N Sentence Hypotheses", *Proc. DARPA Speech and Natural Language Workshop*, pp. 199-202, October 1989.

[8] Cohen, J., "Application of an Auditory Model to Speech Recognition," *Proceedings, Montreal Symposium on Speech Recognition*, p. 8, July 1986.

[9] Cole, R., R. Stern, M. Phillips, S. Brill, A. Pilant and P. Specker, "Feature-Based Speaker-Independent Recognition of Isolated English Letters," *Proc. ICASSP-83*, pp. 731-733, Boston, MA, April 1983.

[10] Cole, R., R. Stern, and M. Lasry, "Performing Fine Phonetic Distinctions: Templates Versus Features," *Variability and Invariance in Speech Processes*, J. S. Perkell and D. H. Klatt, Eds. Hillsdale, NJ: Lawrence Erlbaum Assoc., 1985.

[11] Cole, R., "Spoken Letter Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 385-390, Hidden Valley, PA, June 1990.

[12] Delgutte, B., "Speech Coding in the Auditory Nerve: II. Processing Schemes for Vowel-like Sounds," *J. Acoust. Soc. Amer.*, vol. 75, no. 3, pp. 879-886, March 1984.

[13] Doddington, G. D., and T. B. Schalk, "Speech Recognition: Turning Theory to Practice," *IEEE Spectrum*, vol. 18, no. 9, pp. 26-32, September 1981.

[14] Dutoit, D., "Evaluation of Speaker-Independent Isolated-Word Recognition System over Telephone Network," *Proc. European Conference on Speech Technology*, pp. 241-244, Edinburgh, Scotland, September 1987.

[15] Fletcher, H., "Auditory Patterns," *Review of Modern Physics*, vol. 12, pp. 47-65, 1940.

[16] Glass, J., and V. Zue, "Signal Representation for Acoustic Segmentation," *Proc. First Australian Conference on Speech Science and Technology*, pp. 124-129, November 1986.

[17] Glass, J. R., "Finding Acoustic Regularities in Speech: Applications to Phonetic Recognition," Ph.D. thesis, Massachusetts Institute of Technology, May 1988.

[18] Goldhor, R., "Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features," Ph.D Thesis, Massachusetts Institute of Technology, 1985.

[19] Hedelin, P. and D. Huber, "The CTH Speech Database: An Integrated Multilevel Approach," *Speech Communication*, vol. 9, no. 4, pp. 365-374, August 1990.

[20] Hunt, M., and C. Lefebvre, "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model," *Proc. ICASSP-88*, pp. 215-218, April 1988.

[21] Jelinek, F., L. Bahl, and R. Mercer, "Design of a Linguistic Decoder for the Recognition of Continuous Speech," *Proc. IEEE Symposium on Speech Recognition*, pp. 255-266, Pittsburgh, PA, April 1974.

[22] Kiang, N. Y-S., T. Watanabe, E. Thomas, and L. Clark, *Discharge Patterns of Single Fibers in the Cat's Auditory Nerve*, Research Monograph No. 35, The M.I.T. Press, Cambridge, MA, 1965.

[23] Klatt, D., "Speech Processing Strategies Based on Auditory Models," *The Representation of Speech in the Peripheral Auditory System*, R. Carlson and B. Granström, Eds. New York: Elsevier/North-Holland, pp. 181-196, 1982.

[24] Kubala, F. and R. M. Schwartz,"A New Paradigm for Speaker-Independent Training and Speaker Adaptation," *Proc. DARPA Speech and Natural Language Workshop*, pp. 306-310, Hidden Valley, PA, June 1990.

[25] Kuwabata, K., K. Takeda, Y. Sagisaka, S. Katagiri, S. Morikawa, and T. Watanabe, "Construction of a Large-Scale Japanese Speech Database and Its Management System," *Proc. ICASSP-89*, pp. 560-563, Glasgow, Scotland, May 1989.

[26] Lamel, L., R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, Report No. SAIC-86/1546, pp. 100-109, February 1986.

[27] Lee, K. F., *"Automatic Speech Rcognition: The Development of the SPHINX System,"* Kluwer Academic Publishers, Boston, 1989.

[28] Lennig, M., "An 86,000 Word-Recognizer Based on Phonemic Models," *Proc. DARPA Speech and Natural Language Workshop*, pp. 391-396, Hidden Valley, PA, June 1990.

[29] Leung, H. and V. Zue, "Phonetic Classification Using Multi-Layer Perceptrons," *Proc. ICASSP-90*, pp. 525-528, Albuquerque, NM, May 1990.

[30] Leung. H., J. Glass, M. Phillips and V. Zue, "Detection and Classification of Phonemes Using Context-Independent Error Back-Propagation," *Proc. ICSLP-90*, pp. 1061-1064, Kobe, Japan, November 1990.

[31] Lyon, R., "A Computational Model of Filtering, Detection, and Compression in the Cochlea," *Proc. ICASSP-82*, pp. 1282-1285, Paris, France, 1982.

[32] Mariani, J., "Recent Advances in Speech Processing," *Proc. ICASSP-89*, pp. 429-440, Glasgow, Scotland, May 1989.

[33] Meng, H. and V. Zue, " A Comparative Study of Acoustic Representations of Speech for Vowel Classification using Multi-Layer Perceptrons," *Proc. ICSLP-90*, pp. 1053-1056, Kobe, Japan, November 1990.

[34] Mérialdo, B., "Speech Recognition with Very Large Size Dictionary," *Proc. ICASSP-87* pp. 364-367, Dallas, TX, April 1987.

[35] Mermelstein, P. and S. Davis, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-28, no.4, August 1980.

[36] Nooteboom, S. G., "Some Observations on the Temporal Organization and Rhythm of Speech," *These Proceedings*.

[37] Pallett, D., "Benchmark Tests for DARPA Resource Management Database Performance Evaluations," *Proc. ICASSP-89*, pp.536-539, Glasgow, Scotland, May 1989.

[38] Pallett, D., J. Fiscus, and J. Garafolo, "DARPA Resource Management Benchmark Test Results June 1990," *Proc. DARPA Speech and Natural Language Workshop*, pp. 298-305, Hidden Valley, PA, June 1990.

[39] Phillips, M.," Automatic Discovery of Acoustic Measurements for Phonetic Classification," *J. Acoust. Soc. Am.*, vol. 84. S216, 1988.

[40] Poritz, A.,"Hidden Markov Models: A Guided Tour," *Proc. ICASSP-88*, pp. 7-13, New York, NY, 1988.

[41] Rabiner, L. R.,"An Introduction to Hidden Markov Models," *IEEE ASSP Magazine*, January 1986.

[42] Russel, M.,"Recent Results from the ARM Continuous Speech Recognition Project," *Proc. DARPA Speech and Natural Language Workshop*, pp. 397-402, Hidden Valley, PA, June 1990.

[43] Sachs, M., and E. Young, "Effects of Nonlinearities on Speech Encoding in the Auditory Nerve," *J. Acoust. Soc. Amer.*, vol. 68, no. 3, pp. 858-875, September 1980.

[44] Seneff, S., "A Joint Synchrony/Mean Rate Model of Auditory Speech Processing," *Journal of Phonetics*, vol.16, no.1, pp. 55-76, 1988.

[45] Seneff, S., "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," *Proc. ICASSP-89*, pp. 711-714, Glasgow, Scotland, May 1989.

[46] Soclof, M. and V. Zue, " Collection and Analysis of Spontaneous and Read Corpora for Spoken Language System Development," *Proc. ICSLP-90*, pp. 1105-1108, Kobe, Japan, November 1990.

[47] Soong, F., and E. Huang, "A Tree-Trellis Based Fast Search for Finding the N-best Sentence Hypotheses in Continuous Speech Recognition," *Proc. DARPA Speech and Natural Language Workshop*, pp. 199-202, June 1990.

[48] Wilpon, J., C.H. Lee, and L.R. Rabiner, "Improvements in Continuous Digit Recognition Using Higher Order Spectral and Energy Features," *Proc. ICASSP-91*, Toronto, Canada, May 1991.

[49] Zue, V., Glass, J., Phillips, M., and Seneff, S. "Acoustic Segmentation and Phonetic Classification in the SUMMIT System," *Proc. ICASSP-89*, pp. 389-392, Glasgow, Scotland, May 1989.

[50] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. "The VOYAGER Speech Understanding System: Preliminary Development and Evaluation," *Proc. ICASSP-90*, pp. 73-76, Albuquerque, NM, May 1990.

[51] Zue, V., Glass, J., Goodine, D., Phillips, M., and Seneff, S. "The SUMMIT Speech Recognition System: Phonological Modelling and Lexical Access," *Proc. ICASSP-90*, pp. 49-52, Albuquerque, NM, May 1990.

[52] Zue, V., Glass, J., Goodine, D., Leung, H., Phillips, M., Polifroni, J., and Seneff, S. "Recent Progress on the VOYAGER System," *Proc. DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June 1990.

[53] Zue, V., S. Seneff, and J. Glass,"Speech Database Development at MIT: TIMIT and Beyond," *Speech Communication*, vol 9, no. 4, pp. 351-356, August 1990.

Report on the paper SIGNALS TO SYMBOLS TO MEANING: MACHINE UNDERSTANDING OF SPOKEN LANGUAGE

Renato De Mori
McGill University, School of Computer Science
3480 University street - Montreal - Quebec - Canada

ABSTRACT
This paper is a report on V. Zue's invited paper : FROM SIGNALS TO SYMBOLS TO MEANING : ON MACHINE UNDERSTANDING OF SPOKEN LANGUAGE

1. CONSIDERATIONS ON THE PAPER'S CONTENT

The paper of Professor Zue is an accurate and concise review of the State of the Art in Automatic Speech Recognotion (ASR). Discussing results in terms of intervals of language model perplexities is interesting even if, as Zue points out, lower perplexities do not necessarily imply easier tasks.

Zue's reasons why there is a stochastic component that has to be added to speech and linguistic knowledge are also pertinent. In fact speakers may convey the same underlying message with many choices and linguistic constructs.

Another important aspect emphasized by Zue's paper is the need to compare experimental results by using public domain speech corpora.

Also of interest are considerations on the use of an ear model, feature extraction and the description of the Voyager system.

2. SOME ADDITIONAL CONSIDERATIONS ON HIDDEN MARKOV MODELS

Hidden Markov Models (HMM) have been very popular and highly successful tools for acoustic modelling. The following aspects seem to be of interest:

1) How many different models are required for recognizing large vocabularies?

Triphone models seem to be the solution adopted by many researchers, but their number is very high, making parameter estimation not very accurate with the available data. Various techniques for smoothing and clustering have been proposed (an interesting clustering algorithm has been recently presented by Bahl [1]). Interesting ideas have been proposed by Paul [2] and Bartakova and Jouvet [3] trying to take into account phonetic contexts inspired by phonetic knowledge in such a way that the number of units is kept in the order of magnitude of a few thousand. It is also important to note that in many recent systems triphone models are highly influenced by the word used to extract their parameters and tend to be word dependent.

2) Is corrective training a valid approach?

Recently, various types of algorithms have been developed for such a purpose, based, for example, on Maximum Mutual Information Estimation [4].

3) Are there new acoustic parameters worth using?

There seem to be a tendency of considering new dynamic parameters, like the second derivative of energy and melscaled cepstral coefficients [5]. The parameters and the approach for feature extraction mentioned by Zue in his review are also worth mentioning.

4) What is the role of Artificial Neural Networks (ANN)?

ANNs are essentially function approximators. They can approximate classifiers or data compressors that compute a reduced set of acoustic parameters based on a large set of parameters considered by a designer relevant for performing some acoustic classification. It is important, for this purpose, to have algorithms for global parameter estimation, i.e. for estimating at the same time the parameters of an ANN and the parameters of an HMM which takes the output of the ANN as an observation. Interesting algorithms can be found in [6-7].

ANNs have not succesfully modeled so far the dynamic characteristics of speech processes. Nevertheless, they have produced promising results as classifiers/detectors of some phonetic features. This possibility may be useful in future. A system could be conceived with a set of networks, each network being specialized in the detection of a complete set of features. The network specialized for a set can be fed by acoustic parameters suitable for them in terms of groups of features and their contexts, making different units only for those contexts for which feature variations imply important coarticulation effects.

3. LANGUAGE AND DIALOGUE MODELLING

In Automatic Speech Recognition, the words of a sentence are not available when the sentence has to be interpreted, or if they are available, they are affected by errors. This makes the understanding problem a search problem in which partial interpretation theories are scored. A coherent scoring methodology has been developed based on probabilities. So far linguistic knowledge has been mostly represented by trigram probabilities of having a word or a Part Of Speech (POS) given the two preceding ones. Estimating the probabilities required by these models presents some problems even if large corpora are available. These problems have found interesting solutions in recent years. Nevertheless, these models take into account only a limited context for a word, while it is well known that the expectation of a word in a sentence may depend on the entire sentence structure and, more generally, on the state of the conversation.

An interesting approach emerging now is based on stochastic grammars. Relevant problems along this line include for example the computation of the probability that a grammar generates a sentence only a part of which is known (see [8] and [9] for examples).

Another interesting problem is related to the opportunity of accepting only partial parses of a spoken sentence instead of forcing a complete parse of it.

The role of semantics is also of fundamental importance in speech decoding as a filter, as Zue pointed out, for example, in situations which a recognizer produces the N best word sequences (an efficient algorithm has been proposed recently for this purpose [10]) using only a language model based on bigram probabilities. Another important role of semantic knowledge could be that of predicting new words to be detected in the signal. For this purpose, new mathematical frameworks for Language Modeling have to be developed. A relevant problem is also that of linguistic knowledge acquisition from written corpora.

Finally there is an emerging interest in Dynamic Language Models in which the expectation of a word is considered as a function of the state of the verbal message or the dialogue. Cache memories can be used for this purpose [11].

4. REFERENCES

[1] L. R.Bahl, P. V. De Souza, P. S. Gopalakrishnan, D. Nahamoo and M. A. Picheny, Context-Dependent Modeling of Phones in Continuous Speech. Proc. 4th DARPA Speech and Natural Language Workshop, Asilomar, CA, Feb. 1991

[2] D. B. Paul, New Results with the Lincoln Tied-Mixture HMM-CSR System. ibid.

[3] K. Bartakova and D. Jouvet , Allophone Modelling 12th International Phonetic Congress, Aix en Provence, Aug. 1991

[4] Y. Normandin and S. Morgera, An Improved MMI Training Algorithm for Speaker-Independent, Small Vocabulary Continuous Speech Recognition. IEEE Intl. Conf. On Acoustics, Speech and Signal Processing, Toronto, Ontario, May 1991

[5]J. C. Wilpon, C. H. Lee and L. R. Rabiner, Improvemennts in Continuous Digit Recognition Using High Order Spectral and Energy Features. ibid.

[6] J. Bridle, A Recurrent Neural Network Architecture with a Hidden Markov Model Interpretation. Speech Communication, vol. 9, no. 1, Feb. 1990, pp.83-92.

[7] Y. Bengio, R. De Mori, G. Flammia and R. Kompe, Global Optimization of a Neural Netwotk-Hidden Markov Model Hybrid. Mc Gill University, School of Computer Science, Internal Report, Dec. 1990.

[8] F. Jelinek, Computation of the Probability of of Initial Substring Generation by Stochastic Context-Free Grammars. IBM Internat Report, April 1990.

[9] A. Corazza, R. De Mori, R. Gretter and G. Satta, Computation of Probabilities for an Island Driven Parsers. IEEE Transactions on Pattern Analysis and Machine Intelligence (to appear).

[10] F. K. Soong and E. F. Huang, A Tree-Trellis Based Fast Search For Finding N best Hypotheses in Continuous Speech recognition. IEEE Intl Conference on Acoustics, Speech and Signal Processing, Toronto, Ontario, 1991

[11] R. Kuhn and R. De Mori, A Cache-Based Natural Language Model for Speech Recognition IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, no. 6, June 1990, pp. 570-583.

THE RELATIONS BETWEEN SPEECH PRODUCTION AND LEVELS OF REPRESENTATION: INTRODUCTORY REMARKS

Björn Lindblom

Department of Linguistics
University of Texas at Austin

ABSTRACT

This session can be seen as complementary to a previously organized symposium on **Speech processes in the light of action theory and event perception** [1]. It is aimed at examining the role of perceptual processes in the motor control of "phonetic gestures".

1. THE INVARIANCE ISSUE: IN-PRINCIPLE SOLUTIONS.

One of the classical problems in phonetics is the difficulty of specifying physical invariants corresponding to linguistic categories [2]. There are currently several favored research paradigms that, implicitly or explicitly, take a stance on that issue and can be said to offer programs for the in-principle resolution of it. They seem to fall in either of two categories.

Theme 1: **Phonetic invariance is in the signal.** According to this approach phonetic invariance for linguistic categories will ultimately be established once we learn to look at the signal in the right way and to make the right kind of measurements be they articulatory, acoustic or auditory. The tacit hope is that one day discoveries will be made that render Hockett's often quoted Easter Egg metaphor inappropriate [3]. Recent formulations of the Motor Theory [4], Direct Realism [5], Coordinative Structures [6], the Quantal Theory [7] and the notion of "Icebergs" [8] appear to come close to this approach.

Theme 2: **Phonetic invariance is <u>not</u> in the signal.** This alternative links the variability of speech signals to the adaptive organization of speech. According to this view, the listener's short-term demands for explicit signal information do not stay constant during and across utterances. Thus the lack of signal invariance is seeen to arise as a consequence of the talker's tacit recognition of variations in short-term perceptual and situational demands and his/her adaptive response to them. The so-called H&H theory exemplifies this type of reasoning [9].

2. IMPLICATIONS FOR SPEECH MOTOR CONTROL

What are the implications of these alternatives with respect to the task of the speaker? What are the parameters that (s)he actively controls?

Our answers to those questions are closely tied to the assumptions we make about the nature of speech perception. For instance, if speech perception is assumed to be based on the extraction of higher-order signal invariants - be they gestural, acoustic or auditory - then speech motor control must be seen as aimed at producing those gestural, acoustic or auditory invariants.

If, on the other hand, it is <u>not</u> based on signal invariance, we must envision control parameters in a different way. Let us suppose that the role of the signal is to supplement knowledge already available to the listener (in short- and long-term memory) and that its purpose is to discriminate among competing candidates in the listener's lexicon. Hence the speaker's task is to control the phonetic discriminability - rather than the invariance - of signal attributes. In other words, the talker should generate signals that are sufficiently rich to facilitate correct identification.

3. CONSTANCY AND ADAPTIVE TUNING OF PHONETIC GESTURES

Within the time limits of this symposium it is not possible to do justice to all the paradigms that are currently explored in phonetics and that bear on the invariance issue. The selection of the present contributions was deliberately made so as to promote a discussion of issues related to Theme 2. The justification for that decision is that so far, although neither particularly novel nor counter-intuitive, theme 2 seems to represent the scientifically less traveled research avenue.

Some of the issues that follow from theme 2, can be stated as follows: What is the status of "phonetic gestures"? Are they the theoretical primitives from which explanations of on-line phonetic variability are to be derived? Are they the ultimate control units of speech motor control? Or are "phonetic gestures" in no way prime constructs but themselves derivable from the dynamic tug-of-war between production and perception demands? If so, to what extent are they tuned to meet perceptual demands? If perceptually motivated tuning of gestures can indeed be demonstrated, what is the extent of such transforms? In other words, what is the scope of cooperative behavior in speaker-listener interaction (cf Nooteboom)? If, as argued by Kohler, the varying degrees of reduction that speech exhibits in response to situational and perceptual conditions, how do we describe those listener-oriented conditions in a quantitative and language-independent way (cf Diehl)? And in a similarly language-independent manner how do we quantify processes of reduction? Does Articulatory Phonology (Browman and Goldstein) present a way of addressing those questions?

4. EXISTENCE AND SIGNIFICANCE "CLEAR SPEECH"

In the present context I would like to draw attention to a recent study of "clear speech" by Seung-Jae Moon of the University of Texas at

Austin [10]. Moon made measurements of vowel and consonant formant patterns in the stressed /w_l/-syllables of mono-, bi- and tri-syllabic words. These test items were chosen so as maximize locus-to-target distances, maintain stress constant and produce variations in the duration of the stressed vowel. The idea behind this design was to induce "undershoot" effects in the vowel formant patterns. Two speaking styles were investigated: citation forms and "clear", overarticulated speech. His results, which will be summarized in a paper contributed to this conference, indicate that "clear speech" tokens are not simply louder citation forms but involve reorganization of acoustic patterns and the underlying articulatory gestures. Acoustically, this reorganization takes the form of removing contextual effects of the /w_l/ environment, that is reducing undershoot, and shifting "clear speech" formant patterns closer to null-context reference values for the various vowel categories. Intelligibility tests showed that "clear speech" is more resistant to noise than citation forms.

I am going to allow myself to generalize from Moon's results regarding them as a correct and generally valid description of "clear speech". By way of conclusion, let me raise the following questions for our session: Why should there be such a thing as "clear speech", that is a style of speech that apparently has different acoustic properties from those of neutral citation-form speech? Why do speakers bother to change their pronunciation when attempting to clarify? Given that they do change their articulatory patterns, what does such behavior imply about the organization of speaker-listener interactions? And what does the very existence of "clear speech" imply about perceptual processing and the invariance issue?

5. REFERENCES

[1] LINDBLOM B (ed) (1986): Speech processes in the light of action theory and event perception, theme issue, Journal of Phonetics 14:1, 1-196.
[2] PERKELL J and KLATT D (1986): Invariance and variability in speech processes, LEA:Hillsdale, N J.
[3] HOCKETT C F (1958): A manual of phonology, Bloomington, Indiana:Indiana University Press.
[4] LIBERMAN A M and Mattingly I G (1985): "The Motor Theory of Speech Perception Revised", Cognition 21:1-36.
[5] FOWLER C A (1986): "An Event Approach to the Study of Speech Perception from a Direct-Realist Perspective", Journal of Phonetics 14:1, 3-28.
[6] KELSO J A S, SALTZMAN E L and TULLER B (1986): "The dynamical perspective on speech production: Data and theory", Journal of Phonetics 14:1, 29-59.
[7] STEVENS K N (1989): "On the quantal nature of speech", Journal of Phonetics 17:1/2, 3-45.
[8] FUJIMURA O (1990): "Articulatory perspectives of speech organization", in Hardcastle W J and Marchal A (eds): Speech Production and Speech Modeling, Kluwer Publishers, Dordrecht.
[9] LINDBLOM B (1990): "Explaining phonetic variation: A sketch of the H&H theory", 403-439 in Hardcastle W J and Marchal A (eds): Speech Production and Speech Modeling, Kluwer Publishers, Dordrecht.
[10] MOON S-J (1991): An Acoustic & Perceptual Study of Formant Undershoot in Clear & Citation-Form Speech, Ph.D. diss. University of Texas at Austin.

6. ACKNOWLEDGEMENTS

This research was supported by a grant from the Advanced Research Program of the Texas Board of Coordination and grant No. BNS-9011894 from the NSF.

LISTENER-ORIENTED CONSTRAINTS ON ARTICULATORY ORGANIZATION

Randy L. Diehl

Department of Psychology
University of Texas at Austin

ABSTRACT

A good many phonetic regularities reflect a strategy of talkers to maintain sufficent auditory distinctiveness among speech segments. In defending this claim, I discuss a number of converging methods for evaluating the notion of "auditory distance" in a language-independent way.

1. INTRODUCTION

It is often asserted that many phonetic and phonological regularities reflect a kind of trade off between the needs of the talker (minimal necessary effort) and the needs of the listener (sufficient perceptual contrast). Although most phoneticians would not object in principle to this general claim, in practice they have tended to treat it as more of a slogan than as a basis for genuine phonetic explanation. (Of course, there are notable exceptions to this tendency, as the other papers in this symposium effectively demonstrate. See also [15].) Various reasons no doubt exist for this state of affairs, but two come to mind immediately. First, there is the long tradition within linguistics of viewing the study of phonetics as largely devoted to pure physical description of speech sounds. Within this descriptivist tradition, the study of functional aspects of speech communication is often seen as secondary or even irrelevant. Second, there has been considerable skepticism that notions such as "minimal effort" or "sufficient contrast" can be formulated with adequate precision and validity to have significant explanatory content. For example, Ladefoged [14] recently dismissed the possibility of devising a language-neutral measure of auditory distinctiveness on the grounds that such a measure is inevitably confounded by the language bias of the observer.

I wish to offer a rather more optimistic view of the prospects for a program of functional explanation in phonetics and phonology. Although both talker- and listener-oriented selection pressures are involved in the shaping of sound systems and "on-line" speech behavior, the focus in this brief discussion will be on phonetic regularities that appear to reflect mainly the requirements of listeners.

2. THE AUDITORY ENHANCEMENT HYPOTHESIS

It is a striking fact about articulatory organization that the phonetic properties of vowels and consonants tend to covary in a highly regular manner. Across languages, back vowels are usually rounded while front vowels are usually unrounded, high vowels tend to be produced with a higher fundamental frequency (F0) and with a higher velar position than low vowels, voiced consonants are usually preceded by longer vowels and followed by a lower F0 than voiceless consonants, and so on. For many such regularities, phoneticians have sought explanations in terms of putative physical or physiological constraints on production. Thus, for example, the F0 correlate of tongue height has been attributed to a passive mechanical coupling between the tongue body and larynx (hence the phrase "intrinsic vowel pitch") [8, 13].

The problem with many of these explanations is that, although they may be superfically plausible, they fail to account in detail for the relevant facts. For example, the F0 correlate of tongue height shows up even in the esophageal speech of laryngectomized patients [7]. Because such speakers obviously lack laryngeal cartilages as well as a hyoid bone, the F0 effect cannot be explained by the coupling hypothesis. Moreover, recently several groups of investigators have reported that higher vowels are associated with increased levels of cricothyroid activation [9, 18], which suggests that the F0 correlate is under active control by the talker.

My colleagues and I [5, 6] have been investigating an alternative hypothesis concerning the origin of many significant types of phonetic covariation. We claim that the phonetic properties of vowels and consonants covary as they do largely because language communities tend to select properties that have mutully enhancing auditory effects. The obvious result of such a selection strategy is to produce segments that are more distinctive auditorily. Of course, in communication situations involving low noise and relatively high redundancy, talkers can afford to trade away some distinctiveness for greater ease of articulation. But the potential for greater distinctiveness must be built into the sytem to be exploited when necessary.

How might this auditory-enhancement hypothesis explain, for example, the F0 correlate of tongue height? A principal acoustic correlate of high vowels is a low-frequency first formant (F1). However, evidence suggests that the best predictor of *perceived* vowel height is not F1 per se, but rather the difference in Bark units between F1 and F0, with smaller differences yielding perception of a higher vowel [17]. A possible auditory basis for this effect has recently been suggested by Beddor [1]. On the basis of the notion of a 3.5 Bark spectral integrator [3], she hypothesized that when F0 is raised sufficiently close to F1, the spectral center of gravity associated with F1 shifts downward, contributing to a perceived raising of the vowel. If this hypothesis is correct, then the so-called "intrinsic vowel pitch" may actually reflect a strategy of enhancing the auditory distinction between high and low vowels.

Elsewhere [6] we have argued more generally that the auditory-enhancement hypothesis can account for the salient phonetic properties of the most common vowels. In the case of canonical productions of the vowel /u/, for example, almost every independently controllable articulatory parameter is set to enhance the distinctive lowering of the first two formant frequencies. In fact, virtually all of the theoretical options for lowering the first two resonant frequencies of a tube-like configuration appear to be exploited. These options include vocal-tract lengthening through lip protrusion and larynx lowering, constriction near the antinodes of the standing volume-velocity waveforms corresponding to these resonances (i.e., at the lips and velopharyngeal area), and dilation near the nodes of the same standing wave patterns (i.e., at the midpalate and lower pharynx). Of these, only palatal dilation can properly be argued to be a by-product of other vocal-tract gestures, in this case, tongue-body retraction. The rest of the distinctiveness-enhancing gestures appear to be actively selected.

3. METHODS OF EVALUATING AUDITORY DISTINCTIVENESS

Let us now return to Ladefoged's argument that any attempt to devise a metric of auditory distinctiveness will inevitably be entangled in the language biases of the observer. A preliminary answer to this is that one might use measures of acoustic distance in place of an auditory metric, and such physically defined measures would presumably be language independent. For example, the above auditory-enhancement rationale for the detailed articulatory properties of the vowel /u/ is framed entirely in terms of acoustic distinctiveness. In the final analysis, however, acoustic distance metrics are theoretically insufficient. The auditory system is nonlinear in many respects, and a fully adequate distance metric must incorporate these nonlinearities. So the question remains: How can we circumvent the problem of observers' language bias in evaluating the notion of auditory distinctiveness? I will suggest some possible approaches that have been used in our laboratory and elsewhere.

3.1. Speech/Nonspeech Comparisons

In attempting to test the auditory-enhancement hypothesis, my colleagues and I have often compared listeners' categorization performance on speech stimuli to their performance on nonspeech stimuli that are acoustically analogous in certain relevant respects. Consider the following example. In most languages, [+voice] consonants in medial position are distinguished from [-voice] consonants in having *inter alia* a shorter constriction interval and significant glottal pulsing during the constriction interval. We [16] hypothesized that the presence of glottal pulsing makes the short constriction appear even shorter, thus enhancing the distinctiveness of the constriction-duration correlate of [+voice] consonants.

To evaluate this claim, we first had listeners identify two series of /aba/-/apa/ stimuli varying in closure duration. In one series, the items contained a segment of glottal pulsing during closure, while in the other series, the closure interval contained only silence. As expected, the presence of pulsing shifted the /b/-/p/ category boundary, yielding more /b/ responses. A second group of listeners were asked to identify two corresponding series of nonspeech stimuli, each consisting of two square-wave segments separated by a medial gap of varying duration. Like the speech stimuli, one series contained a segment of glottal pulsing during the gap, while the other did not. These stimuli are highly nonspeechlike and they do not correspond to any obvious natural categories, so it is necessary to provide training in labeling the stimuli. Initially, the subjects learn by means of feedback to press one response key when they hear the series-endpoint stimulus with the shortest medial gap and a different key when they hear the series-endpoint stimulus with the longest medial gap. After this training, the listeners are asked to identify the entire series on the basis of similarity to either end-point stimulus. (We refer to this task as end-point similarity matching.) Interestingly, for these nonspeech stimuli, the presence of pulsing during the gap shifted the category boundary in the same direction as observed for the /aba/-/apa/ stimuli. We took this as evidence that glottal pulsing does in fact make the medial gap appear shorter. Notice that the parallel between the speech and nonspeech results suggests that the effect is not to be explained in terms of linguistic experience but rather in more general auditory terms.

3.2. Adult/Infant Comparisons
In a related study [4], we found that prelinguistic infants showed categorical discrimination of the same /aba/-/apa/ stimuli used in the adult experiment just described. In addition, there was evidence that the presence of glottal pulsing during closure caused a category boundary shift comparable in magnitude to that of the adult subjects. The average age of the infant subjects (7 1/2 months) was younger than the age at which effects of linguistic experience are typically first observed. Therefore, the results tend to confirm our earlier conclusion that the effect of glottal pulsing on the perception of the closure-duration cue is of a general auditory nature rather than being a product of linguistic experience.

3.3. Cross-Native-Language Comparisons
Mandarin syllables carrying a mid-high-rising F0 contour (Tone 2) tend to be shorter than those carrying a low-falling-rising contour (Tone 3). We [2] hypothesized that talkers use length differences to enhance the perceptual contrast between Tones 2 and 3. A primary difference between the two tone categories is that Tone 2 has a relatively short initial period of nonrising F0 prior to a rising interval, whereas Tone 3 has a relatively long period of nonrising F0 before its rising interval. The effect of proportionally lengthening one of these contours is to increase the absolute duration (and hence detectability) of the initial nonrising F0, making the contour perceptually more like Tone 3. We tested this hypothesis by comparing perceptual judgments by native Mandarin and native English speakers on various synthetic series of Mandarin tones ranging incrementally from a Tone 2 contour to a Tone 3 contour, with stimuli from the different series varying in length. (The Mandarin-speaking subjects performed lexical labeling, whereas the English-speaking subjects were trained and tested on the end-point similarity matching task, described earlier.) Both groups of subjects had very similar category boundaries, and both showed the predicted effect of syllable length (i.e., longer stimuli being more likely to be assigned to the Tone 3 category). The cross-native-language similarity suggests that language experience is not a main factor in the length effect. Rather, general auditory factors seem to be responsible.

3.4. Human/Animal Comparisons
Along with others such as Kuhl [12], we have conducted a series of experiments comparing speech categorization performance of humans and animals. Most of the results to date indicate rather striking cross-species similarities. For example, in his dissertation work, Kluender [11] had both humans and Japanese quail categorize a /g/-/k/ stimulus set that varied orthogonally in both voice onset time and F1 onset frequency. For both groups of subjects, a lower F1 onset frequency produced a reliable shift in the labeling boundary, yielding more responses corresponding to the [+voice] category. This similarity suggests that the low-frequency F1 typical of voiced stops enhances the perception of voicing for reasons that have little to do with linguistic experience.

3.5. Auditory Modeling
Recently, a highly realistic model of mammalian auditory-nerve response [10] has been made available to us. This model, together with standard distance metrics applied to sets of its output representations, can provide language-independent estimates of auditory distances among speech stimuli. These estimates can in turn

be used to evaluate auditory-enhancement accounts of regularities involving phonetic covariation.

4. SUMMARY

There are reasons to be optimistic that functional explanatory notions such as "auditory distance" can be formulated with considerable precision and validity. In the context of discussing evidence for the auditory-enhancement hypothesis, I outlined a number of convergent approaches to the evaluation of auditory distance, each of which apparently avoids the confounding effects of observer bias.

5. REFERENCES

[1] Beddor, P.S. (to appear). "On predicting the structure of phonological systems", *Phonetica*.

[2] Blicher, D.L., Diehl, R.L. & Cohen, L.B. (1990). "Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement", *J. Phonetics*, 18, 37-49.

[3] Chistovich, L.A. & Lublinskaya, V.V. (1979). "The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustic study of perception of vowel-like stimuli", *Hear. Res.*, 1, 185-195.

[4] Cohen, L.B., Diehl, R.L., Oakes, L.M. & Loehlin, F.C. "Infant discrimination of /aba/ versus /apa/", unpublished manuscript.

[5] Diehl, R.L. & Kluender, K.R. (1989). "On the objects of speech perception", *Ecol. Psych.*, 1, 121-144.

[6] Diehl, R.L., Kluender, K.R. & Walsh, M.A. (1990). "Some auditory bases of speech perception and production", *Advances in speech, hearing and language processing*, 1, 243-267.

[7] Gandour, J. & Weinberg, B. (1980). "On the relationship between vowel height and fundamental frequency", *Phonetica*, 37, 344-354.

[8] Honda, K. (1981). "Relationship between pitch control and vowel articulation", in D. Bless & J. Abbs (eds.), *Vocal-fold physiology* . San Diego: College-Hill.

[9] Honda, K. & Fujimura, O (1989). "Phonological vs. biological explanation--intrinsic vowel pitch and phrasal declination", Paper presented at the 6th Vocal Fold Physiology Conference, Stockholm.

[10] Jenison, R.L., Greenberg, S., Kluender, K.R. & Rhode, W.S. (to appear). "A composite model of the auditory periphery for the processing of speech based on the filter response functions of single auditory-nerve fibers", *JASA*.

[11] Kluender, K.R. (1988). "Auditory constraints on phonetic categorization: trading relations in humans and nonhumans", Unpublished doctoral dissertation, University of Texas at Austin.

[12] Kuhl, P.K. (1988). "Auditory perception and evolution of speech", *Hum. Evolution*, 3, 19-43.

[13] Ladefoged, P. (1968). "*A phonetic study of West African languages*", Cambridge: Cambridge University Press.

[14] Ladefoged, P. (1990). "Some reflections on the IPA", *J. Phonetics*, 18, 335-346.

[15] Lindblom, B. (1986). "Phonetic universals in vowel systems", in J.J. Ohala & J.J. Jaeger (eds.), *Experimental phonology*, Orlando, Fl: Academic.

[16] Parker, E.M., Diehl, R.L. & Kluender, K.R. (1986). "Trading relations in speech and nonspeech", *Percept. Psychophys.*, 34, 314-322.

[17] Traunmüller, H. (1981). "Perceptual dimension of openness in vowels", *JASA*, 69, 1465-1475.

[18] Vilkman, E., Aaltonen, O, Raimo, I, Arajärvi, P. & Oksanen, H. (1989). "Articulatory hyoid-laryngeal changes vs. cricothyroid activity in the control of intrinsic F0 of Vowels", *J. Phonetics*, 17, 193-203.

LIP ROUNDING AS SIDE CONTACT

Louis Goldstein

Yale University and Haskins Laboratories
New Haven, Connecticut, USA

ABSTRACT

Data are examined to determine the "goal" of the articulatory gesture(s) for lip rounding. It is hypothesized that rounded vowels differ from unrounded vowels in the presence vs. absence of contact between the upper and lower lips along their sides.

1. ARTICULATORY PHONOLOGY

In the articulatory phonology approach developed in the last few years [2,3,4], phonological units are *gestures*. Each is modeled as a dynamical regime that controls the formation of a constriction within one of the relatively independent vocal tract subsystems (i.e., the lips, tongue tip/blade, tongue body, glottis, and velum). The constriction goals for a given gesture are defined, not at the level of individual articulators, but at a *task* level [10,11], where the task specifies the degree (and for oral articulators the location) of the vocal tract constriction. For example, the goal of the bilabial closure gesture at the beginning of the word "bad" is defined in terms of the task or *tract variable* of *lip aperture*, the vertical distance between the upper and lower lips.

One consequence of this approach is that phonological units can be defined relatively invariantly in terms of tract variables. Contextual variation in the relative contributions of individual articulators to a given gesture emerges automatically as a consequence of the temporal overlap among invariantly specified gestures, because an individual articulator's motion is determined by the entire ensemble of concurrently active gestures to which it is relevant. Thus, in the case of the bilabial closure gesture, the relative contribution of the upper lip, lower lip and jaw to lip aperture will automatically differ depending on the jaw requirements of an overlapping (co-produced) vowel gesture [4, 11].

2. LIP GESTURES

Lip constrictions can be described in three dimensions:

LA, Lip Aperture: Vertical distance between the upper and lower lips measured at the center of the lips when viewed from the front.

LW, Lip Width: Side to side measure of lip opening when viewed from the front.

LP, Lip Protrusion: Protrusion of the upper or lower lips from the teeth, as seen in profile. More generally, this this is taken to be anterior-superior positioning of the lips with respect to the teeth, encompassing both "protrusion" and "retraction".

The object of this paper is to investigate which of these dimensions are specified in the "task" control of various lip gestures.

2.1 Consonant Constrictions

As noted above, lip closure gestures can be specified using LA [2,4,10,11]. It is unlikely that any additional specification of LP or LW is required. However, LP is clearly important for labiodental fricatives, in which there is retraction (of the lower lip).

2.2 Rounding and Vowels

While it is possible to hypothesize a relatively invariant tract variable goal for lip closure gestures (in terms of LA), it is more difficult to do so in the case of lip rounding gestures. All three constriction dimensions have been investigated [1,6,9] as potentially relevant to distinguishing vowels on the basis of lip rounding. However, each has been shown to vary considerably from vowel to vowel, for both rounded and unrounded vowels.

The hypothesis proposed here is that none of these constriction dimensions is used to specify the "task" goals for gestures that distinguish rounded from unrounded vowels. Rather, what is specified is whether or not the *upper and lower lips touch along their sides*. Specifically:

(1) Phonologically "rounded" vowels must be produced with contact along the sides (upper and lower lips touching).

(2) Phonologically "unrounded" vowels must be produced with *no* contact along the sides.

If side contact is what is specified for vowel gestures (positively for "rounded" vowels, negatively for "unrounded" vowels), then direct measurement of the length of contact along the sides of the lips should categorically divide the set of rounded vowels from the set of unrounded vowels. Distance from the corner of the mouth to the most forward point of contact was measured in Linker's [9] cross-language study of rounding, and her data for Cantonese, Finnish, French, and Swedish are plotted here in Fig. 1. In general, all the vowels that are phonologically rounded have substantial side contact, while unrounded vowels have virtually none. If .9 millimeters is set as an absolute (speaker- and language-independent) criterion for contact, then all but three of the 272 vowel tokens from eight speakers of four languages are appropriately classified. Note that for Swedish, this means that both the "inrounded" as well as the "outrounded" vowels (as traditionally described [5]) are classified as rounded. (Note that the symbol /ɯ/ is used here for the high front "inrounded" vowel). Differences between the two types of rounding will be discussed below. Also, Swedish /ɑː/ is here classed with the rounded vowels. There are differences in the literature as to the status of rounding in this vowel [5,7] (and in any case, rounding is not contrastive for low vowels in Swedish).

Given this side contact specification, the patterns of variation shown by LA and LW both within and across rounding classes are predictable, as will be argued below. In addition, it is possible for LP to be specified independently of the *touching/no touching* specification. In such cases (e.g., Swedish, below), LP may contrast within the class of rounded vowels as defined by (1). In other cases, LP may contribute, as an "articulator," to side contact (or its absence), with the exact amount of LP determined in a language-specific or speaker-specific fashion.

Finally, in this analysis, rounding and consonant constrictions can be seen as complementary. Consonants control the vertical opening between the two lips along the midline, while rounding for vowels controls the opening along the sides (Rounding for consonants has not been analyzed from this perspective).

3. UNROUNDED VOWELS

For unrounded vowels, LA has been observed to vary considerably without any concomittant variation in LW. This can be seen in the English data presented by Fromkin [6]. This independence of LW and LA follows from the hypothesis that the sides of the lips do not touch in such vowels. As illustrated in the frontal views in Fig. 2, if the sides don't touch, LA can change substantially without automatically changing LW.

The differences in LA among the unrounded vowels probably do not have to be specified as part of the task-dynamic control for these vowels. As Fromkin notes, LA variations are predictable from the different jaw heights found for the vowels. In fact, the slope of the relation between LA and jaw height appears to be about one in her data. Thus, millimeter for millimeter, all of the variation in LA can be attributed to the jaw positioning. In the task dynamic model, these different jaw heights result, in turn, from the different requirements of tongue positioning. Thus, no active control of LA would be required.

Spreading. The vowel /i/ has been shown to involve active retraction of the corners of the lips in English [12] and in

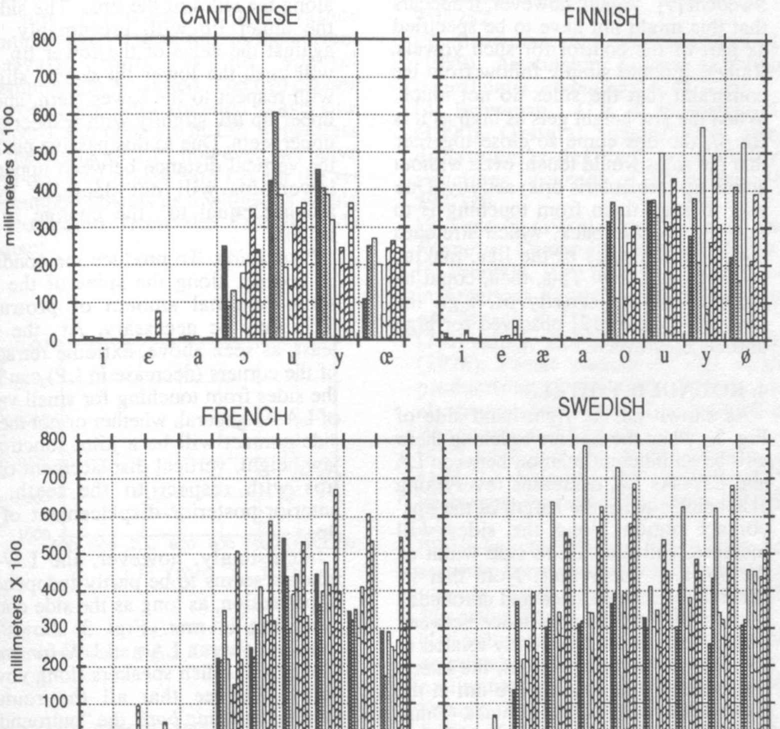


Figure 1. Length of contact along sides of lips. For each vowel, separate bars represent single tokens from each of eight speakers. (All Cantonese and Finnish vowels are long). Data from Linker [9].

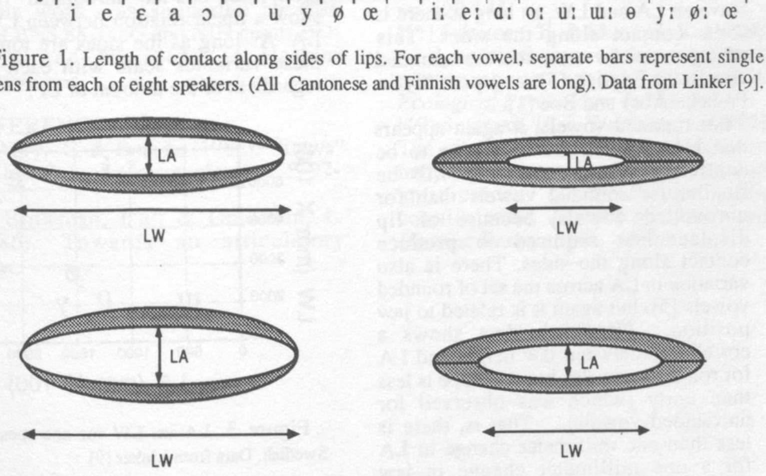


Figure 2. Hypothesized effects of changing LA as a function of side contact.

Swedish [7]. Again, however, it appears that this might not have to be specified as part of the control for such vowels. Rather, it could simply follow from the constraint that the sides do not touch. When the jaw height gets as high as it is for /i/, the lips come so close together that the sides would touch, even without active lip (orbicularis oris) activity. One way to keep them from touching is to pull the corners back, which stretches and thins the sides of the lips, making contact less likely. This, then, could be the goal of the activity of, e.g., the *risorius* muscle [12] observed for high unrounded vowels.

4. ROUNDED VOWELS

As shown on the right-hand side of Fig. 2, when the lips are touching there will be an inherent relation between LA and LW. As LA decreases (everything else being equal), the length of the lips' contact region along the sides will increase, and the side-to-side width of the *opening* decreases. Note that in general, for both rounded and unrounded vowels, the horizontal distance between the *corners* is not inherently related to LA. But for rounded vowels, the lateral endpoints of the opening are not at the corners, because of the contact. Thus, there should be an inherent relation between LA and LW, as long as there is some contact along the sides. This relation can be seen for the rounded vowels in English [5], and possibly in French, Abry and Boe [1].

For rounded vowels, it again appears that LA, per se, does not have to be controlled. In general, LA will be smaller for rounded vowels than for unrounded vowels, because of lip displacement required to produce contact along the sides. There is also variation in LA across the set of rounded vowels [5], but again it is related to jaw position. Fromkin's data shows a correlation between jaw height and LA for rounded vowels, but the slope is less than unity (which was observed for unrounded vowels). That is, there is less than one millimeter change in LA for a one millimeter change in jaw height. Again, however, this may follow from the fact that the sides of the lips are touching for these vowels. Imagine what happens as the jaw raises in a configuration in which there is contact along the sides of the lips. The sides of the upper lip will, presumably, push against the sides of the lower lip. This will push the lower lip *down* slightly with respect to the lower teeth, and the upper lip *up* slightly with respect to the upper teeth. Due to this passive pushing, the vertical distance between upper and lower lips will not decrease by an amount equal to the change in jaw height.

Protrusion. To produce the condition of contact along the sides of the lips, some minimal amount of protrusion (LP) may be necessary. At the very least, as seen above, extreme retraction of the corners (decrease in LP) can keep the sides from touching for small values of LA. In general, whether or not there is side contact will be a joint function of jaw height, vertical displacement of the lips with respect to the teeth, and anterior-posterior displacement of the lips.

Interestingly, however, the LW-LA relation seems to be partly independent of protrusion, as long as the side contact condition is met. Fig. 3 shows the relation between LA and LW for one of Linker's Swedish speakers (long vowels only). We see that all the rounded vowels including *both* the "outrounded" (y, ø, o, ɑ) and the "inrounded" (u, ɯ), show a linear relation between LW and LA. As long as the sides are touching, these variables scale with each other, regardless of the amount of LP.

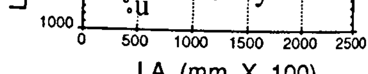


Figure 3. LA vs. LW for one speaker of Swedish. Data from Linker [9].

The two rounding groups *do*, however, differ substantially in LP, as can be seen in Figure 4. This is consistent with Ladefoged and Maddieson's [8] characterization of "outrounded" vowels as [protruded], and "inrounded" vowels as not [protruded] (but rather [compressed]). Note that in the current analysis, there is no special "compression" required. The two classes of vowel are both rounded (have side contact), but contrast in LP. Regardless of the status of LP, however, we have seen that LW scales with LA for the set of rounded vowels as a whole, as would be expected from the fact of side contact.

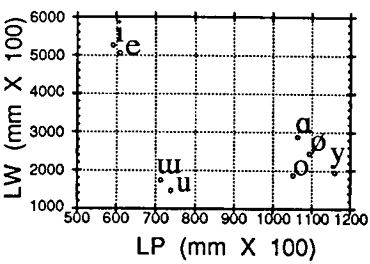


Figure 4. LP vs. LW for the speaker of Swedish, shown in Fig. 3. Data from Linker [9].

AKNOWLEDGEMENTS

This work was supported by NSF grant BNS 8820099 and NIH grants HD-01994 and NS-13617 to Haskins Laboratories.

REFERENCES

[1] Abry, C. & Boe, L. (1986). "Laws" for lips. *Speech Communication, 5*, 97-104.

[2] Browman, C.P. & Goldstein, L. (1986). Towards an articulatory phonology. *Phonology Yearbook, 3*, 219-252.

[3] Browman, C.P. & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201-251.

[4] Browman, C.P. & Goldstein, L. (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics, 18*, 299-320.

[5] Fant, G. (1973). *Speech Sounds and Features*. Cambridge, MA: MIT Press.

[6] Fromkin, V.A. (1964). Lip positions in American English vowels. *Language and Speech, 7*, 215-225.

[7] Hadding, K, Hirose, H. & Harris, K. (1976). Facial muscle activity in the production of Swedish vowels: an electromyographic study. *Journal of Phonetics, 4*, 233-245.

[8] Ladefoged, P. & Maddieson, I. (1990). Vowels of the world's languages. *Journal of Phonetics, 18*, 93-122.

[9] Linker, W. (1982). Articulatory and acoustic correlates of labial activity in vowels. *UCLA Working Papers in Phonetics, 56*, i-ii and 1-134.

[10] Saltzman, E. (1986). Task dynamic coordination of the speech articulators: a preliminary model. In H. Heuer & C. Fromm, *Experimental Brain Research Series 15*, New York: Springer-Verlag, pp. 129-144.

[11] Saltzman, E. & Munhall, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology, 1*, 333-382.

[12] Sussman, H. & Westbury, J. (1981). The effects of antagonistic gestures on temporal and amplitude parameters of anticipatory coarticulation labial coarticulation. *Journal of Speech and Hearing Research, 46*, 16-24.

THE ORGANIZATION OF SPEECH PRODUCTION
CLUES FROM THE STUDY OF REDUCTION PROCESSES

K. J. Kohler

Institut für Phonetik und digitale Sprachverarbeitung
Kiel, Germany

ABSTRACT

Starting from various different views on levels of phonetic representation a unitary approach is presented which centres on the explanatory power of articulatory reduction processes and their auditory control.

1. LEVELS OF REPRESENTATION

The Journal of Phonetics devoted its July 1990 issue to the theme "Phonetic Representation" [1] and thus provided a good opportunity for the phonetics world to find out more about the heterogeneity of its views concerning the relationship between phonetics and phonology. The most encouraging outcome of this discussion is the unanimous realisation that phonology and phonetics need each other. This means that there has been progress in superseding not only the a-linguistic treatment of speech as purely physiological or acoustic events by the first experimental phoneticians, but also the exclusive reliance on symbolic phonological representations in the wake of Trubetzkoy's humanities/science dichotomy. What remains an open issue is the specification of the relationship. Is it to be seen as a mapping from one level of representation to another (following Keating [1, 321ff] or Pierrehumbert [1, 375ff]) or as an integration (following Browman and Goldstein [1, 299ff] or Fowler [1, 425ff])? How many levels of representation should there be: just two - phonology and phonetics, or four - phonology, categorical phonetic representation, articulatory and acoustic parametric representations (Keating), or multiple, with as many phonetic representations as there are different measures of interest for a particular question in hand (Pierrehumbert)?

Intimately linked with the number of levels of phonetic representation is the question as to the ontological status of these levels. Phoneticians and phonologists are interested in finding out how speech communication works and therefore aim at the categories that make it possible, i.e. which are relevant for the language user and form the basis for communication. So they all tacitly or explicitly assume that the levels of phonetic representation are not just heuristic devices for making descriptions of data, but have a reality outside linguistics and phonetics in language and speech behaviour. It is at least doubtful whether this goal has been reached in all cases. Pierrehumbert's multiple phonetic levels are clear instances of descriptive frames, rather than inherent components in speech communication. Keating's categorical phonetic interface between phonology and parametric representations is likewise a surface descriptive device which does not address the question of phonetic explanation in phonology, e.g. with regard to the voicing contrast (cf. [2]).

But even Browman and Goldstein [1], in spite of their laudable attempt at integration in an articulatory phonology, do not go far enough, as Mattingly [1] so succinctly points out: "What is called for is a functional rather than an anatomical organization, one in which gestures are grouped according to the nature of their tasks." (p. 450) In such a functional approach, output considerations do actively constrain the processes of variation in speech production; variation with respect to speaking style and prosodic context does not simply follow from very general principles of gestural overlap and magnitude that are blind to their acoustic consequences [1, 303-4]. This becomes clear the minute we transgress the limitations of lab speech word contrasts and look at processes in connected, continuous and spontaneous speech, where gestures not only overlap to the extent that one becomes completely hidden by others, thus losing its acoustic consequences, but where gestures are constantly eliminated altogether, by comparison with more careful speech.

The most obvious case of heuristics is Ladefoged's discussion of the IPA classification and transcription framework [1, 335-346]. L. is quite right in stressing the pragmatic rather than theoretical basis of the IPA alphabet, but he is wrong in the conclusions he draws from this state of affairs for an anthropophonic approach. His argument runs as follows:
(1) Given the phonological basis of the IPA - whatever its precise nature may be - it is necessary to classify those sounds as different that are used in any one language to distinguish between words as well as those that are contextually conditioned allophones.
(2) When sounds occurring in different languages have to be classified the problem may arise of deciding whether they are the same, as, e.g., dental or interdental fricatives in different varieties of English.
(3) There are no language-independent criteria such as articulatory ease and auditory phonetic similarity that could solve this problem, because due to observer bias there is no principled way of setting up an auditory theshold or of measuring degree of articulatory effort in a general language perspective. What general reasons could we give for treating bilabial and labiodental fricatives as different, but dental and interdental ones as the same, both pairs involving small differences in place of articulation, but only the first being attested as a phonological opposition in languages, e.g. Ewe.
(4) It follows from (3) that all we can do is listing the speech sounds observed in all the world's languages.

In response, I would like to raise the following counter-points:
(a) Although establishing the segmental contrasts for word differentiation is an important phonetic task, it has to be supplemented by research in two further areas to be of validity for the elucidation of the speech communication process:
1. the analysis of contrasting segments in all possible phonetic parameters and in their differentiated influences on the segmental environment;
2. the manifestation of phonological oppositions in connected, continuous and spontaneous speech.
(b) It follows from (a)1. that, with regard to the fricative distinctions under (3), the comparably slight differences of place of articulation in the two pairs can have very different consequences on the inherent acoustic parameters of fricative intensity and spectrum as well as their temporal courses, with auditory results of different magnitudes (cf. [5], plate 11).

This point is further strengthened by the reference in [8, 79] that [ɸ] may alternate with [p] in Ewe, which means that the fricative can be related to a weakening of a closure in an articulatory reduction process, resulting in a weak fricative. This is similar to the change of intervocalic [b] to a voiced bilabial (frictionless) approximant [β] as a stylistic variant in casual German speech of, e.g., 'liebe' (*dear*), compared with a labiodental fricative [v] in 'primitive'. The very genesis of [β] under the conditions of reduction in less formal spontaneous speech of (a)2. rules out the development of a fricative.

(c) Thus the study of synchronic reduction processes in connected, continuous and spontaneous speech of a variety of languages will give us insights into general language-independent scales of articulatory effort, particularly when we can find independent physiological motivations. These synchronic data can be supplemented with data from diachronic sound change adding to their explanatory power.

(d) Phonemic oppositions do not all show the same stability. Some exhibit lower resistance to coalescence than others, for articulatory and/or acoustic-auditory as well as environmental (e.g. noise) reasons. Even when [ɸ] and [f] constitute separate phonemes for the differentiation of isolated words their auditory distance is smaller than the one that separates either of them from [s]. Similarly, [θ] and [f] are auditorily closer than [s] and either of these sounds, as can be judged from the sound change [θ] --> [f] in some varieties of English. It is indeed possible to make language-independent statements about the auditory distances or similarities of sounds, provided the techniques of investigation go beyond contrasting isolated words by linguistically biassed observers and include the native speaker reaction in articulation score type experiments [7], in direct similarity assessment, in slips of the ear during spontaneous speech, in the study of auditory constraints on articulatory reduction, for a variety of languages.

2. ARTICULATORY REDUCTION

I will now give a brief summary of reduction processes in German (for further details see [3, 4]) and draw conclusions from them for the organization of speech production and levels of representation. In the sequence of the preposition 'mit' (*with, by*) and the definite article 'dem' (dat.) in 'mit dem Auto' (*by car*) we can get the following series of reductions from most to least careful.

I. mɪt^{h}de:m
mɪt^{h}dem
mɪt^{h}dəm
mɪtdm

II. mɪt̚p:m
mɪp:m

III. mɪpm
mɪb̥m
mɪbm
mɪmm
mɪm

This series of segmental changes contains a great number of phonemic switches , and, therefore, the question arises as to whether we are here dealing with a single continuous scale of reduction and one large range of signal variability, related, say, to the most elaborate form (taken as the underlying abstract invariant) or whether this scale is discontinuous, with each phonemic change constituting a new invariant reference point for signal variability. The series can be divided into the indicated three groups according to speech production criteria. In I., we get a progressive shortening of the opening-closing movements between two oral closures, until there is no longer an open phase between them. In II., coarticulation occurs between two successive closure gestures (apical and labial/dorsal), and the apical one can be more and more reduced until it is finally eliminated. In III., we then find a progressive shortening of the oral plus velic closure configuration, i.e., an increasingly earlier descent of the velum for the nasal release. With this shortening, passive voicing can continue through the stop closure phase, and eventually the complete blockage of air is eliminated altogether, the velum being lowered as early as the onset of the labial closure. Seen in this way, the phonemic switches do not throw any light on the significant changes in speech parameters, because they cut across them.

We thus have a continuous natural progression in the reduction of articulatory movements from the full form [mɪtʰ de:m] to the fully reduced one [mɪm]. Although there are a number of discrete phonemic changes along this articulatory continuum they do not capture the essentials of the speech production processes. In these cases, speakers do not set out to alter an underlying phonemic chain in a series of steps before they hand it over to the articulators for execution, nor do they select different forms from a lexicon, and the changes are not caused by simple peripheral vocal tract constraints either. It may be assumed that speakers start from a segmental representation of the full forms [mɪtʰ] and [de:m] taken from a lexicon, and attribute to the sequence a reduction coefficient at a high processing level before actual execution. The strength of this reduction coefficient determines the extent of articulatory smoothing, always beginning with the group I processes and moving further and further down the continuum with increasing strength. The phonemic representation of these reduction stages makes it possible to provide a first description of the phenomena, but it is misleading from a speech production point of view, because it is also far too complicated. Instead, we can think of a general programme for articulatory reduction, whose components are specified and hierarchically ordered for a particular language or dialect (some may even be universal) according to the degree of reduction to be achieved. The components of this programme are triggered by the reduction coefficient. It is not necessary to specify all the possible types of assimilations etc. for individual segmental sequences, but quite general instructions along the lines given in the characterisation of the three groups of processes suffice in this general reduction programme, whose specific output depends on the application of the general phonetic rules to specific segmental sequences according to the strength of the reduction coefficient. Browman and Goldstein's articulatory phonology complemented by Mattingly's functional approach and extended to connected speech would provide a good basis for such a reduction programme.

But it would also have to be supplied with an auditory control component, because speakers not only control reduction with regard to the physiological and articulatory potentials contained in the dynamics of sound production but also take listeners into account and adapt to their needs [6] in two ways:
(a) Reduction processes are favoured that show a low degree of perceptual salience. That is the reason why apical fricatives, released apical plosives and syllable or word initial apical nasals and plosives are not assimilated in German (e.g. 'Beamte'[bə'ʔamtʰə] vs. 'Beamten' [bə'ʔampm] (*civil servants*). Fricatives have more distinctive acoustic structures separating the different places of articulation than nasals and stops, and among the latter, unreleased ones are

still less salient than released or even aspirated ones. Furthermore, the syllable or word initial position has a higher signalling value for a listener and must therefore be given a more precise articulation by a speaker. Thus the final position has a higher reduction coefficient than the initial one, allowing for instance final [bm] in 'geben' (*to give*), but preventing the sequence [pm] across the word boundaries in 'gib nicht' (*don't give*). What is not very distinctive for a listener anyway may be reduced by a speaker more easily to yield to the principle of economy of effort.

(b) Different communicative situations put different demands on the perceiver of speech, and speakers have to tune their performance to these conditions to guarantee a successful language interaction ([6]). This means that speaking styles in keeping with different speech environments exhibit varying degrees of reduction oriented towards the listener's needs

3. CONCLUSIONS

Instead of the dichotomy of phonetics and phonology or of multiple levels of phonetic representation I propose a unitary explanatory level, which I call "Phonetics in Language and Speech Communication". Besides this domain of an ontological status there may be as many heuristic ones as one deems necessary for preliminary descriptive orientation. In the pursuit of this communicative goal the study of word contrasts and lab speech has to be supplemented by the analysis of connected, continuous and spontaneous speech, which will make the investigation of articulatory reduction processes possible. They will in turn throw light on the organization of speech production and allow us to give a more precise account of the concept of articulatory effort in a language-independent perspective. They will also provide a direct link with historical sound changes. One aim must be the quantification of reduction coefficients for different speaking styles, replacing the atomisation of a multitude of articulatory parameters by global functional coordinative structures. They in turn need an auditory control unit that regulates the ease of articulation in relation to the perceptual demands of the communicative situation and incorporates language-independent perceptual distance measures.

4. REFERENCES

[1] "*Journal of Phonetics*", *18(3)* (1990).

[2] KOHLER, K. J. (1984), "Phonetic explanation in phonology: the feature fortis/lenis", *Phonetica*, *41*, 150-174.

[3] KOHLER, K. J. (1990), "Segmental reduction in connected speech in German: phonological facts and phonetic explanations"; in HARDCASTLE and MARCHAL (eds.) "*Speech production and speech modelling*", pp. 69-92, Dordrecht/Boston/London: Kluwer Academic Publishers.

[4] KOHLER, K. J. (1991), "The phonetics/phonology issue in the study of articulatory reduction", *Phonetica*, *48*.

[5] LADEFOGED, P. (1964), "*A Phonetic study of West African languages*", Cambridge: Cambridge University Press.

[6] LINDBLOM, B. (1990), "Explaining phonetic variation: a sketch of the H & H theory", in HARDCASTLE and MARCHAL (eds.) "*Speech production and speech modelling*", pp. 403-439, Dordrecht/Boston/London: Kluwer Academic Publishers.

[7] MILLER, G. A. and NICELY, F. E. (1955), "An analysis of perceptual confusions among some English consonants", *JASA*, *27*, 338-352.

[8] WESTERMANN, D. and WARD, I. C. (1933), "*Practical phonetics for students of African languages*", London: Oxford University Press.

PERCEPTUAL GOALS OF SPEECH PRODUCTION

Sieb G.Nooteboom

Research Institute for Language and Speech, Utrecht University

ABSTRACT

In this paper predictions are made about the production of words embedded in fluent connected speech, based on a principle of cooperative behaviour combined with insights in perception. It is concluded that the effects of such cooperative behaviour, in as far they are not brought about by linguistic means, are real but relatively small.

1. INTRODUCTION

Sometimes speakers speak in order to be understood. On those occasions, they are well advised to adapt their pronunciation to the estimated needs of their audience. I will call this the principle of cooperative behaviour. It entails that a speaker spends more time and effort on the pronunciation of parts of his message that are essential to recognition and comprehension than on parts of his message that are to some degree redundant (output- oriented control when necessary versus system-oriented control when permitted [14]

In this paper I will predict regularities in speech production from insights in speech perception. In some cases these predictions are corroborated by experimental evidence, in other cases evidence is controversial, in other cases again it is still lacking, and has to be filled in by others who know more than me, or by future research. I will focus on embedded words and discuss lexical redundancy, word onsets, word boundaries, and contextual redundancy.

2. LEXICAL REDUNDANCY

A monosyllable like English PAT contains little redundancy. Each of its constituting phonemes is not fully predictable from the other ones. This is different for a polysyllabic word like HIPPOPOTAMUS, that is highly redundant, in the sense that the word remains recognizable even when a few of its constituting phonemes are missing out completely.

From a listener's point of view this means that the word HIPPOPOTAMUS remains recognizable when it is rapidly and sloppily pronounced, whereas the word PAT can only be recognized correctly when it is pronounced carefully, and therefore, I assume, more slowly, so that all three phonemes are identifiable.

In general, lexical redundancy increases with word length. Therefore the need for slow and careful articulation decreases with increasing word length. From this we may predict that speed of articulation increases with increasing word length. This seems a plausible explanation of the well known phenomenon of time compression in polysyllabic words. From hereon I will take for granted that, other things being equal, monosyllabic words are more vulnerable to communication noise than polysyllabic words, and therefore liable to be pronounced more slowly and more carefully.

3. WORD ONSETS

Lexically redundant words can be recognized on the basis of a fragmentary stimulus, for example when a considerable portion of either the end or the beginning of the word stimulus is missing. Some time ago I demonstrated that words are more easily and faster recognized from initial than from final word fragments, when initial and final fragments nominally contain equal amounts of lexical information [18]. This difference appears to be related to the fact that word onsets ensure fast and proper alignment of the stimulus with word candidates, whereas word endings do not [20]. I predict that cooperative speakers spend more time and effort on onsets than on endings of embedded words, especially when the words concerned are contextually little redundant.

I have no clear and direct evidence in favour of this prediction, and would be grateful to anyone who does. Indirect evidence would be the prevalence of regressive over progressive assimilation and coarticulation on word boundaries, as seems to be case in Germanic Languages. Whether this tendency is universal, I do not know.

The relative importance of word onsets to recognition should also make it profitable for a language to have word initial stress. For English it has been argued that a listener's strategy considering each stressed syllable as a potential word onset is profitable [6]. From this it has been predicted that special word onset markers are more to be expected when a word starts with an unstressed than when it starts with a stressed syllable. This expectation was to some extent corroborated in a production study [4].

In clear speech, for example in noisy environments, informationally important words are sometimes set off by speech pauses. From the relative importance of word onsets one would predict that such speech pauses are more liable to be made before than after important content words. This seems to occur in certain styles of reading aloud [8]. The evidence is easily confusing, however, because informationally important words are often at the end of phonological phrases, potentially followed by a phrase boundary marking speech pause.

4. WORD BOUNDARIES

Whatever the relative importance of word onsets and endings, word boundaries in connected speech are potentially important to word recognition. Knowing where a particular word ends, is knowing where the next word begins, and this saves an awful lot of trouble in lining up potential word candidates with the incoming signal, as is well known from problems in the machine recognition of connected speech. Real perceptual ambiguities such as **lettuce** versus **let us**, or **budget** versus **budge it** [5], however, attest to the fact that word boundaries need not (always) be clearly marked.

Yet, it seems reasonable to predict that cooperative speakers, also in fluent, perhaps even fast, connected speech, tend to make subtle, little time consuming phonetic word boundary markers that aid listeners in finding word boundaries. This was demonstrated to be correct for ambiguous Dutch two-word combinations of the type **known ocean** versus **no notion**. Excised from fast connected speech, such speech fragments were recognized correctly in a binary forced choice task well above chance (80%) [22]. 80% is not 100%, however. In the absence of disambiguating context, ambiguity remains.

Perceptual ambiguity can also arise due to assimilation and/or degemination on word boundaries, as in **hold back** being confusable with **whole back**. This type of perceptual confusion is an immediate function of global speech tempo. In rapid connected speech the **whole back**-stimulus leads as easily to a **hold back**-response as the **hold back**-stimulus leads to a **whole back**-response [16]. Apparently, optional assimilation and degemination on word boundaries is fully incorporated in the word recognition strategies employed by listeners. One expects that the actual occurrence of assimilation and degemination is a function of lexical and contextual redundancy. This has not yet been tested.

In normal conditions, long polysyllabic words often are recognized before the end of the stimulus word has come in [15], but short monosyllabic words, if not predictable from preceding context, can rarely be recognized before the end of the word, and, if embedded, are recognized only during or after the processing of the immediately following word [1]. A speech pause immediately following a monosyllabic content word would therefore be much more helpful to recognition than a speech pause following a lexically redundant polysyllabic word. This was experimentally shown to be correct [19]. From this I predict that cooperative speakers are more liable to insert speech pauses after monosyllables than after polysyllables. I do not know whether this is actually so.

5. CONTEXTUAL REDUNDANCY

An effect of contextual redundancy seems confirmed by Lieberman's famous example **a stitch in time saves nine**, used to show that highly predictable words are less clearly pronounced than unpredictable words [12]. Lieberman's findings were more recently confirmed and extended [10],[11]. However, none of these studies pulled apart the effects of contextual redundancy and accentuation.

It has been argued that accented words are processed faster by listeners than unaccented words [3],[7]. This is not always so. Terken and Nooteboom [21] showed that 'new' words are processed faster when (correctly) accented than when (incorrectly) unaccented, but 'given' words are processed faster when (correctly) unaccented than when (incorrectly) accented. From this it follows that cooperative speakers should take care to produce accents on words carrying new information, but not on words carrying given information.

The issue is somewhat more complicated, however, because of the phenomenon of 'unit accentuation' [2]. A word group like **french cheese** can be marked as informationally important or as carrying new information, by a single accent on **cheese**. The word **french** is then 'new' but not accented. The principle of cooperative behaviour predicts that it will nevertheless be more carefully and more slowly pronounced than when it is 'given'. This prediction is falsified by Eefting [9], who, taking advantage of the phenomenon of unit accentuation, varied contextual redundancy and context-induced accentuation independently in a production study with read aloud text. She found that, other things being equal, accented words are considerably longer than unaccented words, but unaccented 'new' words are not significantly longer than unaccented 'given' words. Apparently sometimes the consequences of cooperative behaviour on one level, in this case the level of the word, are suppressed by the consequences of cooperative behaviour at another level, in this case the level of accent patterns, the correct realization of which constrains temporal patterning.

6. CONCLUSION

In this contribution I have focussed on the level of the word, because of my belief that the struggle between cooperative and self-indulgent behaviour of speakers will be most apparent where form and meaning come together. In quite a few cases evidence was found for systematic variations in word production due to the alternation between output- and system-oriented control.

Yet, the most striking result of this enterprise is in my own judgement that these variations are often relatively small and unimportant, at least when they are not supported by or drawing on conventionalized rules or structures of the language. Of course, the effect of time compression in polysyllabic words is not a small effect, but does not seem to reflect spontaneous adaptive behaviour. It rather is canonized in the sense that it belongs to what native speakers know about their language: shrinking monosyllables and stretching polysyllables are perceived as incorrect [17]. Non-phonological acoustic-phonetic word boundary cues are there, but relatively small and not very effective in normal conditions. Phonotactic word boundary cues might be effective, but tend to be obliterated by regular conventionalized rules of assimilation and degemination. Contextual informativeness and redundancy, if not expressed by conventionalized and rule-governed accent patterns, have only marginal effects on pronunciation.

Summarizing, we can say that the effects discussed are real but not very impressive. It seems that for normal speech communication conditions, the tools a speaker finds at his disposal in the set of structures and rules that make up his language, are by and large sufficient for his purposes, and that he has some but relatively little need of adapting his speech in ways that are not rule-governed.

It is plausible, of course, that this state of affairs is the result of the adaptive nature of language [13]. Strong and regularly occurring adaptive behaviour of speakers is easily conventionalized and thereby becomes part of the language. If this is correct, the adaptive behaviour of speakers can be studied as a source of language change. Such adaptive behaviour should perhaps more than we have done so far be studied in less favourable but real communication conditions. But we should always be aware that the expected adaptations on one level can be severely constrained by the requirements of another level, as in the example of word durations being controlled by accent patterns.

7. REFERENCES

[1] Bard, E.G., Shillock, R.C., and Altmann, G.T.M. (1988), "The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context", *Perception & Psychophysics, 44, 5,* 395-408.

[2] Chafe, W. (1974), "Language and consciousness", *Language 50,* 111-133.

[3] Cutler, A. (1976), "Phoneme-monitoring reaction time as a function of preceding intonation contour", *Perception & Psychophysics, 20,* 55-60.

[4] Cutler, A. (1990), "Production and perception of word boundaries", paper presented at the ATR Workshop on Speech Perception and Production, held in Kyoto, november 15-16, 1990.

[5] Cutler, A. and Butterfield, S. (1990), "Syllabic lengthening as a word boundary cue", to be printed in the Proceedings of the SST-90, Melbourne.

[6] Cutler, A. and Carter, D.M. (1987), "The predominance of strong syllables in segmentation for lexical access", *Computer, Speech and Language, 2,* 133-142.

[7] Cutler, A. and Foss, D.J. (1977), "On the role of sentence stress in sentence processing", *Language and Speech, 20,* 55-60.

[8] De Rooij, J.J. (1979), *"Speech punctuation"*, unpublished doctor's thesis, Utrecht.

[9] Eefting, W. (1991), "The effect of "information value" and "accentuation" on the duration of Dutch words, syllables and segments", *Journal of the acoustical Society of America, 89, 1,* 412-424.

[10] Fowler, C.A. and Housum, J. (1987), "Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction", *Journal of Memory and Language, 26,* 489-504.

[11] Hunnicut, S. (1985), "Intelligibility versus redundancy-Conditions of dependency", *Language and Speech, 28,* 45-56.

[12] Lieberman, P. (1963), "Some effects of semantic and grammatical context on the production and perception of speech", *Language and Speech, 6,* 172-187.

[13] Lindblom, B. (1989), "Phonetic invariance and the adaptive nature of speech", *In: "Working Models of human Perception", edited by B.A.G.Elsendoorn and H.Bouma,* Academic Press, London, 139-173.

[14] Lindblom, B. (1990), "Explaining phonetic variation: a sketch of the H&H theory", *In: "Speech Production and Speech Modelling", edited by W.J.Hardcastle and A.Marchal,* Kluwer Academic Publishers, The Hague, 403-439.

[15] Marslen Wilson, W. (1980), "The temporal structure of spoken language understanding", *Cognition, 8,* 1-71.

[16] Menert, L. (1989), "Perceptual ambiguity as indicator of voice assimilation", *In: "OTS-Yearbook 1989", edited by P.Coopmans, B.Schouten, W.Zonneveld,* 63-73.

[17] Nooteboom, S.G. (1973), "The perceptual reality of some prosodic durations", *Journal of Phonetics 1,* 25-46.

[18] Nooteboom, S.G. (1981), "Lexical retrieval from fragments of spoken words: beginnings versus endings", *Journal of Phonetics, 9,* 407-424.

[19] Nooteboom, S.G., Scharpff, P. and Van Heuven, V.J. (1990), "Effects of several pausing strategies on the recognizability of words in synthetic speech", *Proceedings of the First International Congress on the Processing of Spoken Language,* Kobe, Acoustical Society of Japan, 385-387.

[20] Nooteboom, S.G. and Van der Vlugt, M.J. (1988), "A search for a word-beginning superiority effect", *Journal of the acoustical Society of America, 84, 6,* 2018-2032.

[21] Terken, J. and Nooteboom, S.G. (1987), "Opposite effects of accentuation and deaccentuation on verification latencies for Given and New information", *Language and Cognitive Processes, 2,* 145-163.

[22] Quené, H. (1991), "Acoustic-phonetic cues for word segmentation", submitted for publication.

PHONETICS IN THE NEXT TEN YEARS

Patricia A. Keating

Department of Linguistics, UCLA, Los Angeles, CA, USA

ABSTRACT
One perspective on trends in research and training in phoentics -- that of an American linguist -- is presented.

1. INTRODUCTION

Our charge in this session is to discuss the immediate future of phonetics: what will be the main topics of research, what advances will occur, how the field will change, and similar questions. When one tries to talk about the future, it is hard not to simply survey the present, and assume that basically the same will continue. At best, one finds small, recent changes, and extrapolates from these to infer large, future changes. These limitations will hold in my paper. Also, as an American and a linguist, I will have more to say about phonetics in America and phonetics as part of linguistics; I will trust to my colleagues in this session to better represent other perspectives. (See also Kuhl [2] for yet another view.) However, to broaden my horizons, I circulated a questionnaire asking for thoughts from other phoneticians. Although this paper will not consist of a systematic report on responses to that inquiry, the responses nonetheless influence my ideas here, and a list of respondents is thanked below. A second influence has been my service as phonetician/phonologist on the U.S. National Science Foundation Linguistics Program advisory panel. Like the survey, this position gives me a wider perspective on what other individual researchers take to be the most interesting and promising areas and methods of research in the immediate future.

2. CHANGES IN WHAT PEOPLE DO AND HOW THEY DO IT

2.1. Introduction

Work in phonetics is in part driven by the development and availability of new technologies. This is very much the case today, in several ways.

2.1.1. Speech Production. First, outside of industry there is a current great enthusiasm for work on speech production, especially articulator movements, due to the X-ray microbeam facility and to the various magnetometer systems, ultrasound, and other emerging possibilities. These devices are now available primarily in a few specialized labs, but magnetometers and ultrasound may come to be found in other well-equipped labs (as are the Selspot and other movement-tracking systems). The interest of these systems is that they provide data over time which is easily processed and interpreted, and thus they fit in with the study of speech dynamics. I will rely on Professor Fujimura for more extensive discussion of this topic.

2.1.2. Databases. Second, especially outside of academia, there is a growing use of large speech databases for work in acoustic phonetics and speech recognition. Virtually all industry phoneticians mention this as an important factor for the direction of future work. Such databases, as a consequence of cheaper computer memory and faster access devices, will become more common outside of industry as well, and more of them will be assembled, particularly in Europe. This development in the availability of large corpuses of speech is symbiotic with an interest in natural, connected, speech, which I will discuss below: with carefully annotated databases, analysis of large numbers of instances of any kind of acoustic phonetic event is possible. Researchers then need not limit their studies to minimal pairs or otherwise carefully controlled wordlists. However, it must be noted that the databases available today are not typically based on natural speech, nor do they span a range of styles and genres. While it may be hoped that large speech databases will in the future contribute to the study of more natural speech, much careful thought is needed to plan databases for this purpose.

2.1.3. Computers. Third, the range of acoustic analyses available outside of the major research labs is vastly improved. Ten years ago a DEC PDP-11 computer (standard in linguistics department phonetics labs) provided one user at a time access to not-always-flexible DSP, acoustic analysis, and synthesis programs. A spectrograph provided spectrograms and power spectra, again to a single user. The typical phonetics research lab would have both kinds of machines, and non-specialist departments would have just the spectrograph. Today, these two machines have merged in function, and any department or lab acquiring a "spectrograph" can have easy access to a wide range of fast acoustic analysis. But better yet, personal computers provide other, even less expensive, means for acoustic analysis, thus allowing multiple users and even portable labs. It is possible to have an IBM- or Macintosh-based system that performs most or all of the functions of the DEC machine of ten years ago. Even five years ago these were not commonly used for speech work in N. America (Keating and Anderton [1]), but now they are being acquired even by non-specialist departments. Another factor here is the availability of several commercial packages for acoustic analysis for these machines. One further change to be expected in this regard is that the availability of inexpensive processors and grey-scale printers will make computer-generated spectrograms the norm. This development will feed another development to be discussed below, the growth in acoustic phonetic description. New computers do not just make the old phonetics lab less expensive. They also make new applications available to more users. In addition to large speech databases, these include digital imaging techniques, multi-channel data collection, articulatory modelling and synthesis, especially complex models of the vocal cords and tongue, and parallel distributed processing/neural networks.

2.2. Theory-Driven Change

2.2.1. Speech Production. It seems clear that right now the most energy and enthusiasm in academic phonetics, at least, is focused on speech production studies, and this situation is likely to continue for at least a few more years. The contribution of technology to this effort was already mentioned. Another contributing factor would appear to be some waning of interest in the dominant approaches to speech perception from the 1970s. There are two theoretical driving forces as well. Some research is based in a concern for more general biological properties of movement and motor coordination (especially at Haskins Labs). Other research seeks to provide a full description of speech, including time-varying properties. There is relatively little information at hand about time-varying properties of articulation, and this creates a need for many new studies. The one problem I see for this area of research is that it is in conflict with the goal of moving away from "laboratory speech", towards the study of more natural speech, on which more below.

The theory and techniques developed as part of a dominant research trend generally have a "trickle-down" effect. By this I mean that work on normal adults is carried over into the areas of e.g. speech development and clinical research. We can expect this to occur in the present case of speech production. Thus some responses to my questionnaire from linguists trained in speech production research mentioned these as promising areas for future work. There have always, of course, been speech production studies of children and clinical populations, given the fact that many researchers in speech production are in Speech and Hearing departments. In the future, however, such studies should be more common. These will complement the acoustically-oriented studies of the 1980s.

2.2.2. Speech Perception. The remarks above on speech production are not meant to imply that speech perception has been abandoned. One trend is to discuss speech perception in terms of more general models of the auditory system; another is to discuss it in terms of more general models of psychological category formation. It seems reasonable to expect that these two trends will be brought together in the near future, so that the physical and psychological aspects of perception will be given equal attention. A further topic of interest is the use of dynamic information in speech and, more generally, the perception of connected speech.

2.2.3. "Real" Speech. There is growing interest in the acoustic properties of more natural speech, especially in connection with work in machine and human speech recognition and text-to-speech synthesis systems. This interest is in part a reaction against work on the perception of isolated segment contrasts, and analysis of isolated acoustic characteristics of nonsense items produced under controlled laboratory conditions. It takes several paths. One is the move away from ideal and idealized "laboratory speech". This will be a big area in the next ten years --acoustic analysis and perception of casual speech, connected speech, spontaneous (vs. read) speech; women and children, regional dialects -- the full range of speech styles and voices encountered by human and nonhuman speech recognizers. We might also see more cooperation between phoneticians and social scientists as a result.

Another path is interest in lexical access --going beyond nonsense syllables and acoustic cue manipulations to get to the task that real listeners actually use speech perception for. Here we can point to a convergence of interest between, say, Ken Stevens and William Marslen-Wilson. I think that anyone who works on representational questions will come to take lexical access more into account. And this should contribute to better interaction between phoneticians and psychologists or cognitive scientists.

A third path is interest in the prosody of connected, fluent speech. It is always said that prosody is the great problem facing natural synthesis, and also that better use needs to be made of prosodic information in recognition work. At the same time, as a result of developments over the last 10 years, there are well-established linguistic theories of prosodic structure, e.g. in the "Prosodic Hierarchy" tradition ([4], [6]), just waiting to be applied to problems of phrasing and fluent speech. I see a real opportunity

here that could have as one result improved academic industry cooperation.

At the acoustic end of research on prosody, an interesting new development is attention to the time-varying characteristics of the voice source as a function of prosodic structure. Source characteristics may be seen as yet another dynamic, coarticulating property of connected speech, determined in part by "inherent" segmental specifications, in part by prosody, and in part by speaker characteristics. It seems possible to me that the same trend might extend to another global property of utterances, amplitude: modeling the combined effects of respiratory energy, source, and articulation on this continuous function could contribute to more natural speech synthesis.

2.2.4. Phonology. Phonology was one of the "phonetic sciences" that the International Congresses were originally intended to bring together with core phonetics/ speech science, and while some of the other "phonetic sciences" are not so actively represented here today (e.g. music, sociology), phonology has always been included in the Congress. In the early days (the 1930s) phonology meant the Prague School, and relations between phonetics and phonology at the Congresses were apparently good. Relations declined when the mantle of phonology passed to generative phonology. However, the relation of phonetics and phonology seems to be cyclic, and for the past few years we have largely returned to a period generally characterized by mutual interest and cooperation. I feel confident that this will continue for at least a while because much of the new generation of linguists goes strongly in this direction. Some of the new work by these young people is hard to categorize clearly as "phonetics" or "phonology", and I take this as a good thing, generally. (See also [3] and [5].) There are two overall research programs in this area today. One might be called the "explanation" program, because its goal is to find phonetic explanations for cross-linguistic phonological patterns -- for example, for generalizations about segment inventories. This kind of research is represented in the recent special issue of <u>Phonetica</u>. The other research program might be called the "grammar" program, because its goal is to establish the phonetic structure of individual utterances in a language and the general principles that underlie it. This kind of research is represented in the recent special issue of <u>Journal of Phonetics</u>.

2.2.5. Cross-Language Research. There has been a good deal of cross-language comparative work recently. Some of this is concerned with testing phonological theories of rules or representations, but much of it tests aspects of strictly phonetic theories, such as theories of speech production. At the same time, the current funding mechanism for speech research in Europe (Esprit projects) seems designed to encourage cross-language research, because it encourages multinational collaboration. This research is likely to be more descriptively oriented.

2.2.6. Cognitive Science. In responses to my survey, no one had a strong reaction on this topic. Indeed, phoneticians are generally not in the core of this research community, despite their strong ties to perceptual and cognitive psychology, which <u>are</u> in the core, and their use of computers, sometimes seen as a prerequisite to cognitive science. Why are phoneticians not active cognitive scientists? I am not sure, but perhaps our work, by its nature, is already so interdisciplinary and has ties to so many fields that we do not need to reach out for a broader forum. Perhaps in the next decade younger phoneticians will do a better job of making the connection between phonetics, especially speech production, and cognitive science.

2.3. Other Change

2.3.1. Language Description. There is surprisingly little thorough phonetic data available on most languages. Among phoneticians, linguists have this responsibility; yet they often work only on English. However, the increased availability of good systems for acoustic analysis should change this situation somewhat, allowing descriptive linguists who are not from major phonetic research centers to carry out more instrumental phonetic studies. Also, the revival of the Journal of the IPA as an outlet to publish descriptive phonetic work will help. Thus we should see more basic data, especially acoustic phonetic data, available on many of the more "exotic" sound types. At the same time, the development of speech databases for certain major languages e.g. in Europe will result in more thorough and comprehensive acoustic phonetic descriptions of these languages. Another development is that portable computers can be used to digitize and store audio and physiological signals, and also for acoustic analysis. Such systems make it possible to use computers for instrumental phonetic fieldwork. But it's not just that more cross-language work will be done, and that perhaps we will build up more comprehensive acoustic descriptions of certain languages. It's also that this descriptive work will be more sophisticated, in that it will be tied in more directly with basic theoretical models and with analysis-by-modeling. We are beginning to see more kinds of acoustic measures being used, especially applied to less-common sounds and sound contrasts, in the way that Ken Stevens, Peter Ladefoged, and their colleagues have been doing since the 1960s. To some extent this change follows from developments in the acoustic theory of speech production, and its application to a broader range of cases.

It is also possible that we will see changes in which languages are studied by phoneticians. At least in the U.S., languages of the Pacific, especially SE Asian languages, are taking on more importance, yet these are not much studied by phoneticians. At the other extreme, there seems to be more sensitivity among linguists generally to the fact that many indigenous languages are in or near their last generation of fluent speakers. Perhaps phoneticians will play a larger role in preserving a record of such languages.

2.3.2. Non-Phonetician Users. Another change I see happening now is that there are more of what I will call "users" of phonetic data. By "users" I mean people who learn how to use certain equipment and make certain measurements without having a more general interest in phonetics, without first having general training in phonetics, and with no long-term interest in setting up, maintaining, or perhaps even being affiliated with, a phonetics lab. The use of speech data is much more common among non-phoneticians now than before, I think. First, psycholinguists who do not study speech per se are increasingly interested in working in an auditory modality. This means they use speech signals to create their experimental stimuli. This is a striking aspect of the speech lab set-up at the Max-Planck-Institute for Psycholinguistics, for example. Second, phonologists and linguists in language-area specialties seem ready to rely more on instrumental data at times where the ear alone, especially the untrained ear, is not reliable. For example, it is once again becoming the norm to see f0 traces in essentially phonological studies of intonation or of tone. (Use of instrumental data by phonologists is generally cyclic. A facilitating factor these days is the familiarity of personal computers to non-phoneticians, which makes the phonetics lab less threatening than it once might have seemed.) Third, second language researchers seem to be increasingly interested in acoustic analysis as a tool for contrastive analysis and for assessment. (This also will contribute to language description and to cross-language work.) As a result, researchers in all of these areas will want access to phonetics instrumentation. My impression of this situation is that laboratory phonetics is maturing as a "service" discipline, exactly parallel to the way that traditional impressionistic phonetics has become a service discipline. (I say this despite the fact that phonetic transcription is a theoretical construct, as Ladefoged has recently made great efforts to stress.) The use of phonetics lab equipment for speech data analysis or manipulation is a service that phoneticians offer to other phonetics "consumers". The good side of this is that more people get involved in phonetics, bringing more data and new issues to the phonetician's attention. The bad side of this is that non-phoneticians will come to think that this is the whole of instrumental phonetics. They may thing that, just like impressionistic auditory phonetics, instrumental phonetics is an essentially non-theoretical discipline where everything has already been worked out, and where the only remaining challenges are descriptive. They may also think that qualifications in phonetics consist mainly of technical training. Such an impression of phonetics among non-phoneticians would not be good for our field. For example, the best students would tend not to choose to go into phonetics.

2.3.3. Historical Phonetics. My own impression is that there is renewed interest among younger phoneticians in the area of sound change and historical phonetics, but not yet any unifying themes or strong theoretical innovations in their work. Generally this interest springs from work in particular language families or areas. Not only might we hope for increased phonetic sophistication in theories of sound change, but data from sound change might be expected to provide test data for general phonetic theories.

2.3.4. Funding. It seems clear that phoneticians able and willing to work with engineers have a much better time in the funding situation. The same obviously applies to phoneticians interested in more applied kinds of work. Our European colleagues benefit financially from collaborations with industry, but caution us about the problems that arise (lack of freedom to do basic research, and bureaucratic/ administrative problems). Many (American) academic respondents to my questionnaire already complain that speech technology seems to be driving the field more than they would like. Almost everyone in academia responding to my questionnaire thought that funding considerations were going to become more important in the future, that is, that they would have to choose their topics based more on considerations of funding possibilities than they have in the past.

3. CHANGES IN EMPLOYMENT, ETC.

3.1. Retirements, Hiring Patterns

Although all the new reports assure us that academic jobs will be booming in this decade, and although there is more hiring going on in U.S. universities, a disturbing pattern is that there has been no increase in long-term jobs for new phonetics Ph.D.s in linguistics. Linguists are going to Speech and Hearing departments (which seems to be a genuine growth area for linguists just now), and to industry. The situation in Europe, Australia, etc. is also not especially encouraging, and essentially jobs are scarce everywhere. Perhaps we faculty could help our students by doing more -- or leaning on these very students to do more! -- to integrate phonetics better into cognitive science, since that does seem to be an academic growth area. Another strategem is to convince phonologists in more departments that they would benefit from having a phonetician as a colleague, while also pointing out that a reasonable phonetics lab can be started for less money than might have been expected. The other thing we can do is to call more attention to the availability of absolutely top-rate women and minority phoneticians; a university can do well in its affirmative action programs by expanding in phonetics.

3.2. What Students Should Be Studying For the Future

In my questionnaire I asked respondents to advise current students on how to prepare for their futures, including what they themselves look for in hiring. The responses were quite revealing, and can be summarized as follows. Most important for phoneticians generally is a solid ground in acoustic theory and speech acoustics. Also important are basic knowledge of experimental design and statistics and user skills in acoustic analysis and synthesis. The more computer experience a student has, the better, even in academia; programming in C and in an AI language are recommended for industry. Some background in linguistics is helpful for industry, and substantial background in phonology is helpful for academic lingustics departments. Another recommendation is experience with modeling and simulation; it reinforces knowledge of theory and demonstrates computer skills. Finally, students interested in an industry career should try to arrange an internship or some other kind of student job at a company or industry-related research center. My hope is that students can use this report as a guide in considering possible directions for their own research or future careers. There are many exciting research questions to choose from, most of which can be directed to either academic or industry careers. The classic advice, to find an interdisciplinary niche, remains relevant today for both academia and industry. Students should also think about how theoretical developments from various areas might translate into applications, or contribute to progress in applications-oriented research. Students should not wait for industry to realize that their skills are needed, because they themselves can help create the market for those skills.

4. References

[1] KEATING, P. and ANDERTON, A. (1987), "Survey of phonetics computers in North American laboratories", UCLA Working Papers in Phonetics, 66, 104-123.

[2] KUHL, P. (1990), "Hot topics in speech communication", J. Acoust. Soc. Am. 88, Suppl. 1, S69.

[3] LIBERMAN, M. (1983), "In favor of some uncommon approaches to the study of speech", The Production of Speech, ed. P. MacNeilage, New York: Springer-Verlag.

[4] NESPOR, M. and VOGEL, I. (1986), Prosodic Phonology, Dordrecht: Foris Publications.

[5] OHALA, J. (1990), "There is no interface between phonology and phonetics: a personal view", J. Phonetics, 18, 153-272.

[6] SELKIRK, E. (1980), "Prosodic domains in phonology: Sanskrit revisited", Juncture, ed. M. Aronoff and M.-L Kean, Saratoga: Anma Libri.

5. ACKNOWLEDGMENTS

Thanks to those who responded to my survey: S. Baum, M. Beckman, P. Beddor, S. Boyce, C. Browman, G. Bruce, F. Chen, K. Denning, F. Ganon, J. Gilbert, S. Hawkins, C. Henton, M. Huffman, M. Jackson, H. Javkin, K. Johnson, R. Krakow, P. Ladefoged, I. Maddieson, J. Ohala, B. Poser, P. Price, B. Repp, C.-Y Tseng, M. Withgott and a couple of unidentified respondents. Thanks also to Louis Goldstein and to members of the UCLA Phonetics Lab.

RECORDING AND INTERPRETING ARTICULATORY DATA-- MICROBEAM AND OTHER METHODS

Osamu Fujimura

The Ohio State University
Division of Speech and Hearing Science
Columbus, OH 43210-1002, U. S. A.

ABSTRACT

Keating's paper "Phonetics in the Next Ten Years" is discussed with two foci: (1) instrumentation in speech production research and (2) the phonetics-phonology interface. These comments supplement two recent review articles by the author. A contributed paper in this conference by the author and his colleagues provides an alternative theory of phonetic implementation.

1. DEVICES FOR SPEECH PRODUCTION RESEARCH

Upon Keating's invitation, I shall discuss instrumentation issues in speech production research in the coming years in the context of a new emphasis on speech dynamics. Obviously I am biased toward the use of the x-ray microbeam for speech studies, because I have spent about 30 years proposing, designing, implementing and using the system. At that time, it seemed to be the only feasible method for observing the dynamics of speech production with little disturbance to the subject, in conformity with appropriate safety considerations. It took us a long time to make the system widely available; the U. S. Government (National Institute of Health) implemented a special support for the nationally shared research facility at the University of Wisconsin (Principal Investigator: J. H. Abbs) ten years ago. The system has now started producing systematically controlled articulatory data. A wide use of the system in full operational capacity is now foreseeable. In the mean time, there are some discussions of the possibilities of implementing similar facilities in Europe and in Japan.

Recently, two additional devices for observing tongue movements in speech have emerged. Both are still in a developmental stage, but are indeed promising, and may well be close to productive research use. I do not believe, however, that new methods will entirely replace the x-ray method in the next ten years. The advantages and disadvantages of all these methods shall be discussed below in some detail. My general view is that speech production research is so complex and difficult that no single method fulfills the purpose. Different approaches, including instrumental, theoretical, and computational, must be exploited as much as possible. For example, despite remarkable progress in technology, the traditional film analysis of cineradiographic images of the tongue contour still provides very useful information, particularly as a supplemental means for interpreting articulatory data obtained otherwise. The palatographic methods, now computerized for dynamic observations, will continue to provide three-dimensional information which most other methods cannot provide, at least as easily. There is much to be done in devising and combining different methods of articulatory observation and measurements. For instance, direct evaluation of muscle contraction states is one largely unexplored area, electromyography being severely limited in its applicability.

On the other hand, we have already seen much progress in our understanding of speech production. As Keating mentions, research has been progressing quickly with very important changes of direction.

While research results obtained by the use of ultrasonic and magnetic devices are only preliminary at this stage, so too are those from the microbeam studies, despite a much longer history. I hope this will be no longer true in a few years. Examples of good speech production research using new observational methods convince us of the critical importance of this type of research for understanding what speech is and how it is organized.

I have discussed this issue fairly extensively in a recently published review paper [5]. Therefore, I shall only add some remarks which have come to my attention since the time I wrote that paper. It is my understanding at this time that the magnetic method is particularly promising as a competitor to the x-ray microbeam. However, I still remain to be convinced with regard to its reliability and facility of use. Schönle's method [15] is even commercialized, and I think the device could have wide application in clinical and pedagogical areas. Perkell's method [13] seems to be producing some early research data. I would not be surprised at all if the magnetic device, quite possibly along with the palatographic device which already has proved remarkably successful in practical situations [15], becomes widely used in training deaf children to speak. However, my information is that, as a rigorous research tool for exact measurements of flesh point positions on the tongue, the system needs substantial improvement. At the time of the conference, perhaps, more convincing demonstrations may be given.

In principle, magnetic methods are superior to x-ray methods in that they do not involve any ionizing disturbances to the human body at all, and that the interference of metal objects in the mouth should be less severe. The current version of Perkell's system seems to be using very small coils, equivalent to the pellets for sampling flesh points on the tongue, probably at a severe cost of signal to noise ratio in the position determination. With regard to the microbeam, the pellet can be made smaller, also at some cost of detection reliability. The gold pellet being used at the University of Wisconsin facility is typically 2.5 mm in diameter. We should be able to reduce the size and weight considerably, possibly deviating from the spherical shape which we use now. The size issue may sound trivial, but it has major consequences. For example, if we reduce the diameter from 2.5mm to 2.0mm, keeping the same shape, the weight of the pellet is just about halved. Such a reduction in weight can bring about a qualitative difference in the usefulness of the system. In particular, the force required to retain the pellet in position will be reduced accordingly, and the choice of adhesive materials will be much easier, since the surface area is not reduced in the same proportion as the volume or weight. Also, it is quite possible that we can dispense with the string attached to each pellet for safety precautions. The chance of detachment becomes smaller partly because of less protrusion of the object embedded on the tongue surface (also causing less articulatory disturbance), and also partly because of the reduced inertia reacting to large acceleration of the tongue surface. Probably more importantly, the.chance of inspiration of a pellet, once detached, will be reduced substantially when the weight is reduced, because of the reduced gravity force relative to surface tension.

Currently, from a subject's viewpoint, the most disturbing aspect of the use of pellets is the string. If we can dispense with the string entirely, or if we can limit the use of the string to a short segment of string attached to the immediate surrounding area of the tongue surface, as opposed to the attachment outside the mouth on the facial skin as we do now, most of the disturbance will disappear except when we try to fix the pellet at the very tip of the tongue. This possibility is a distinct advantage of the x-ray method over the magnetic method. The magnetic method requires not just a retaining string, but electrically conductive lead wires due to its very principle (unless a small radio transmitter with a small battery together with the coil can be made comparable in size).

Ultrasonic methods have the distinct advantage of wide use in general medical diagnosis with consequent availability of commercially developed apparatus. The spatial and temporal resolutions available by this method, however, seem not quite satisfactory from our research point of view. The strongest concern I have, assuming some devoted effort to optimize details of the apparatus for the specific purposes of speech research, is the wave reflection at the boundary between flesh/water and air. After writing the review paper mentioned above, I have communicated with Dr. Stone, who recently published a new article on tongue movement [16], discussing some technical details of her method; I still maintain the same opinion about this point. It is inherently difficult for the ultrasonic method to avoid interference with the movement of the lower surface of the mandible, which easily deforms according to the force applied by the contacting device, even when a highly compliant material is used as an impedance matching medium placed between the solid and the skin. When the skin moves, the internal tissue becomes displaced, resulting in possibly quite significant distortion of the three dimensional shape of the tongue surface. This is particularly difficult in a dynamic situation, where the transmitter can not move according to the movement of the mandible (or even muscle contraction inside). As I mentioned in my paper, it is technically possible to avoid this problem by using a servomechanism in order for the device to follow exactly the moving skin with minimal force. Such a method apparently has been studied by Müller and his colleagues many years ago [12] in lip movement measurement, but I have no followup information about this interesting attempt. Servomechanism can be extremely effective in such devices, but requires very advanced engineering involving highly mathematical analyses. For ultrasonic methods to be reliably useful for general articulation research, however, I think it is necessary to resort to such advanced techniques.

The current method of computer-controlled x-ray microbeam for pellet tracking uses an extremely small radiation dose, thus making it possible to use the same subject for rather extensive speech material. As research makes progress, and the scope of study expands, from the main focus on robust segmental characteristics to more general issues of speech organization and principles of speech organization beyond the minimal scope of segmental concatenation, we need more and more data, with a great number of the factors of speech utterance under control. The data, to a large extent, must be collected from the same subject; yet subject-to-subject variation requires still larger amounts of data obtained from many subjects. Given more and more powerful data processing computational tools in combination with efficient and hazardless acquisition methods, the amounts of data we should use are expanding rapidly. This is probably the most remarkable change we have seen in speech production research. Until just recently, an articulatory study typically involved one to three speakers' data for about 100 seconds worth net total of speech materials each. Now, with the microbeam facility, we count on having several to even twenty speakers, each a total net (with convenient breaks between utterances) of typically 1,000 seconds per speaker in one session. As the data processing/interpreting methods advance, using for example neuronetwork and other AI type processing techniques, backed up by inexpensive computer memory, substantially more data can be effectively used for our studies without much difficulty. In this context, it is worthwhile to reexamine the radiation dose problems, which once appeared almost completely solved. One approach is to look at non-ionizing measurement methods, perhaps even at the cost of high accuracy of position assessment. There is, however, still some room for further improvement in this respect with regard to the x-ray microbeam method also.

One such innovation, which I have proposed in connection with the Wisconsin microbeam project, is the use of scattered photons within the scheme of x-ray microbeam. The current method attempts to capture all the photons being transmitted through the subject's tissues in the area immediately surrounding each pellet, thus optimizing the use of the radiation dose by not wasting any useful photon that comes through to the detector. The scintillation counter as the photon detector is almost ideal in capturing penetrating photons. What about the photons that do not penetrate, then? Those are not "useful" photons, from this point of view, but still cause ionization within the body. In fact, it is the non-penetrating photons, rather than the penetrating photons, that form potential harm to the body. Fortunately, a large portion of those photons are absorbed by the metal pellet. A significant portion of the photons that hit the pellet, however, are scattered. Some of these scattered photons, particularly those with characteristic energy of the metal used as the pellet, can be detected and identified as coming from the pellet. This would constitute a positive identification of the location of the pellet, rather than the current negative method which identifies the shadow of a pellet. In order to make use of these photons, we need a supplemental detector which covers a large solid angle other than the directly penetrating direction of the microbeam. I expect this new method to enhance the accuracy and reliability of the pellet identification, further reducing the necessary dose. An additional advantage is that with an appropriate choice of pellet material, interference from metal elements in the mouth due to dental work will be eliminated. Particularly if one side of the mouth is free of gold and other heavy metals in the useful area of the head profile, this method may prove critically helpful for the microbeam method, retaining its high accuracy and reliability.

2. MODELS OF PHONETIC IMPLEMENTATION

Keating, in Section 2.2 of her paper, discusses theory-related issues, emphasizing strongly the emerging importance of speech production studies in this context. There have been many discussions and advocations of new ways of relating phonology to phonetics, and some insights have been acquired indicating the future direction of research in this area. From my point of view, however, we still are short of any explicitly formal phonetic theory in general phonetics. The only complete (though still vague in many ways) theory of phonetic implementation is segment concatenated-smoothing (coarticulation theory basically since Lindblom, [9.10.11].see also, Fujimura [6]). On the other hand, the theory of nonlinear phonology has declared a basic departure from the segmentalism of this approach, but it is still not clear what this departure means. I suspect this issue will not be clear until we know exactly what the alternative, or supplement, to the coarticulation theory is going to be, or even could be. Browman and Goldstein [1,2] have proposed a very interesting idea, which seems radical enough to achieve the needed change; unfortunately, the picture is only conceptual and remains very vague, and many difficulties including some apparent internal inconsistencies must be resolved [3,7].

I have discussed some of these issues in my recent articles [4,5]. With my colleagues at OSU, I am currently engaged in designing a new model of the phonetic implementation process as a comprehensive (but rough and tentative) quantitative model, relating phonological and other specifications to articulatory control organization and then to acoustic signals. A contributed paper in this conference provides a sketch of this theory. This model is being used for an application of the abduction method [8], to automatically interpret microbeam pellet data of prosodic control [17,18]. Hopefully, some specific details of the model can be determined by this abduction data processing; at the same time, an empirical validation of the theory will be provided. Such studies crucially require a large database of articulatory recordings from natural utterances from systematically controlled speech materials covering a large number of factors, both phonological and extraphonological. Hopefully, as Keating suggests, theory-driven experimental work with extensive articulatory data, including this approach, will pave the way to understanding the linguistic and paralinguistic organization of natural speech.

3. REFERENCES

[1] Browman, C.P. & Goldstein, L. (1985). Dynamic modeling of phonetic structure. In V.A. Fromkin (ed.)*Phonetic Linguistics--Essays in Honor of Peter Ladefoged* (pp. 35-53), New York: Academic Press.

[2] Browman, C.P. & Goldstein, L.

(1989). Tiers in articulatory phonology, with some implications for casual speech. In J. Kingston and M.E. Beckman (*eds.*) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 341-376). Cambridge: Cambridge University Press.

[3] Fujimura, O. (1989). Towards a model of articulatory control: Comments on Browman and Goldstein's paper. In J. Kingston and M.E. Beckman (*eds.*) *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (pp. 377-381). Cambridge: Cambridge University Press.

[4] Fujimura, O. (1990). Articulatory perspectives of speech organization. In W.J. Hardcastle & A. Marchal (*eds.*) *Speech Production and Speech Modelling* (pp. 323-342). Dordrecht: Kluwer Academic Publishers.

[5] Fujimura, O. (1990). Methods and goals of speech production research, *Lang.& Speech* **33**, 195-258.

[6] Fujimura, O. (1991).Comment: Beyond the segment. In I.G. Mattingly & M. Studdert-Kennedy (*eds.*) *Modularity and the Motor Theory of Speech* (pp.25-31). Hillsdale, N.J.: Lawrence Erlbaum Assoc. Publishers.

[7] Fujimura, O. (in press). Comment on Beckman-Edwards- Fletcher's paper. In Docherty & D.R. Ladd (*eds.*) *Laboratory Phonology II*. Cambridge: Cambridge University Press.

[8] Josephson, J.R. (1987). A framework for situation assessment: using best-explanation reasoning to infer plans from behavior. In *Proceedings of Expert Systems Workshop* (pp. 76-85). San Diego, CA: Science Applications International Corporation (SAIC-87/1069).

[9] Lindblom, B. (1963). Spectrographic study of vowel reduction, *J. Acoust. Soc. Am.* **35**, 1773-1781.

[10] Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. J. Hardcastle and A. Marchal (*eds.*), *Speech Production and Speech Modelling* (pp. 403-439). Dordrecht: Kluwer Academic Publishers.

[11] Lindblom, B. (1991). The status of phonetic gestures. In I.G. Mattingly and M. Studdert-Kennedy (*eds.*) *Modularity and the Motor Theory of Speech Perception* (pp. 7-24). Hillsdale, N.J.: Lawrence Erlbaum Assoc. Publishers.

[12] Müller, E., Abbs, J.H., Kennedy, J., &Larson, C. (1977). Significance of perioral biomechanics to lip movements during speech. *Paper presented to the American Speech and Hearing Association, Chicago.*

[13] Perkell, J.S. & Cohen, M.H. (1986). An alternating magnetic field system for tracking multiple speech articulatory movement in the midsaggital plane. *Technical Report 512. Research Laboratory of Electronics, MIT.*

[14]Schönle, P.W. (1988). *Electromagnetische Artikulographie. Ein neues Verfahren zur klinischen Untersuchen der Sprechmotorik*, Berlin: Springer-Verlag.

[15]Shibata, S., Ino, A., & Yamashita, S. (1979). *Teaching Articulation by Use of Electro-Palatography.* (English Translation by K. Kakita, H. Kawasaki, & J. Wright, *ed.* by S. Hiki, 1982), Tokyo: Rion Co., Ltd.

[16] Stone, M. (1990). A three-dimensional model of tongue movement base on ultrasound and x-ray microbeam data. *J.Acoust.Soc.Am.* **87**, 2207-2217.

[17] Westbury, J., & Fujimura, O. (1989). An articulatory characterization of contrastive emphasis in correcting answers. *J Acoust.Soc. Am.* **85**, Suppl. 1, S98.

[18] Westbury, J. and Fujimura, O. (1990).Articulatory correlates of contrastive emphasis in correcting answers in English. *ATR Workshop in Speech Production and Perception, November 1990, Kyoto, Japan.*

PHONETICS IN THE NEXT TEN YEARS

Francis Nolan

Department of Linguistics, Cambridge University, UK

ABSTRACT

This paper deals selectively with a few themes which I think will be of importance. It begins by suggesting that phonetics shows signs of renewed identity as a discipline, and arguing that this has important intellectual consequences. Section 2 deals briefly with the relation between phonetics and phonology, which will continue to become closer. Section 3 predicts increasing interest in the principles of traditional phonetic description as a consequence of developments elsewhere in the phonetic sciences. Next, expected changes in the kind of data which phonetic research focuses on are dealt with in section 4. Finally, it is suggested there will be greater recognition of phonetics as having a wider purview than the phonic realisation of language.

1. A DISCIPLINE AGAIN?

One of the major intellectual attractions of phonetics is its readiness to draw on, and contribute to, other disciplines. At best, this gives the phonetician the opportunity to aspire to be a latter-day "Renaissance man" – or woman – a polymath who is at home in any branch of learning. The reality of this interdisciplinarity may be a little less impressive, according to Peter Ladefoged in his opening address to XIth International Congress of Phonetic Sciences in Tallinn [9]: he cited the camelion-in-reverse behaviour of the phonetician who, when among specialist acousticians, professes to be rather more a kind of psychologist, and when among psychologists, speaks with greatest authority on matters of vocal anatomy – and so on. In the English proverb, "the jack-of-all-trades is master of none", but here I will deal not with problems which arise from phoneticians' incomplete mastery of other "trades", but rather with whether they lack a trade of their own, and if so, what the consequences are.

One can imagine phonetics lying at the hub of a wheel, with spokes linking to a large number of disciplines on the rim – linguistics, psychology, acoustics, pedagogy, forensics, and others. The question then arises as to whether phonetics constitutes a discipline in its own right, or whether it is merely an umbrella term for those parts of other disciplines which happen to deal with speech, or perhaps a name for a node through which information flows between them – a kind of intellectual telephone exchange. Such existential doubt among phoneticians has been promoted on the one hand by institutional factors – in the UK at least, a process of absorbing the independent phonetics departments of the first half of the century into linguistics, or other, departments, continues today; and on the other hand by scientific factors – it is clear that today's phonetician is crucially dependent on collaboration with those disciplines which develop technology, and experimental methodology.

Disciplinary divisions have no intrinsic merit. Does it matter then if phonetics is merely a handy label for various disciplines' speech work? I think it does. Viewing phonetics as merely the intersection of others' domains can lead to an abdication of responsibility, particularly in the discussion of theories and principles. Phonetics in the recent past has been characterised by the borrowing, albeit often profitably, of theories from elsewhere; and, more worryingly, perhaps, by a relative lack of concern with questioning and developing its own basic principles, which have changed little in the course of a century. This latter issue I shall deal with in section 3 below; first, I shall consider briefly the borrowing of theories.

In speech production, for instance, the direction of research through the 1980s was stongly influenced by theories taken from the field of motor control – Action Theory and Task Dynamics, for instance [4,7]. In speech peception, the concept of neurophysiological Feature Detectors stimulated much research from 1971 onwards [1]; and, more recently, the philosophy of Direct or Ecological Perception from the visual domain has been adapted for speech (e.g. [3]). Perhaps, too, the relative success in speech technology of general pattern processing strategies such as Hidden Markov Modelling, and to a smaller extent Neural Networks, in tackling the task of speech recognition will lead increasingly to their adoption into the phonetician's conceptual armory. Such borrowing enriches phonetics; but it has to be asked whether it perhaps leads to a neglect of fundamental theoretical development internal to phonetics.

It could be argued that the Motor Theory of speech perception [10,11] provides an instructive alternative model of progress. The search for explanations to puzzling phonetic facts about the absence of invariant acoustic cues to sounds, and to the non-linear relation of perceived phonetic categories to acoustic continua, led to a specifically phonetic hypothesis. I cite this not as a proponent of the Motor Theory; nor to argue for the position it entails that "speech is special". Rather I see it as evidence that the speech researcher, the phonetician, who more than anyone else is the person with the advantage of an all-round (if imperfect) view of the different aspects of speech, can alone contribute a particular kind of theory, a kind without which the phonetic sciences would be impoverished.

What has this to do with the next ten years? There are signs, I think, of an increasing self-confidence, and sense of identity, among phoneticians. The high level of activity generated by the 1989 IPA Convention in Kiel (of which more in section 3), the renewed vigour of the *Journal of the International Phonetic Association* (which Pat Keating mentions in her paper in this session), the devotion of a wide-ranging theme issue (18/3) of *Journal of Phonetics* to the notion of "phonetic representation", are just a few of the indications that phoneticians have rediscovered a sense of identity, and realised that they need not aspire merely to be "a kind of" engineer, psychologist, or linguist, but that there is an identifiable core of problems, and solutions, to be researched which are specifically "phonetic".

A consequence of renewed intellectual identity will be, I hope, that phoneticians will increasingly seize the initiative in defining research questions, and specifying the criteria which answers must satisfy.

2. PHONETICS & PHONOLOGY

One of the clearest innovations of the last decade has been the emergence of empirical work which blurs, both conceptually and methodologically, the distinction between phonetics and phonology. A range of such work has been brought together under the heading of "Laboratory Phonology" (e.g. [8]).

Some strands remain close in spirit to one aspect of existing experimental phonetics, concentrating on testing in the physical domain the predictions of linguistic phonological models. Others challenge traditional notions of phonological representation: Browman and Goldstein [2], for instance, propose as phonological primes the "gestures" of a dynamic articulatory model. At the same time phonologists are exploring the value of replacing the familiar rules and representations with radical alternatives, which superficially at least look susceptible of direct quantitative phonetic interpretation – for instance Goldsmith [5], who proposes replacing the "metrical grid" with a "connectionist" model in which the "activation levels" of syllables are computed arithmetically from an initial value and lateral inhibition between adjacent syllables.

In one possible view, the interface of phonetics and phonology is where, traditionally, discrete, symbolic, descriptions meet continuous, quantitative models. There, too, are found a range of phenomena (such as coarticulation, micro-intonation, and incomplete assimilation) whose status as natural consequences of physical mechanisms, or linguistically controlled effects, is ambiguous. This interface will continue to throw up new problems, the solutions to which may be found in radical new types of model.

3. TRADITIONAL PHONETICS

For all the technological advances of recent decades, it is clear that traditional phonetic methodology, based on carefully trained listening (and looking), will remain crucially important in the next decade. This is for two main reasons. First, because in the great majority of practical applications of phonetics this is the only form of analysis available; and second, because the whole framework of phonetic classification generally used is initimately bound up with this methodology.

The purpose of the 1989 IPA Convention in Kiel was to review and update the International Phonetic Alphabet, and hence the descriptive and classificatory principles which it embodies. In many ways the Convention was a great success, stimulating wide debate both in the year or so before it and during the actual meeting, and resulting in a new IPA chart, incorporating a number of rationalisations and improvements.

However, it quickly became clear in the run-up to the Convention that the practical requirement of achieving a revised chart at the end of proceedings would override any desire for extensive debate on the fundamental principles of phonetic description. Although as convenor of the "vowel group" I included in my pre-convention survey questionnaire items concerning the principles of vowel classification [12], by the time of the Convention it was clear that no-one wished to embark on a fundamental review of descriptive principle. The same was, I think, generally true of the other groups, and so the Convention ended up approving a revised chart which is based on the same conceptual framework as previous ones.

Of course change for its own sake is not of value. Phoneticians sometimes look askance at the rapid succession of (equivalently) teddy boys, mods, rockers, hippies, skinheads, punks, and so on who have battled it out on the streets of theoretical phonology; and take pride in the stability of traditional phonetics. But arguably stability is only to be valued where it is the outcome of reasoned evaluation of alternatives. It is far from clear that such evaluation has taken place in traditional phonetics.

Recent developments in phonology should mean that debate over the conceptual framework behind phonetic classification can no longer be postponed. While phonologists argued over features versus segments, phoneticians could remain aloof – the IPA framework implicitly embodies both. While phonologists fought over the abstractness of representations, phoneticians could turn a blind eye – it was merely a matter of how the IPA categories were used or misused. But nowadays "non-linear" phonologists are proclaiming that phonologically relevant properties do not, after all, line up in discrete phoneme-sized segments. This, of course, is not news to phoneticians. Clumsy notational devices like [pʰæǣn] *pan* and [ss̬ɔːzz̥] *saws* have been used by phoneticians to highlight the tensions between the phonetic continuum and a segmentation based on phoneme-sized slices. But I suspect any serious qualms phoneticians had about the segmental principle at the very basis of the descriptive phonetic framework could be soothed by the reassuring knowledge that, ultimately, phoneme-sized chunks were what phonologists wanted, and, basically, were what speech was all about.

Not any more. The new phonologies are breaking away from strict segmentation, and it is time to ask whether the phoneme-sized units of IPA phonetic practice have independent justification, or are only a by-product of the phonological world-view with which the IPA grew up. Maybe the debate has already begun. Kelly and Local [6], in examining the data of phonological analysis, push phonetic notation beyond its usual limits, and, whilst not actually replacing segmental impressionistic phonetic representations and their interpretations, substantially augment them with descriptive mechanisms having domains other than the segment. I expect the next ten years to see a more explicit examination of the principles of IPA description in the light of phonological thinking.

It is not only phonology that has changed since the inception of the IPA. Understanding of the physical events of speech has advanced considerably, much of the progress being enabled by technological developments which have aided the study of speech production and acoustics. But again, there has been little explicit consideration of the impact which this progress might have on the basic descriptive mechanisms of the IPA. A few recent minor changes might be indirectly attributable to technology – for instance, the readiness of the Kiel Convention to accept for the first time the need for symbolisation of "advanced/retracted tongue root" may have been influenced by the existence of x-ray evidence to back up the reality of this dimension. But major issues, such as the possibility of developing a new framework for vowel classification in light of acoustic and articulatory studies, were clearly too massive to even be considered in the time available.

Bloomfield is reported by Twaddell [13] as stating that "the physical (acoustic) definition of each phoneme of any given dialect can be expected to come from the laboratory within the next decades". The notion of seeking an acoustic definition of an abstract phonological unit may be conceptually flawed (although it still lies behind a lot of experimental work); but how about definitions of phonetic categories? I am of course far too circumspect to predict that the definition of the Cardinal Vowels will emerge from the laboratory in the next ten years; but with the increasing use of computers not only in phonetic analysis, but also in teaching practical phonetics, the possibility of more quantitative phonetic reference values will have to be seriously considered.

The near future may see the first revolution in IPA description. Alternatively, the status quo may continue. But I think the health of traditional phonetics depends on active evaluation of developments elsewhere in the phonetic sciences.

4. WHAT KIND OF SPEECH?

Pat Keating, in her paper to this session, mentions the move away from "laboratory" speech to more natural data. I too believe this will be a continuing trend in the next decade. Part of the reason is practical – technological advances allow much larger stretches of speech to be stored in manipuable form, and liberate phoneticians from the 2.4 s utterance. But more important are the theoretical motivations.

It is not simply, as sometimes implied, that studying phonetically-explicit citation forms is misleading in the way that, if one needed to find out how people drive, it would be misleading to study the both-hands-on-the-wheel, constant-checking-in-the-mirror style used only by driving-test candidates. The problem is more that studying only laboratory speech provides a flat view of a phenomenon which has a third dimension – a dimension of variation. We need to know how much speakers can put in to speech (and for this it may be necessary to push speakers beyond careful speech to see what they do, say in noisy situations, to make their speech super-explicit); but it is equally important, for the understanding both of production and perception, to discover the principles governing what they are prepared to leave out in less explicit styles.

5. WHAT CAN WE TELL?

Given its historical ties to linguistics, phonetics is often seen as dealing with the pronunciation of words, or, in a more sophisticated version, with the phonetic realisation of grammatical strings. Beyond that, it is nowadays widely accepted that related matters of linguistic "performance", specifically the production and perception of utterances, fall centrally within the scope of phonetics.

But I would argue that phonetics will increasingly find its identity as the discipline which deals with the broad question "what can we tell when a person speaks?" Thus it is concerned not merely with the encoding of linguistic information, but with information about the speaker (age, sex, health, psychological state, identity, and so on).

One aspect of this is of particular interest to me. Because of the increasing availability of recorded speech samples in connection with crimes (ranging from hoax emergency telephone calls to kidnap and serious fraud) there is a rising demand for expert opinions on speaker identity (and also other factors, such as disputed utterance content and tape tampering). Many phoneticians are rightly cautious about the application of phonetic knowledge to the identification of individuals, when comparatively little is known about how far the association of a speech sample with an individual can be demonstrated under different conditions, and when so much is demanded by those in the forensic arena. However those demands will be satisfied one way or another, sometimes by self-appointed experts who lack the all-round view of the speech event which a true phonetician enjoys. Only if phonetics accepts its responsibility in this area will it be listened to as the discipline which speaks with authority on speech in the forensic context.

6. CONCLUSION

So I hope that the next decade of phonetics will be characterised by an increased sense of identity, a wider acceptance of a broadened scope of the discipline, the development of specifically phonetic theories, and greater concern for the principles underlying the traditional descriptive core of phonetics. If so, phonetics should emerge more clearly as the science of speech. Ten years from now, we should even, perhaps, occasionally hear an engineer, a psychologist, a researcher in motor control, claiming to be "a kind of phonetician".

REFERENCES

[1] ABBS, J.H. & SUSSMAN, H.M. (1971) "Neurophysiological feature detectors and speech perception. A discussion of theoretical implications", *Journal of Speech & Hearing Research, 14*, 23-36.

[2] BROWMAN, C.P. & GOLDSTEIN, L. (1989) "Articulatory gestures as phonological units", *Phonology, 6/2*, 201-51.

[3] FOWLER, C.A. (1986) "An event approach to the study of speech perception from a direct-realist perspective", *Journal of Phonetics, 14*, 3-28.

[4] FOWLER, C.A., RUBIN, P., REMEZ, R.E. & TURVEY, M.T. (1980) "Implications for speech production of a general theory of action", in B. Butterworth (ed.) "*Language Production*", Vol. 1.

[5] GOLDSMITH, J. (1990) "Local modeling in phonology", to appear in S. Davis (ed.) "*Connectionism: Theory and Practice*".

[6] KELLY, J. & LOCAL, J. (1989) "*Doing Phonology*", Manchester: Manchester University Press.

[7] KELSO, J.A.S., SALTZMAN, E.L., & TULLER, B. (1986) "The dynamical perspective on speech production: data and theory", *Journal of Phonetics, 14/1*, 29-59.

[8] KINGSTON, J. & BECKMAN, M.E. (1990) "*Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*", Cambridge: Cambridge University Press.

[9] LADEFOGED, P. (1988) "A view of phonetics", *UCLA Working Papers in Phonetics, 70*, 41.

[10] LIBERMAN, A.M., COOPER, F.S., SHANKWEILER, D.P., & STUDDERT-KENNEDY, M.G. (1967) "Perception of the speech code", *Psychological Review, 74*, 431-61.

[11] LIBERMAN, A.M. & MATTINGLY, I.G. (1985) "The motor theory of speech perception revised", *Cognition, 21*, 1-36.

[12] NOLAN, F. (1989) "Vowels", *JIPA, 18/2*, 69-74.

[13] TWADDELL, W.F. (1935) "On defining the phoneme", *Language Monograph* 16.

SPEECH PERCEPTION IN THE NEXT TEN YEARS: TECHNOLOGICAL SOLUTIONS vs. ACQUIRING ACTUAL SPEECH KNOWLEDGE

Louis C.W. Pols

Institute of Phonetic Sciences
University of Amsterdam, The Netherlands

ABSTRACT

Technological developments will continue to have a strong influence on basic and applied research in phonetic sciences. Solutions that improve performance in speech technology systems, so far have contributed little to our knowledge about human speech communication processes. That is why in the future more speech perception research for its own sake will be required, not just with speech(-like stimuli) under controlled laboratory conditions but also with 'real' speech.

1. INTRODUCTION

In her review for this conference, about 'Phonetics in the next ten years', Keating [2] emphasizes that, instead of predicting future research activities, one generally extrapolates from present and past situations. Doing so one can safely say that technological developments will continue to have a strong influence on basic and applied research in phonetics. Nowadays it is much easier to analyze many different aspects of speech, with the consequence that more can be measured. This 'more' is both in terms of all kinds of speech characteristics, as well as in terms of different languages, speakers, conditions, and styles. However, measuring more is not necessarily knowing more.

In looking ahead it is wise to look backward as well, in order to have a point of reference for judging progress in acquiring phonetic knowledge. Subsequent International Congresses of Phonetic Sciences (ICPhS) are good occasions for that because of their rather long, four-year, span and because of their emphasis on phonetics.

In my short contribution I would like to emphasize the need to improve our knowledge about the process of human speech perception. In the past, speech perception was a research topic in its own merit, presently it is frequently considered background knowledge or a by-product of research for improving automatic speech recognition, spoken language understanding, and synthesis-by-rule.

2. SPEECH PERCEPTION vs. SPEECH RECOGNITION

The speech databases, used to train those speech recognition systems that are based on neural nets or hidden markov models, provide means to acquire a lot of implicit knowledge.

This knowledge is stored in network structures and transition probabilities. However, most of the time there is no systematic and easily-accessible relation between a certain variable, such as speaker, speaking style, speaking rate, phonemic or sentence context, and the parameters of the network. Despite that, the performance of the most advanced of these systems is surprisingly high and their resistance against various sources of variation is steadily improving.

In a way this is unfortunate, since it does not force researchers to go and study these relations in more detail. Instead, speaker variability is tackled by putting a greater variety of speakers in the training data, context variability is solved by introducing triphone models, rate variation and duration variation is handled by self loops, etc.

The human listener is much more adaptive to all this (systematic) variability and finds ways to normalize. Although, for the time being, technical solutions have been found in the speech recognition domain, our knowledge about how exactly the human listener acts, has hardly been improved. The few researchers that still adhere to formalized acoustic-phonetic knowledge for performing automatic speech recognition have been far less successful, again indicating the complexity of the problem.

3. SPEECH PERCEPTION vs. SPEECH SYNTHESIS

Similar developments can be signalled in text-to-speech synthesis-by-rule [5]. The topic of 'many speakers' and 'different voice characteristics' is immediately shifted aside by choosing one voice or few voices only. Local rate changes are generally not used at all. Still, many segmental, supra-segmental, and linguistic features have to be modelled to make synthetic speech somewhat acceptable.

Only recently some progress has been made in modelling vowel duration [8]. So far the Klatt rules for American English, dating back from the 9th ICPhS in Copenhagen 1979 [3] and before, were the best available.

Our limited knowledge about context-specific dynamic formant changes has not been improved a lot in the last ten years, despite the fact that that knowledge is indispensable for rule synthesis in every single language. A common way to avoid this deficiency is to choose larger basic units such as diphones in which the nearest-neighbor transitions are already incorporated. This means another missed chance to improve our knowledge about how humans produce and perceive these transitions.

Easier access to large and multi-lingual, segmented and labeled (phonetic, prosodic, and linguistic), speech databases, such as becoming available in the ESPRIT projects SAM and Polyglot, will hopefully provide enough data to give it another try [1,6]. This holds not just for the segmental domain but, for instance, also for prosody in order to improve intelligibility and naturalness of synthetic speech. A better understanding of how humans produce and perceive conversational speech, for instance with respect to phonetic and linguistic reduction, stress assignment (given vs. new information), and prosodic phrasing, would also contribute to more natural synthetic speech.

4. SPEECH PERCEPTION FOR ITS OWN SAKE

In my opinion, the next ten years require a renewed and growing interest in studying the basic processes of human speech perception. Speech perception of course has many sides, from perceiving simple basic speech signal attributes such as pitch, duration, and vowel quality, via dynamic attributes such as formant transitions, and aspects of normalization, segregation, and trade-off, to word perception and lexical access. Although the acoustic analysis and perception of 'real' speech might become a fashionable research topic in the next decade [2], this should not prevent us from studying speech and speech-like stimuli under controlled laboratory conditions as well. For instance, if we knew better how context-specific formant transitions are produced [10] and what the variable and invariant components are that determine their role in speech perception [4,7,11], then we would have acquired some generally-applicable universal knowledge. This will certainly contribute also to improved speech technology products. However, that progress might not be astounding and should not be the main reason for doing it. Otherwise, applying acquired knowledge is an excellent way to test whether it is already complete and formalized. Just as ICPhS always had a rather strong link with phonology, the link with psycho-acoustics and hearing, as well as with psycho-linguistics was also apparent in the last few congresses. These domains are excellent bases for studying speech perception as well [9].

5. REFERENCES

[1] FOURCIN, A. et al. (Eds.) (1989), Speech Input and Output Assessment. Multilingual Methods and Standards, Ellis Horwood Lim., Chichester.

[2] KEATING, P.A. (1991), "Phonetics in the next ten years", this volume.

[3] KLATT, D.H. (1979), "Synthesis by rule of segmental durations in English sentences", Proc. 9th ICPhS, Copenhagen, vol. II, 290-297.

[4] PERKELL, J.S. & KLATT, D.H. (Eds.) (1986), Invariance and variability in speech processes, L. Erlbaum Ass., Hillsdale, N.J.

[5] POLS, L.C.W. (1990), "Does improved performance of a rule synthesizer also contribute to more phonetic knowledge?", Proc. ESCA Tutorial Day on Speech Synthesis, Autrans, 49-54.

[6] POLS, L.C.W. (1990), "How useful are speech databases for rule synthesis development and assessment?", Proc. ICSLP-90, Kobe, Vol. 2, 1289-1292.

[7] POLS, L.C.W. & SCHOUTEN, M.E.H. (1987), "Perception of tone, band, and formant sweeps", In: Schouten (Ed.), 231-240.

[8] SANTEN, J.P.H. van & OLIVE, J.P. (1990), "The analysis of contextual effects on segmental duration", Computer Speech and Language 4, 359-390.

[9] SCHOUTEN, M.E.H. (Ed.) (1987), The Psychophysics of speech perception, M. Nijhoff Publ., Dordrecht.

[10] SON, R.J.J.H. van & POLS, L.C.W. (1990), "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", J. Acoust. Soc. Am. 88(4), 1683-1693.

[11] WIERINGEN, A. van & POLS, L.C.W. (1990), "Transition rate-dependent processing of one-formant speech-like stimuli", IFA Proc. 14, 1-16.

WHAT PATHOLOGY TELLS US ABOUT LEXICAL ACCESS IN SPEECH PRODUCTION

Victoria A. Fromkin

University of California, Los Angeles

ABSTRACT

Language deficits which result from brain damage provide insights into the nature of normal speech production and perception. Aphasia data are shown to be of specific importance in revealing the structure of the mental lexicon, the manifold representations of each lexical item, how they interact, and the processes involved in accessing this lexicon during speech production.

1. INTRODUCTION

- As stated almost thirty years ago by Denes and Pinson, [4] speech communication may be viewed as a chain of events starting "in the speaker's brain (where) ... appropriate instructions, in the form of impulses along the motor nerves, are sent to the muscles of the vocal organs, the tongue, the lips and the vocal cords" causing movements which in turn produce speech sound waves. We know a great deal about the physiological, articulatory, and acoustic aspects of these stages of speech production as a result of experimental phonetic research. But we are still far from understanding the processes by which a speaker once he has arranged his thoughts, "puts what he wants to say into linguistic form ... by selecting the right words and phrases to express its meaning, and by placing these words in the correct order required by the grammatical rules of the language.." (p 3)

- One approach to investigating this complicated process is to see if deviant language, such as the speech of brain damaged aphasic patients can provide insights into the normal linguistic processing system.

- The entry into the area of aphasia research and brain-mind-cognition studies was a logical development of the goal to understand the nature and form of human linguistic knowledge and how this system of knowledge -- the mental grammar -- is put to use in speech production and comprehension.

- Interest in brain mechanisms underlying language and speech goes back about 2000 years. Aristotle's false view that the brain is a cold sponge whose primary action is to cool the blood was not shared by the Graeco-Roman physicians, who, writing in the fifth century B.C.E., recognized that the loss of speech and the loss of language could be distinguished. The Hippocratic view was that the brain is "the messenger to the understanding" and the organ whereby "in an especial manner we acquire wisdom and knowledge." [1]

- This recognition of the brain-cognition-language relationship which has endured through the centuries, led in the early part of the 19th century to theories of 'localization' suggesting that different human abilities and behaviors are traceable to specific brain structures. In 1861, in a meeting in Paris, language was specifically related to the left side of the brain in a paper presented by Paul Broca in which he presented autopsy evidence showing that a localized (anterior) left hemisphere lesion resulted in a loss of ability to speak, whereas focal lesions in similar parts of the right brain did not. He managed to convince his Parisian audience (and most of neurology) that "On parle avec l'hemisphere gauche". [3]

- In 1874, Wernicke [12] pointed out that damage in the posterior portion of the left temporal lobe (now called Wernicke's area) results in a different form of language breakdown than that occurring after damage to the frontal cortex (Broca's area). These different kinds of acquired language loss -- aphasias -- continue to be corroborated.

- Aphasia research by linguists and phoneticians has been motivated in part by these findings that focal damage to specific brain areas results in the disruption of distinct cognitive functions as well as motor and perceptual abilities, and that the selectivity appears to be specific as to the parts of language which are effected. This supports a modular conception of the grammar itself, in which the components are interactive but independent of each other, since these components as well as the hierarchy of linguistic units posited by linguists appear to be just those parts which can be differentially destroyed or damaged. Given this fact, the study of the kinds of disruption which follows localized lesions, permits us to investigate the levels of representation at different stages in the memory.

- Jakobson [5] was the first linguist to conduct aphasia research, following up on the insights of de Courtenay in 1885 and Saussure in 1879 who had expressed the belief that a study of language pathology could contribute to linguistics. As this symposium will hopefully show, their views have been corroborated since such research is contributing to our understanding of both the representation and the processing of language and speech.

2.LEXICON

2.1. Lexical Selection

- Aphasia research has become increasingly concerned with lexical representation and access in the attempt to understand "how the right words" are selected. [see, for example, 7,9,10] Simultaneously, current linguistic research is being conducted on the lexicon and the morphological component of the grammar.

2.2. How Many Lexicons?

- In trying to understand the complex problems of lexical selection and phrase construction in speech production, one question of interest is how the lexicon is organized, and whether, for example, content words (open class items) are listed separately from grammatical morphemes (closed class items) -- inflectional and derivational, free and bound. Aphasia research (as well as speech errors produced by normals) supports the proposal that these two classes of formatives are processed at different levels of speech production [7, 8]. It is logical to assume that if this is so, the two categories of morphemes are also stored in separate lexicons.

- The speech output of Broca's and Wernicke's aphasia patients provide some evidence. Broca's aphasic speech is characterized by word-finding pauses, loss of both free and bound grammatical morphemes, and quite often, disturbed word order, but with access to content 'open-class' words. Auditory comprehension for colloquial conversation gives the impression of being generally good, although controlled testing reveals considerable impairments. The term **agrammatism** is often used as a term for Broca's aphasia.

- Wernicke's aphasia patients, on the other hand, produce fluent speech with good intonation and pronunciation, but with many word substitutions (both semantically similar and dissimilar), neologisms as well as phonological errors. They also show comprehension difficulties. Their utterances while often semantically empty (given their difficulties with major category morphemes, e.g. nouns, verbs, adjectives, appear to be well formed syntactically with inflectional and grammatical morphemes intact.

- Thus, these two major classes of aphasics reveal differential impairment in these two classes of morphemes. [2]
- Agrammatism and Wernicke's 'fluent' aphasia are not the only types of aphasia which show differential processing of lexical and grammatical morphemes. The language deficits of some patients after brain injury diagnosed as having acquired dyslexia primarily effect reading and writing, leaving the spoken language intact. These subjects also provide insights into normal speech production since lexical access which is impaired for many of them is involved in spontaneous speech as well as in the reading and writing processes.
- Again we find evidence for the separation of the lexicon into sub-lexicons, one storing major category content words and morphemes, and another where grammatical formatives are listed. A patient of Newcombe and Marshall [10], for example, shows differential impairment in reading words in these two major classes. Errors are made in reading content words, with substitutions of semantically and/or phonologically related words, but grammatical formatives can not be accessed at all as shown in Table 1.

= = = = = = = = = = = = = = = = = =

Table 1. Patient G.R. [10]

Stimulus	Response
WITCH	'witch'
WHICH	'no!'
BEAN	'uh...soup'
BEEN	'no!'
HOUR	'time'
OUR	'no!'
EYE	'eyes'
I	'no!'
HYMN	'bible'
HIM	'a boy? no!'
WOOD	'wood'
WOULD	'no!'
FOUR	'four'
FOR	'no!'
MOOR	'fog..mist?'
MORE	'no!'

- Dyslexics like G.R. who often substitute semantically related words, e.g. "prison" for JAIL appear to bypass any orthographic to phonological or pronunciation rules, going directly to what must be an orthographic sub-lexicon connected to a semantic sub-lexicon. The connection between the semantic and phonological representations remain, but in accessing the semantic "address" a misselection occurs. Thus we have evidence for the separation of components even within the major subcomponent of lexical content words.
- We find that this reading disruption problem is paralleled in normals in the kinds of semantically similar word substitutions which occur in speech errors, e.g. 'downtown' for 'uptown', 'wrist' for 'finger', 'behind my face' for 'behind my back'.
- Some of the substitutions both in the reading errors of acquired dyslexics and the speech errors of normals show phonological similarities between the target and the substituted word rather than or in addition to semantic similarities, e.g. 'fluency' for 'frequency' 'progress' for 'practice', 'persecuted' for 'prosecuted'. Such errors suggest the ways in which the entries in each sub-lexicon are listed e.g. by semantic feature or class in the semantic lexicon, by phonological form in the phonological lexicon. Since the number of phonologically similar substitutions which share initial word onsets is significant, it seems safe to conclude that words are listed according to such onsets. The fact that in addition to the onsets, substitutions show other phonological similarities, e.g. of segments and number of syllables, shows that these phonological factors play a role in both the organization and the access of the lexicon but the nature of the organization requires further investigation; the listing of words in the phonological sub-lexicon according to number of syllables and onsets for each subset seems to be a possible starting point.

- Additional information about the organization and processing of the lexicon is provided by a patient with whom I have worked over the last number of years, referred to as Kram and MS in the literature. [7, 9] Kram shows good language comprehension and fluent intelligible speech production, with greatly impaired reading and writing ability. For example, he will read the word 'fame' as [fæmi] and write it to dictation as FAM; he can neither map the orthography onto a phonological representation in his mental lexicon nor use normal orthographic-to-phoneme rules; he uses his own idiosyncratic rules instead. Furthermore, he can understand the meaning of a word only through its phonology; when he produces nonsense forms, he is unable to state what the written form means or even if it is a 'real' word. If he does read a word correctly he understands what it means; if he reads a homophone correctly, he cannot determine which of the ambiguous meanings is represented by the spelling, as shown in Table 2.

==================

Table 2. Kram's pronunciation and comprehension of written homonyms.

Stimulus	Pronunciation	Meaning
sum	sum/some	"I've got some"
can	san	"don't know"
for	for/four	"I have four fingers and a thumb"
pig	pig	"oink oink"

- Other similar cases are reported in the literature [11]

-We see again that such deviant language, written as well as spoken, provides clues as to lexical representation, structure, and processing.

-The neologistic jargon produced by other aphasics also provides information about normal processing of speech. (1) and (2) are examples of such utterances:

(1) the leg vilted from here down

(2) This is the krebekacks where the frejes get out after the chuw.

- Note that the nonsense forms are well formed both phonologically and morphologically, i.e. appropriately inflected or derived. But some aphasics, i.e. the agrammatic patients with Broca's aphasia, have particular difficulty with inflectional affixes [2,9]; English speaking agrammatics may omit grammatical formatives completely; speakers of other languages, like Hebrew, do not omit bound morphemes but substitute other incorrect inflectional morphemes. This difference was accounted for by Grodzinsky [9] by an explanation of particular interest to those of us concerned with speech production mechanisms. He points out that vowels in Hebrew are predictable, according to inflectional and derivational morphological rules. For example, the vowel in the word for a single male child is "e" *yeled*, is "a" for a female child *yalda*; the plural for these two singular nouns is *yeladim* and *yeladot*, respectively. Since the roots of Hebrew words consist only of consonants, e.g. /y-l-d/ in the examples given, agrammatic aphasic Hebrew speakers would be unable to talk at all if they omitted the inflectional and derivational morphemes which are realized vocalically. Thus, these Hebrew speakers instead of omitting these morphemes, substitute incorrect vowels in words such as those above and omit free-standing grammatical morphemes. This shows the phonetic and speech production constraints which exist.

- The aphasic data which have been cited show us something about how a speaker "puts what he wants to say into <u>linguistic form</u>" even if the 'wrong' words or wrong inflections are selected, or if the right words are distorted. Denes and Pinson 's observations can be extended to cover the production of jargon if one posits that a speaker must first, prior to articulatory processes, generate a string of phonological units, properly inflected according to phrase structures determined by the grammar, which string is then mapped onto the proper motor commands to move the articulators to produce sounds.

3.PRODUCTION MODELS
3.1 Lexical Models

-These data from pathological language and from normal but deviant (speech error) language have led to the construction of first approximation lexical models composed of phonological, orthographic, and semantic sub-lexicons. [9]. Each entry in each of the components is connected to its parallel representation through an addressing system. When the connections between the orthographic representation of a word and its phonological representation is blocked, a speaker is unable to read the word; when the connection between the semantic representation of a word and its phonological representation is disturbed, a semantically similar but incorrect substitute can be produced, or, as in the case of jargon aphasics, the entire phonology may be disrupted.

- The cases of jargon aphasia are particularly telling; it is seldom that the inflectional and free standing grammatical morphemes are mispronounced again supporting the notion of a major division into two lexicons, each, possibly with its own sub lexicons.
- Under conditions of pathology, access to either lexicon and the connections between the sub lexicons may be blocked. It must be the case, then, that these divisions exist in the normal lexicon as well and under certain conditions (which are not clear as yet) partial blocking may occur for normal speakers.

4. REFERENCES

[1] ARBIB, M.A., CAPLAN,D. & MARSHALL, J. C. (Eds.) (1982) *Neural Models of Language Processes*. NY: Acad. Press

[2] BADECKER,B.& CARAMAZZA, A. (1985)On considerations of method and theory governing the use of clinical categories in neurolinguistics and cognitive neuropsychology. *Cognition 10*, 17-24

[3] BROCA, P. (1861) Nouvelle observation d'aphemie produite par une lesion de la moitie posterieure des deuxieume et troisieme circumvolutions frontales. *Bulletin de la Societe Anatomique de Paris, 3* 398-407

[4] DENES, P.B. & PINSON,E.N. (1963) *The Speech Chain*. Bell [

[5]JAKOBSON,R. (1940) *Kindersprache,Aphasie and allgemeine Lautgesetze*. Almqvist u. Wilsells, Uppsala.

[6] FRIEDERICI,A.D. (1983) Perception of words in sentential contexts: Some real time processing evidence. *Neuropsychologia.21* 351-8

[7]FROMKIN,V.A. (1987) The lexicon: Evidence from acquired dyslexia. *Language 63*. 1-22

[8] GARRETT,M.F. (1988) Processes in language production. in F.J. Newmeyer (Ed.) *Linguistics: The Cambridge Survey III* Cambridge Univ.Press. 69-96

[9] GRODZINSKY,Y. (1984) The syntactic characterization of agrammatism. *Cognition 16*. 99-120

[10] NEWCOMBE,F.& MARSHALL, J.C. (1984) Varieties of acquired dyslexia: a linguistic approach. *Seminars in Neurology,4.2* 181-95

[11]PATTERSON,K. (1978 Phonemic dyslexia: Errors of meaning and the meaning of errors.*Quarterly Journal of Experimental Psychology 29*. 515-25

[12]WERNICKE,C. (1874) *Der aphasiche Symptomenkomplex* Breslau: Cohn & Weigert

FROM CEREBELLAR DYSARTHRIA TO NORMAL SPEECH PRODUCTION

Michèle GENTIL

INSERM, Paris, France

The purpose of this paper is to consider what cerebellar dysarthria tells us about the physiological mechanisms involved in speech, from EMG patterns of 13 cerebellar patients. Results are discussed in terms of cerebellar system characteristics regulating speech activity.

1. INTRODUCTION

Careful observations of humans with lesions of the cerebellum or cerebellar pathways have demonstrated a variety of motor deficits including disorders of voluntary movements. Lesions affecting the cerebellar system result in dysarthric speech. Acoustic and X-ray analyses of ataxic dysarthria have shown that the movements of speech lack precision in direction, velocity and extent [1, 3, 4]. A few studies have described the pathological kinesiology of the articulatory organs of the cerebellar patients in EMG terms [2, 5]. We assessed the effects of cerebellar lesions on oro-motor system for a better understanding of the control mechanisms involved in speech. We studied EMG patterns associated with lip and jaw movements during speech production for 13 patients with Friedreich ataxia.

2. METHOD

9 females and 4 males with a diagnosis of Friedreich ataxia and 2 normal subjects (1 female and 1 male) participated in this study. The mean age of patients was 37.6 with a standard deviation (SD) of 12.3. Normal subjects were 25 years old.

Electromyographic signals were recorded simultaneously from 5 muscles using hooked wire electrodes. They were orbicularis oris superior (OOS), orbicularis oris inferior (OOI), mentalis (MENT), depressor labii inferior (DLI) and anterior belly of the digastric (ABD). All electrodes were placed on one side of the subject, usually the left.

The subjects were required to :
- Produce the syllable /ba/ in response to an auditory signal (10 times)
- Repeat similar monosyllabic or four-syllable nonsense utterances /ba/, /epapap ə/. Each utterance was repeated 7 times at 2 speaking rates, conversational and fast.

The amplified data signals were simultaneously recorded on magnetic tape using an 8-channel instrumentation recorder (Euromag model 5423 MP) and on paper using an 8-channel recorder (Gould model ES 1000). The audio signal from a microphone LEM was also recorded on an edge track of the tape, and on paper.

The data were processed using a laboratory computer system including a Digital Equipment Corporation PDP 11 / 34 control processor. All of the data were subjected to ensemble averaging.

3. RESULTS

3.1 Initiation of muscle activities

A delay in the initiation of muscle activities was observed. The patients always showed a much longer interval between an auditory signal and the onset of any muscular activity for the production of the /b/ in the syllable /ba/, than normals. Table 1 summarizes results for 11 patients and 2 normal subjects. Among other things, cerebellar lesions result in hypotonia [3] This appears us a probable explanation for the delay in the initiation of muscular activities.

3.2 Mean durations of muscular activities

The mean durations of muscular activities for patients with Friedreich disease always exceeded those for normals. The lengthening of muscular activities for 11 patients in comparison with normals is shown in tables 2 (for /b/) and 3 (for /a/) in the syllable /ba/ produced at a conversational speaking rate. These results suggest that the cerebellar system is involved in the control of duration parameters of vocalization.

3.3 Muscular synergia

Muscular synergia can be gauged from the performance of alternating movements. At a conversational speaking rate, muscular synergia was better preserved than at a fast speaking rate. In fact, at a conversational speaking rate, 5 of 13 patients showed a normal EMG pattern similar to that of the control subject. Fig 1 illustrates this normal EMG pattern for patient MD. It is observed the synchronization of the MENT, OOI, OOS activities associated with the closing movement of lips on the one hand, and that of the ABD, DLI activities associated with the opening movement of jaw and inferior lip on the other hand. Moreover, this pattern reveals reciprocity between activity of agonists and antagonists. The other 8

patient productions at a conversational speaking rate were abnormal (fig. 2).
Of 13 patients, only 8 could produce the 4-syllable nonsense utterance /epapapə / 7 times at a relatively fast speaking rate, the rapid alternating movements of articulatory organs presented too many difficulties for the other 5. At a fast speaking rate, no patient produced a normal EMG pattern. In other words, the cerebellar system is concerned with control of speech movements requiring coordination of synergistic muscle groups.

4.CONCLUSION

It is known that the cerebellum is responsible for the delicate and precise control of posture and locomotion [6] but our knowledge of its real role in the regulation of speech movements is still limited. Extensive studies are needed to throw light on control of the vocal tract exerted by the cerebellum at multiple levels, including coordination of orofacial, velopharyngal, laryngeal and respiratory activities in speech production.

5.REFERENCES

[1] HIROSE, H. (1986), "Pathophysiology of motor speech disorders", *Folia Phoniat.*, 38, 61-88.

[2] HIROSE, H., KIRITANI, S., USHIJIMA, T. & SAWASHIMA, M. (1978), "Analysis of abnormal articulatory dynamics in two dysarthric patients", *J.Speech Hear. Dis.*, 43, 96-105.

[3] KENT, R.D. & NETSELL, R. (1975), "A case study of an ataxic dysarthric. Cineradiographic and spectrographic observations", *J. Speech Hear. Dis.*,1, 115-134.

[4] KENT, R.D., NETSELL R & ABBS, J.H. (1979), "Acoustic characteristics of dysarthria associated with cerebellar disease", *J. peech Hear. Res.*, 22, 627-648.

NETSELL, R. & ABBS, J.H (1977), "Some possible uses of neuromotor speech disturbances in understanding the normal mechanism" in M. Sawashima and F.S. Cooper (eds), Dynamic aspects of speech production, University of Tokyo Press, 369-398.

[6] BROOKS, V.B. & THACH, W.T.(1981),"Cerebellar control of posture and movement". In J.M. Brookhart, V.B. Mountcastle and V.B. Brooks, (eds) Handbook of physiology Sect 1, The Nervous system, Vol.2, Part 2 (Bethesda American Physiological Society).

Table 1 - Summary of mean reaction times : Intervals between an auditory signal and the onset of the muscular activities for eleven patients and two normal subjects. All values are in msec.

SUBJECTS	MUSCLES	AM	SC	MD	XF	JM	ED	DD	NR	HB	AG	LM	MA	MP
		PATIENTS											NORMALS	
MEAN REACTION TIMES	MENT	255	172	181	284	180	112	252	144	179	144		58	10
	OOI	550			473	212	141	307	144	327		109		29
	OOS	351	211	196	370	141		212		320		160	58	

Table 2 - Mean durations of the muscular activities for the production /b/ in the syllable /ba/. All values are in msec.

SUBJECTS	AM	SC	MD	CB	XF	JM	ED	DD	NR	AG	LM	NORMALS
	PATIENTS											
MENT	334	423	484	510	554	394	280		488	363	495	240
OOS	269	533	408	550	§	353	304	360	500	§	472	152

§ Data were not available.

Table 3 - Mean durations of the muscular activities for the production /a/ in the syllable /ba/. All values are in msec.

SUBJECTS	AM	SC	MD	CB	XF	JM	ED	DD	NR	AG	LM	NORMALS
	PATIENTS											
ABD		364		§		315	388	390	457	266	427	168
DLI	246	372	291	276	495	285	348	300		116	412	174

§ This muscular activity was not recorded.

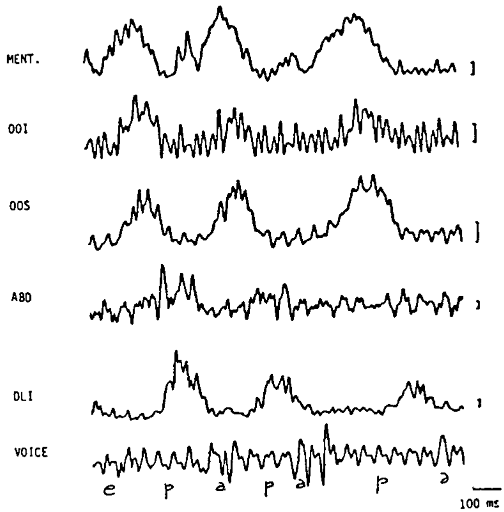


Figure 1 *Averaged integrated EMG: patient M.D. shows a normal pattern during the production /epapapa/ at a conversational speaking rate. Brackets indicate 100 μV.*

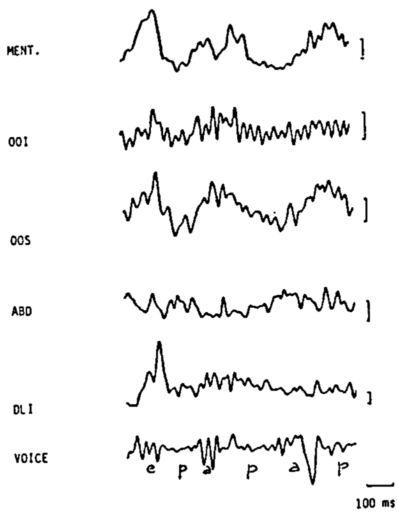


Figure 2 *Averaged integrated EMG: patient A.M. shows an abnormal pattern during the production /epapapa/ at a conversational speaking rate. Brackets indicate 100 μV.*

POSSIBILITIES OF COMPENSATION IN SPEECH PRODUCTION

Shizuo Hiki

School of Human Sciences, Waseda University
Tokorozawa, 359 Japan

ABSTRACT

Possibilities of compensation in speech production are discussed from viewpoints of control of speech organs, realization of acoustical requirements and role of the auditory feedback, taking examples of lip, tongue and jaw movements for the vowel and consonant utterances by normal adults and children and by hearing-impaired children.

1. ASPECTS OF COMPENSATION

It has been widely noticed by researchers in their observations on speech production process that gestures of phonation and articulation can be substituted in various ways in order to compensate for personal deviation in speech organs in normal cases, and especially for their defects in pathological cases.

In this report, aspects of the possibilities of compensation in speech production are discussed with regard to whether the control of speech organs is arbitrary or not, how far the acoustical requirements are realized, and what kind of role the auditory feedback plays.

Examples used here are lip spreading and rounding, lingual contact to the palate and muscle contraction in articulator for vowels and consonants by normal adults and children having different shape and size of the palates, and by hearing-impaired children with normal articulator.

2. ARBITRARY CONTROLS OF THE LIPS

On the three-dimensional matrix consisted of axes of the front and back movements of tongue, the high and low displacements of the tongue, and spreading and rounding of the lips, the five Japanese vowels [i, e, a, o, u] can be represented as shown in the top left of Fig. 1. The characteristic shapes of the mouth basically correspond to the second and third axes (bottom right) rather than to the first and second axes (bottom left). In order to coordinate the configuration and the data of stroboscopic observation of the mouth shape [1], the jaw opening (displacement of the lower incisors) and the separation and width of the lips are also shown in the bottom right.

As the area of the lip opening is nearly proportional to the product of the width and separation of the lips, they can be traded each other for keeping the area required by each kind of vowel (along the bold lines).

This is an aspect of the compensation in speech production in which suitable gesture within acceptable range for each vowel has been arbitrary chosen by each speaker in the development of articulation.

3. CONSTRAINTS BY PALATE SHAPE

3.1. Difference in Palate Shape

Fig. 2 shows the boundary of the area of the maximum lingual contact to the hard palate in the monosyllabic utterance of [i] in solid line by two adult subjects as examples. They were extracted through the frame-by-frame inspection of the recorded electro-palatograms (with sixty-four electrodes), and drawn on the photograph of the frontal view of the plaster cast of the palate of each subject covered with the artificial one [2].

The palate of the left subject was wider and deeper, and the boundary of the contact for [i] was closer to the mid-sagittal plane, while the boundary of the right subject was near the lateral edge of the palate.

In this aspect of compensation, lingual contact has been involuntary adjusted to peronal difference of the palate shape through auditory feedback, in order to realize the acoustical requirement for the vowel.

3.2. Growth of Child Palate

Reference lines were drawn on the horizontal and vertical views of the plaster casts of the hard palates of the fifteen adults (M: male, F: female), thirty children in the dental stages; IIA, C, IIIA, B, C and IVA, and two children (a: boy, b: girl) in different dental stages, as shown in the top of Fig. 3, and used for the measurement of the size of various parts of the hard palate [3]. The size and shape of the hard palate changes significantly as the subject grows, as shown in the bottom. The shape of palate in children has the characteristic that the front part is shorter and shallower compared with that of an adult. Consequently, the surface of the tongue tip is apt to contact a wider area of the front of the hard palate.

Examples of the area of maximum lingual contact at the closure of alveolar flapped [r] are shown in the left column of Fig. 4, and alveo-dental plosives [d] in the right column. The utterance by one of the child subjects is shown in the bottom row, while that of one of the adult subjects in the top. In the utterance by children, the area of maximum contact at the closure of such consonants tends to be closer to the front edge of the hard palate, compared with the adult subject, and does not show the characteristic pattern for each consonantsas compared with the adult subject.

In this aspect, constraints by the palate shape are too strong to be compensated, even though those consonants may be perceptually differenciated each other by normal children and acoustical target for each may be settled in a higher level of motor control.

4. DEFECTS IN AUDITORY FEEDBACK

Measurement of the formant frequencies of the five Japanese vowels uttered by forty-two hearing-impaired children [4] showed that the vowel-space deviated in various way from that of the normal children. They can be classified into the following types; the shape similar to that of the normal children but the range of F2 reduced slightly, the range of F2 reduced and the pentagonal shape rotated, the range of F2 also reduced and [a] and [o] came closer to each other, the range of F2 reduced and [i] and [e] came closer to each other, both F1 and F2 almost neutralized, and the range of F2 reduced and F1 raised, as shown in the left of Fig. 5.

The right of Fig. 5 shows the relation between the change in F1 and F2 and the movements of the jaw, tongue and lips in the normal utterance of the five Japanese vowels as calculated by the computer-simulated muscle contraction model [5]. Also shown in this figure are the schematic mid-sagittal sections of a child's vocal tract during the normal utterance. By referring to these figures, the deviation of the vowel-space of hearing-impaired children from those to the normal children can be explained in terms of the nature of the imperfection of articulatory movement.

The reduction of the range of F2 implies that the forward pull of the tongue for [i] and [e] and the backward pull for [o] and [a] were not sufficient. The deviation of F1 and F2 of [i] toward the normal area of [e] shows that the constriction of the vocal tract was not narrow enough, while that of [o] toward the normal area of [a] suggests that the protrusion of the lips for [o] was not sufficient.

Those deviations of the vowel-space of the hearing-impaired children having basically normal articulator were due to the defects in auditory feedback which is indispensable for the compensation in speech production. Therefore, analysis on how the vowels were contrasted with each other in vowel-space, along with the considerations on the range of arbitrary controls and constraints in articulator for realization of the acoustical requirements, will provide a useful data for the basic nature of compensation in articulation.

5. REFERENCES

[1] Fukuda, Y. and S. Hiki(1982), Characteristics of the mouth shape in the production of Japanese: Stroboscopic observation," J. Acoust. Soc. Japan (E), 3(2), 75-91.

[2] Hiki, S. and H. Itoh (1986), "Influence of palate shape on lingual articulation," Speech Communication, 5, 141-158.

[3] Hiki, S. and R. Kagami(1976), "Some properties of formant frequencies of vowels uttered by hearing-impaired children," J. Acoust. Soc. Amer., 59(S1), S86.

[4] Hiki, S. and J. Oizumi(1974), "Speech synthesis by rule from neurophysiological paramerter," Preprints, Speech Communication Seminar, Stockholm, 219-225.

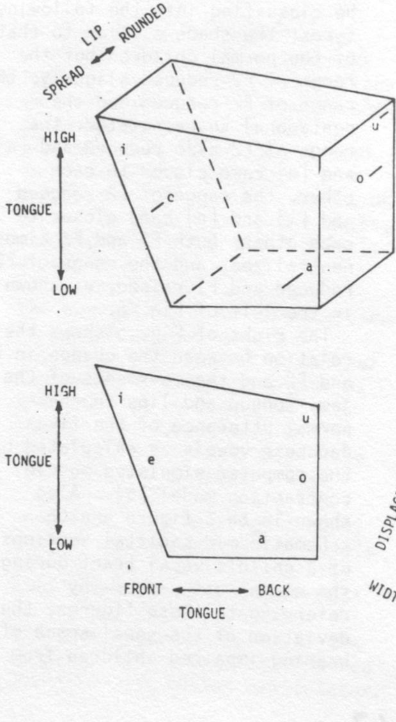
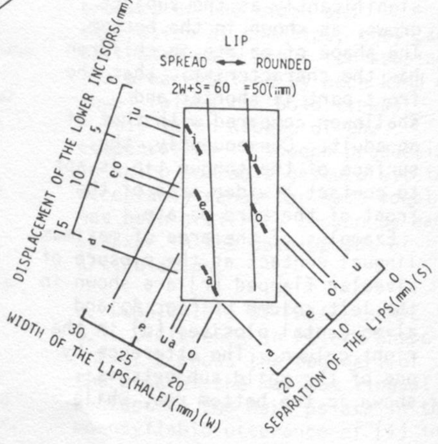


Fig. 1.



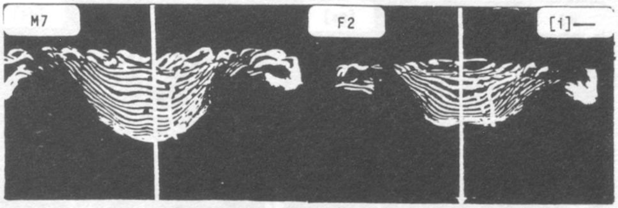
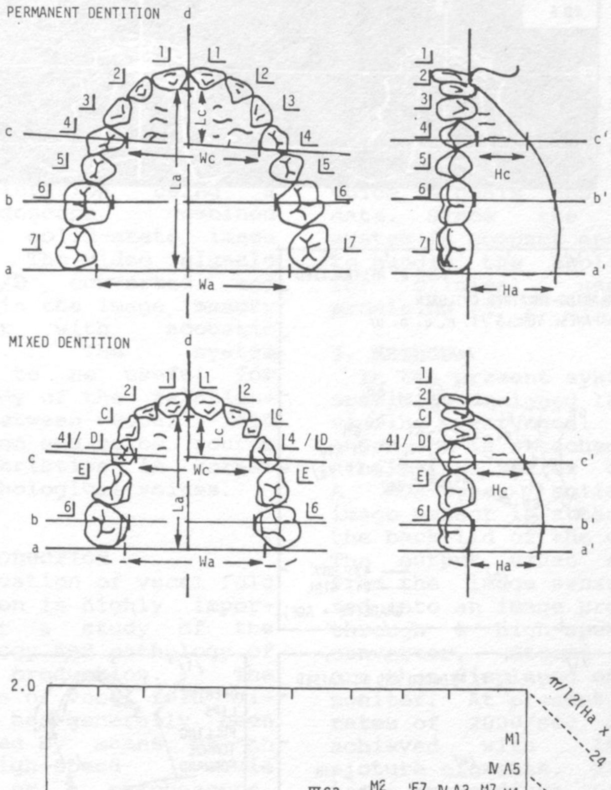

M7
F2
[i]—

Fig. 2.


PERMANENT DENTITION
MIXED DENTITION
d
c
b
a
c'
b'
a'
Lc
La
Wc
Wa
Hc
Ha

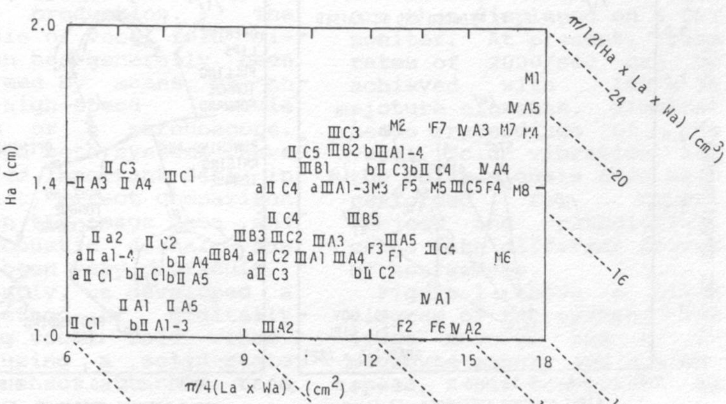

Ha (cm)
2.0
1.4
1.0
6
9
12
15
18
π/4(La x Wa) (cm²)
π/12(Ha x La x Wa) (cm³)
24
20
16
12
10
8
6
4

Fig. 3.

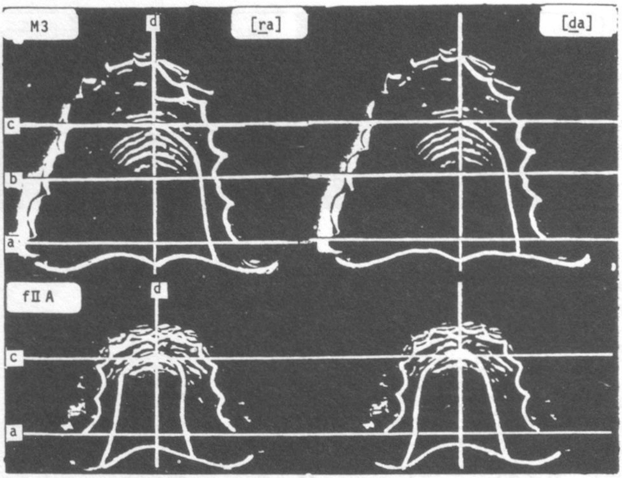
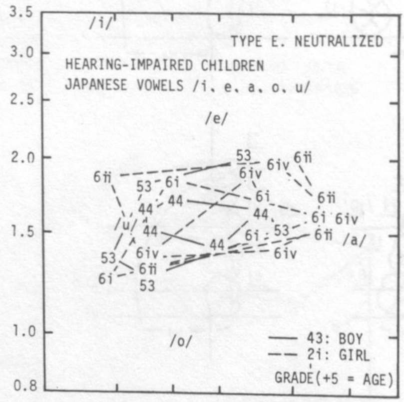
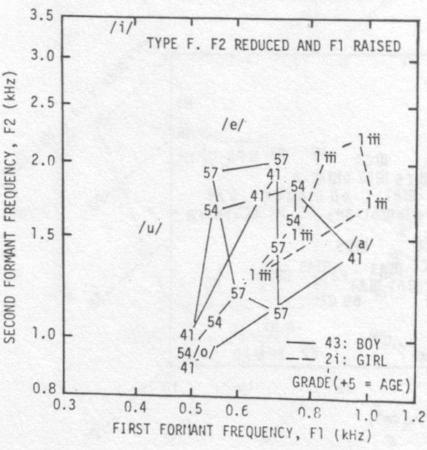

M3
[ra]
[da]
d
c
b
a
fII A
d
c
a

Fig. 4.


TYPE E. NEUTRALIZED
HEARING-IMPAIRED CHILDREN
JAPANESE VOWELS /i, e, a, o, u/
/i/
/e/
/u/
/a/
/o/
43: BOY
2i: GIRL
GRADE(+5 = AGE)


TYPE F. F2 REDUCED AND F1 RAISED
/i/
/e/
/u/
/a/
/o/
43: BOY
2i: GIRL
GRADE(+5 = AGE)
SECOND FORMANT FREQUENCY, F2 (kHz)
FIRST FORMANT FREQUENCY, F1 (kHz)

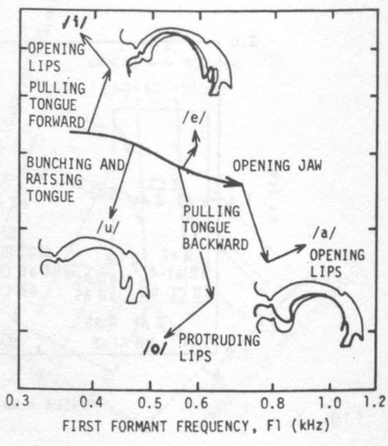

/i/
OPENING LIPS
PULLING TONGUE FORWARD
/e/
BUNCHING AND RAISING TONGUE
OPENING JAW
PULLING TONGUE BACKWARD
/u/
/a/
OPENING LIPS
/o/
PROTRUDING LIPS
FIRST FORMANT FREQUENCY, F1 (kHz)

Fig. 5.

ANALYSIS OF NORMAL AND PATHOLOGICAL VOCAL FOLD VIBRATION WITH REFERENCE TO VOICE QUALITY

H. Hirose, S. Kiritani and H. Imagawa

University of Tokyo, Tokyo, Japan.

ABSTRACT

A digital image-recording system has been developed to facilitate high-speed image recording or vocal fold vibration using a tele-endoscope combined with a solid-state image sensor. The video signals are A/D converted and stored in the image memory together with acoustic signals. The system appears to be useful for the study of the relationship between vocal fold vibration and voice source characteristics in normal and pathological voices.

1. INTRODUCTION

Observation of vocal fold vibration is highly important for a study of the physiology and pathology of voice production. The analysis of vocal fold vibration has generally been performed by means of an ultra-high-speed movie system or a stroboscope. Although both systems have provided important data in the past, direct comparison between the image data and the acoustic signal has often been very difficult.

Recently, we developed a new method of digitally imaging vocal fold vibration using a solid-state image sensor attached to a conventional camera system. This system is relatively free from the mechanical noises and suitable for simultaneous recordings of voice signals and image data. Since the entire system is compact and easy to handle, the application for clinical use is promising [1, 2, 3].

2. METHOD

In the present system, a specially designed lateral-viewing laryngeal tele-endoscope is attached to a single-lens reflex camera. A MOS-type solid-state image sensor is attached to the back lid of the camera. The output video signals from the image sensor are fed into an image processor through a high-speed A/D converter. Stored images are then displayed on a CRT monitor. At present, frame rates of 2000/sec can be achieved with 100 x 36 picture elements. Simultaneous recordings of the vocal fold vibration and the voice signals have been performed for normal subject and pathological cases with different degree of hoarseness.

Figure 1 shows a block diagram of the system. The image memory has a 2-megabyte memory and a high-speed, 8-bit converter. As

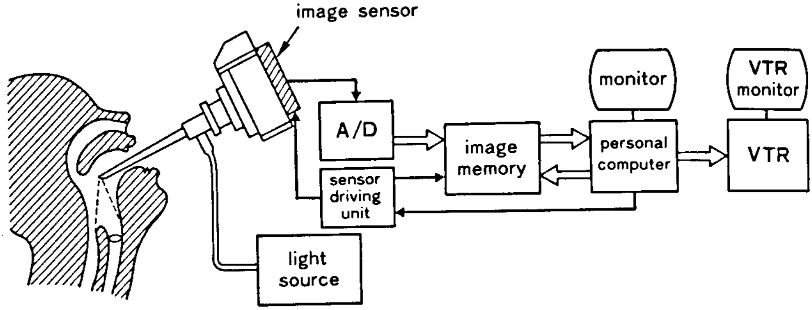


Fig. 1 Block diagram of the present system

a light source, a pair of 250 W halogen lamps are used.

Data recording is made in the same manner as in still photography of the larynx. The larynx is visualized through a view finder with the tip of the scope in the pharynx. The camera shutter is then released for data recordings. During the shutter opening of approximately 150 msec, 200 to 400 data frames are stored in the memory.

For the purpose of clinical application, recordings were made in those cases with organic changes in the vocal fold associated with "rough" quality of voice.

3. RESULTS AND COMMENTS

An application of the present system for the analysis of pathological larynx has proved promising. Incomplete glottal closure and asymmetrical or irregular vibratory patterns were easily identified in cases with recurrent laryngeal nerve paralysis, vocal fold polyp, polypoid vocal fold or sulcus vocalis. Furthermore, asynchronous movement patterns were often noted between the left and right vocal folds and/or between the anterior and posterior parts of the one vocal fold. It was also confirmed that, in most cases, pathological vibratory patterns were accompanied by irregularity in simultaneously recorded acoustic signals.

In the acoustic waveform, these voices show cycle to cycle variations in the waveform. However, in most cases, similar waveforms tend to recur cyclically (namely, at every other cycle, every third cycle etc.). Cyclic fluctuations in the pattern of vocal fold vibration are rather small and, in some cases, it is not easy to identify the pattern of fluctuation through simple visual inspection of the image.

In order to clarify the pattern of fluctuation in the movement of the vocal folds, brightness values at picture elements (pixels) along the horizontal scan line across the selected part of the glottis were plotted by the computer and characteristics of the successive frames were analyzed.

Figure 2 shows acoustic wave forms and brightness curves for Case 1, 20-year-old female with sulcus vocalis.

In this particular case, the right vocal fold showed only a very limited vibratory movement and a complete glottal closure was not obtained during the vibratory cycle. The acoustic signal shows three distinct cycles having different waveforms where a similar waveform appears at every third cycle. In one cycle, the dip in the brightness curve which corresponds to the glottal opening is clearly deeper than that in the other two cycles. The finding would indicate that the glottal opening is larger in that cycle than in the others.

Figure 3 shows acoustic waveforms and brightness curves for Case 2, a 59-year-old male with cyst of the left vocal fold. Observations of the vibratory pattern disclosed that the amplitude of vibration of the left vocal fold was much smaller than the right.

In this case, two distinct periods of strong and clear excitation and weak, noisy excitation alternated with each other resulting in fluctuation in the waveform at every other pitch period.

Inspection of the brightness curves indicates that the duration of the closure period is clearly different in these two cycles. In one cycle (cycle A, hereafter), the closure period is longer and the excitation in the speech waveform is strong. In the other cycle (cycle B), the closure period is short and speech waveform is noisy, suggesting that the glottal closure is incomplete in this cycle. In this particular case, it can also be noted that there is a marked asynchrony between the movements of the anterior and posterior parts of the glottis. In cycle A, the anterior part starts to open immediately after the posterior part closes, while in cycle B, the anterior part remains closed until after the posterior part begins to open. It can be speculated that this imbalance between the anterior and posterior parts of the glottis is

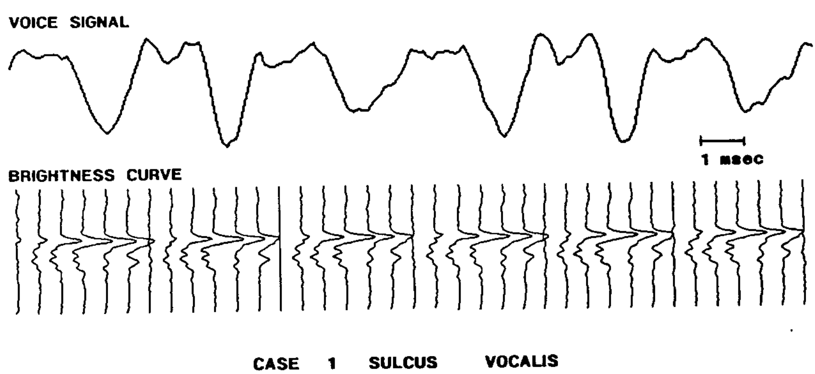


Fig. 2 A comparison between the acoustic waveform and the brightness curves for Case 1

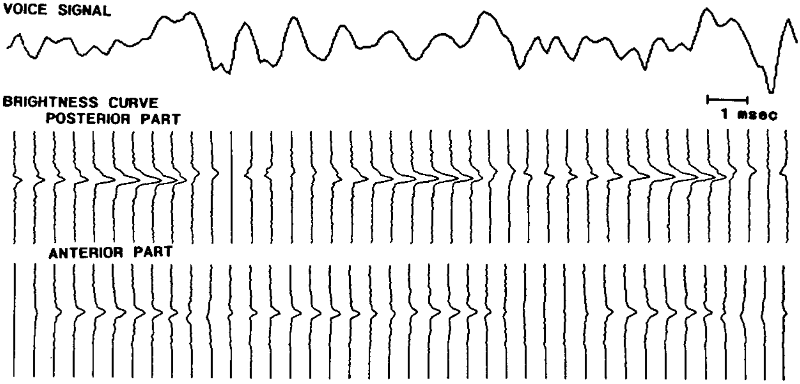


Fig. 3 A comparison between the acoustic waveform and the brightness curves for Case 2

related to the periodic fluctuation in the vibratory movement of the vocal folds.

The procedure for the recording and analysis of vocal fold vibration with the present system is simple compared to the conventional high-speed filming system. While the system is useful for practical purposes, a few technical improvements in the system's performance, particularly in the maximum frame rate, are still needed.

For clinical purposes, however, the present system has sufficient capability for the observation of pathological vibratory patterns and is useful as a practical unit.

It is thus expected that the present system would shed a new light for the understanding of physiological as well as pathological mechanisms of vocal fold vibration during voice production.

4. REFERENCES

[1] HIROSE, H. (1988), "High-speed digital imaging of vocal fold vibration", Acta Otolaryngol (Stockh), Supple. 458, 151-153.

[2] KIRITANI, S., IMAGAWA, H. and HIROSE, H. (1988), "High-speed digital image recording for the observation of vocal cord vibration:, In: Vocal Physiology: Voice Production, Mechanisms and Functions, (Ed. O. FUJIMURA), pp 261-268, Raven Press.

[3] KIRITANI, S., IMAGAWA, H. and HIROSE, H. (1990), "Observation of pathological vocal fold vibration using a high-speed digital image-recording system", Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, 24, 1-6.

FORMANT FREQUENCIES FOR INCREMENTALLY VARYING VOCAL TRACT

Hisashi J. Wakita

Panasonic Technologies Inc.
Santa Barbara, CA 93105, USA

ABSTRACT

This paper concerns the behavior of the formant frequencies for a dynamically varying vocal tract to investigate the sensitivity of the formant frequencies to different parts of the vocal tract. The sensitivity functions for diphthongs were obtained by simulating the varying vocal tract using measured vocal tract area functions for five steady-state vowels.

1. INTRODUCTION

A better understanding of the relationships between vocal tract configurations and their acoustic characteristics has always been important not only in the area of speech production, but also more recently in the inverse problem of estimating the vocal tract shape from acoustic data. Insight into the dynamics of the vocal tract is expected to especially contribute to better speech synthesis based on articulatory models and also is expected to clarify some of the physiological constraints to be imposed on a method for estimating either vocal tract shapes or the midsagittal view of the vocal tract based only on the resonance frequencies of the vocal tract. In some of the recent studies on static vowels, attention has been directed to the clarification of the detailed sound production processes. For instance, some vocal tract spatial parameters derived from the distribution of kinetic and potential energies inside the vocal tract at a particular resonance mode have been studied by Fant and others[2] to clarify the affiliation of formant frequencies with the particular part of the vocal tract and also to clarify the sensitivity of the formant frequencies to local perturbations in the cross-sectional areas of the vocal tract. Along this line of thinking, this paper concerns the behavior of the formant frequencies for a dynamically varying vocal tract. Particularly, the contribution of different parts of the vocal tract to increments in a given formant frequency were investigated together with the sensitivities of the formant frequencies to given parts of the vocal tract.

2. EXPERIMENT

As a preliminary step, the vocal tract area functions for five Russian vowels measured by Fant[1] were used to simulate the five diphthongs which typically appear in American English. Each area function was represented by the concatenation of 34 acoustic tube sections. In order to compute the formant frequencies for a dynamically varying vocal tract, the difference in the areas between two vowels was linearly interpolated into ten intervals. Between each two time successive intervals, the formant frequencies were computed by varying two sections at a time, from glottis to the lips, thus resulting in a total of 170 steps to represent the movement of the vocal tract from one vowel to the other. In computing the formant frequencies, a transmission line analog of the vocal tract was simulated on a digital computer, taking all the energy losses within the vocal tract and the lip radiation into account[3]. The glottis was assumed to be closed in this study. The transfer function of the vocal tract was computed by sweeping the frequency in two Hertz steps, so that the first three formant frequencies could be obtained by peak-picking, with an accuracy of one Hertz.

As a result of this analysis, a set of formant frequency increments, F_{ij}, can be obtained. In this case, ΔF_{ij} represents the increment of the i-th formant frequency affiliated with the j-th section of the area function. Thus, the total formant frequency increment is represented by the sum of all the ΔF_{ij}'s as given by

$$\Delta F_i = \sum_{j=1}^{N} \Delta F_{ij} \qquad (1)$$

The sensitivity of the formant frequency increment is then defined in this study as a ratio of the formant frequency increment to the area increment of the j-th section given by

$$S_{ij} = \frac{1}{K_i}\left|\frac{\Delta F_{ij}}{\Delta A_j}\right| \qquad (2)$$

where S_{ij} is normalized by the total sum of the ratios of the formant frequency increment to the area increment as given by

$$K_i = \sum_{j=1}^{N}\left|\frac{\Delta F_{ij}}{\Delta A_j}\right| \qquad (3)$$

This normalization is made so that the sensitivity of a section of the vocal tract will be bounded by zero and one.

3. RESULTS AND DISCUSSION

Fig. 1 shows the results for the vowel movements from /a/ to /i/ and from /a/ to /u/. The top figures show the distribution of the area increment along the vocal tract from the lips to the glottis. The successive three figures for each vowel movement are the sensitivity functions for the first three formant frequencies. In these two cases, the sensitivity for the vowel movement from /a/ to /i/ is rather dispersed along the vocal tract, whereas a relatively high localized sensitivity is observed in the lip and the pharynx regions for the vowel movement from /a/ to /u/. The most noticeable feature of the sensitivity functions among five vowel movements is that there is a place in the vocal tract where the area increment is relatively small and also the sensitivity experiences a local minimum. This is indicated by the small arrows in the figure. This location corresponds roughly to either the 8-th or 9-th section from the lips.

In the midsagittal view of the vocal tract, this location corresponds roughly to the boundary between the oral and pharyngeal cavities. This seems to be a rather strong physiological constraint in moving the vocal tract configuration from one vowel to another. Particularly, when the vocal tract configuration moves from a front vowel to a back vowel or vice versa, the area in-

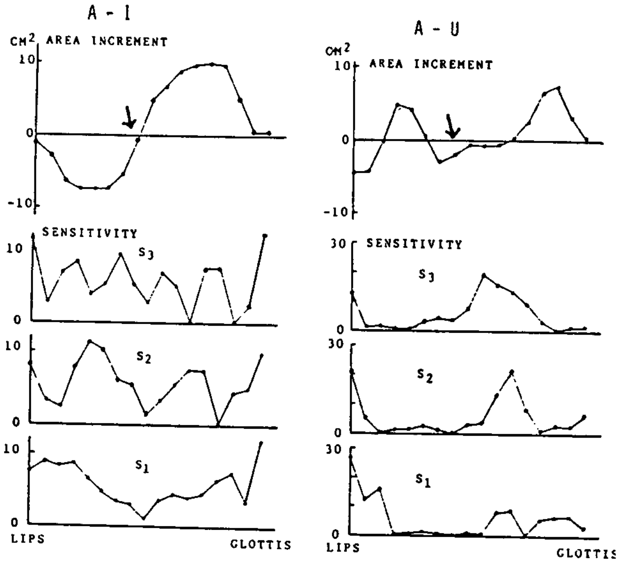


Fig. 1 Area increments and sensitivity functions for varying vocal tracts from /a/ to /i/ and from /a/ to /u/.

crement distribution becomes antisymmetric around this region, to maintain an approximate constant volume of the vocal tract.

Fig. 2 shows the vowel movement from /o/ to /i/ and from /o/ to /u/. The local minima of the sensitivity functions are again indicated by the arrows in the Figure. In the movement from /o/ to /i/, the lip area is highly sensitive, whereas in the movement from /o/ to /u/ the pharynx area as well as the lip area is sensitive.

Fig. 3 shows the vowel movement from /e/ to /i/. A similar tendency is also observed in this case, although the local minima of the sensitivity functions are somewhat obscured, due to the absence of an area increment over a rather broad region around the boundary between the oral and pharyngeal cavities. The sensitivity is high in the front cavity for this case.

4. CONCLUSION

The data obtained in this study will help clarify the relationships between dynamic vocal tract configurations and their acoustic correlates. The data will also be useful first in designing some speech synthesis rules on the basis of an articulatory model, second in gaining some insight into compensatory articulation, and third in finding some physiological constraints in the estimation of vocal tract dynamics from acoustic data.

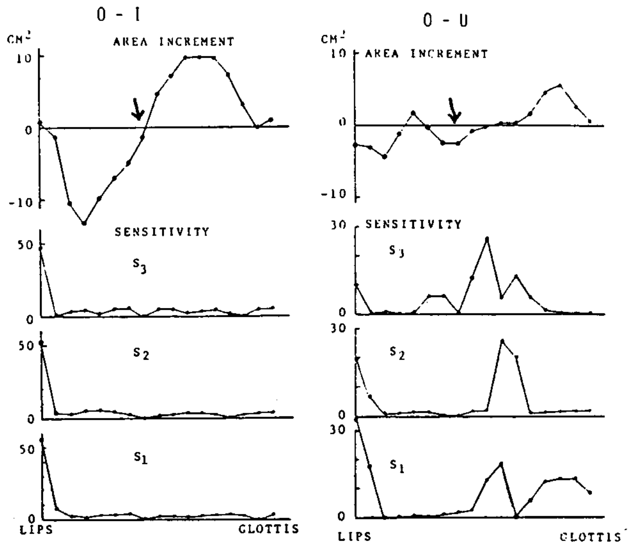


Fig. 2 Area increments and sensitivity functions for varying vocal tracts from /o/ to /i/ and from /o/ to /u/.

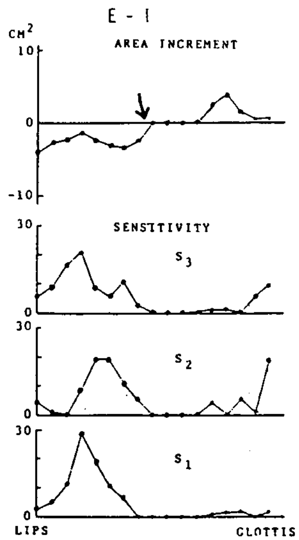


Fig. 3 Area increments and sensitivity functions for a varying vocal tract from /e/ to /i/.

6. REFERENCES

[1]. FANT, G. (1960), "Acoustic theory of speech production", The Hague : Mouton & Co.

[2]. FANT, G. and PAULI, S. (1976), "Spatial characteristics of vocal tract resonance modes", Almqvist & Wiksell Int. : Speech Comm. Seminar, Fant ed.., Stockholm, pp. 121–132.

[3]. Wakita, H. and Fant, G. (1978), "Toward a Better Vocal Tract Model", Quarterly Progress Status Report, Speech Transmission Laboratory, Royal Institute of Technology, Stockholm, Sweden, Number 1, pp. 9-29.

RECENT DEVELOPMENTS IN THE RESEARCH OF THE STRUCTURE OF VOWEL SYSTEMS

L.J. Bonder
Hogeschool van Amsterdam, dept. Logopedie/Akoepedie, Amsterdam

L.F.M. ten Bosch
Institute of Phonetic Sciences, University of Amsterdam, The Netherlands

ABSTRACT

The structure of vowel systems is discussed from the articulatory and acoustic point of view. Also, the relation between acoustic and perceptual properties of vowels is briefly dealt with. It is shown that the positions of monophthongs in a vowel system can be modelled in several ways by means of 'ease of articulation' and 'sufficient contrast'. The important models as found in the literature (e.g. vowel dispersion, Quantal Theory) are discussed in more detail.

1. INTRODUCTION

One of the most important means of communication between human beings, speech, is actually produced by accurate regulation of the subglottal air pressure and manipulation of the shape of the vocal tract. Consonant and vowel systems in languages are organized such that specific linguistic as well as phonetic (articulatory and perceptual) constraints are met: consonants and vowels serve as linguistic units but must be pronounceable and sufficiently contrastive at the same time. In this presentation, we present an outline of the present theories aiming at a structural description of vowel systems in relation to articulatory models. It is split up into three sections (2) the relation between articulation and perception, (3) the boundary of the vowel space, and (4) the internal structure of vowel systems.

2. RELATION ARTICULATION - PERCEPTION

For a phonetic description of vowel systems, one must look for articulatory, acoustic and perceptual differences between vowels that possibly underlie the phonological oppositions between them. At the phonetic side, we have to consider (2.1) the articulation-to-acoustics mapping, and (2.2) perceptual aspects of vowel sounds.

2.1 Articulation-to-acoustics

The computation of acoustic output from the vocal tract shape and the inverse problem, the computation of the shape from the output, constitute one of the main topics in speech production research. Most of the speech production models are based upon the source-filter theory. It has been demonstrated that speech generated by means of such models is hardly discriminable from natural speech ([11]).

The problem, how to relate vocal tract shape and acoustic output can be tackled in different ways: (1) in terms of electric LC-circuits; historically, this has been the usual paradigm originating from transmission engineering. (2) in terms of the n-tube representation of the tract. An n-tube is a tube with n segments with length l_i and area A_i. An example of this approach is given by nomograms [10], and the 'fibre'-concept [1, 5, 7]. It is also applied in the Quantal Theory ([22, 23]). (3) in terms of articulatory-based tract models. This approach, followed by Mermelstein, Maeda, Lindblom, Sundberg a.o., is characterized by the choice of a small subset of 'higher level' articulatory parameters that rule the 'lower level' tube parameters. (4) in terms of eigenfunctions of the Webster horn equation. This approach has been dealt with by [12], and, in a more loose mathematical way, in the distinctive regions theory, in [19].

These four approaches are in fact equivalent, as can be seen by considering the mathematics involved. However, all models have different starting points. Mathematically, the filter behaviour is fully described by:

$$S(z) \cdot A(z) = O(z)$$

where $S(z)$ and $O(z)$ denote the z-transforms of the input and output signal, respectively, and $A(z)$ the inverse filter. Ideally, this filter includes radiation, wall losses, etc. Such a point of view, however, only represents the technical aspects of the articulation-to-acoustics relation. It does not yield a quick insight into the relation between articulators and acoustic features.

Determination of the tract shape, while the acoustic output is given, is equivalent to decomposing $O(z)$ into the factors S and A. It is because of a priori assumptions and knowledge about these factors that such a decomposition makes sense. Thermal and viscous losses, wall vibrations and radiation effects are important when modelling the articulation-to-acoustics relation in detail. On the other hand, main effects can clearly be demonstrated by relatively simple tract models ([4, 5, 9, 10]).

An accurate calculation of the acoustic output of the tract is numerically quite involved but essentially straightforward. The *inverse problem*, however, is much harder to solve. In fact, up to now, the problem has been dealt with succesfully in the case of (sustained) vowels. In the dynamic case, the dynamic gesture is often reduced to a sequence of static articulatory positions.

It is well known that the inverse problem has no unique solution [1]. The acoustics-to-articulation relation is 'one-to-many'. The solution space, i.e. all vocal tracts producing the same acoustic output, is said to define a 'fibre' in the articulatory space. This fibre concept is well known in the general mathematical theory of mappings. In order to specify one unique exemplar from the fibre, additional constraints have to be defined. These constraints may be on the acoustic side (e.g. additional constraints on bandwidths), or on the articulatory side (e.g. minimality of an articulatory effort function).

2.2 Acoustics and perception.

Apart from the question how tube shape is related to acoustic output, the correspondence between acoustic output and perceptual features has also drawn much attention in the past two decades. This latter relation is prominent in the discussion on the structure of vowel systems. Ultimately, the structure of vowel systems is determined by linguistic (perceptual) oppositions between phonemes, and their (allophonic) realisations are bounded by physiological constraints. It has been shown that the structure of actual vowel systems is based upon the principle of perceptual contrast ([15, 16]; also [6, 9]). Several approaches have been suggested in the field of speech production. In [19] a model is proposed in which the first three eigenvectors of the Webster horn equation play a role in the determination of distinctive regions along the vocal tract. Due to the accentuation on symmetry along the tract and the criticism it not being able to describe dynamic details in some CV transients (/da/, /di/, du/) ([4]), this theory of distinctive regions still seems susceptible for some improvement. Strictly, the derivation of the results of [19] is valid in a neighbourhood of the neutral tract only; one cannot draw conclusions for more deviant formant positions.

A theory which dynamically combines articulatory gestures and acoustic output is put forward in the Quantal Theory (QT, [22, 23]). In its pure form, this theory states that the articulatory positions of which the acoustic output (in a way) is less sensitive to articulatory deviations are preferable to other positions (articulatory plateaus). The Quantal Theory predicts, in the case of vowels, the vowels that are likely be a member of a vowel system. The presuppositions of the Quantal Theory, however, still lead to discussion and have been questioned by many authors (cf. Journal of Phonetics, vol. 17), whereas the results are not convincing (cf. e.g. [8, 14]). It is generally believed, however, that the speech signal inherits 'quantal' phonetic properties as a consequence of non-linearities of the articulation-to-acoustics mapping and probably, the categorical perception of speech sounds. If quantality exists, it is probably a result of close approximations of formant frequencies ([2, 14, 21, 23]). In [2], the importance is stressed of the difference between F_3 and F_2 (instead of the classical F_2) as a classifier between front and back vowels. Approximations of formant frequency values are called 'focal points'; there is a relation between these focal points and the notion of 'plateaus' in the Quantal Theory. The cardinal vowels correspond to focal points with respect to the F_1 and F_2 (in case of /u/) and F_3 and F_2 in case of /i/ and /a/.

It may be clear that for a proper theory of the structure of vowel systems, based upon articulatory, acoustic and perceptual features of vowels, relations must be established between very different spaces each with their own metric, any mapping between them introducing non-linearities. We must relate the phonological observations of vowel systems, with the linguistic notion of opposition as a primary tool, with the psycho-physical properties of the human hearing system, with its spectral integration and masking behaviour. In this long sequence, we have to simplify the mappings we encounter on the way in order to be able to handle all relationships.

3. BOUNDARY OF THE VOWEL SPACE

In [17], the notion of "possible speech sound" is elaborated. Phonetically, the set of possible speech sounds is a subset of the total sound-producing potential of the vocal tract. From a phonological point of view, however, 'possibility' is a function of segmental features that are relatable to articulatory and perceptual attributes. Phonologically, the boundary of the vowel space is determined by the features [low], [back, round] and [front, spread], corresponding to the cardinal vowels /a/, /u/ and /i/, respectively. Along the dimensions [height], [backness] and [rounding], all other vowels take an intermediate position. Since the vowel coordinates on these dimensions are not uncorrelated ([back] is positively correlated with [round], [low] with [central], etc.), the dimension of the set of vowels in the phonological 3D space is somewhere between 2 and 3, rather than 3. Phonological analyses, however, are not capable to explain the actual boundary in the phonetic vowel space.

By using Maeda's statistical analyses of articulatory positions ([18]) it has been shown ([20]) that the boundary of the vowel triangle can adequately be simulated by putting specific lower and upper bounds to the tube segment areas. In [5] and [9], this phenomenon is studied by using the n-tube as articulatory model. These studies confirm that articulatory models using the 4-tube are capable of showing relevant details of the mapping articulation-to-acoustics ([4, 5, 9, 10]). In particular, the boundary of the vowel space in the 2D formant space can be described in terms of articulatory constraints. By examining the contour lines of opening degree of lossless 4-tubes, it is shown that the lowness of vowels is determined by this parameter [6]. The inverse problem can always be solved uniquely – in the lossless as well as in the lossy case – by constraining the tube shape ([9]) by means of an articulatory effort function, similar to the one applied in [1].

4. INTERNAL STRUCTURE OF VOWEL SYSTEMS.

Apart from the question how tube shape is related to acoustic output, the correspondence between articulation, acoustic output and perception has drawn much attention in the past two decades, particularly in the discussions on the structure of vowel systems. This structure may be considered to be determined by the articulatory possibilities and constraints on the one hand, and the perceptual demands on the other. One of the rules which vowel systems seem to obey is

the principle of perceptual contrast.

In [15], such a rule was implemented in a computational model for the prediction of vowel systems demanding *maximal* perceptual contrast. The maximal contrast had been established by the minimization of

$$\sum_{i=1}^{N}\sum_{j=1}^{i-1}\frac{1}{D_{ij}^2} \qquad (1)$$

where D_{ij} is the Euclidean distance between any two vowels i and j in the perceptual space; N denotes the number of vowels in the system.

The results of the computation show among others an abundance of high (central) vowels for higher values of N. The L&L-model has been further elaborated in [3], considering a perceptual distance measure based on the difference between the so-called auditory spectra of any two vowels i and j. In their paper, B&L adopt two ways of interpreting the concept of auditory spectrum: (a) as loudness density pattern, and (b) as auditory filter output. The metric of the perceptual measure is no longer Euclidean as in (1) but generalized to the L_p-metric.

The computations have again been carried out under the constraint of maximal perceptual contrast. The modifications appeared to lead to a reduction of the number of high vowels for either interpretation of the auditory spectrum.

The B&L-approach has also been applied under the constraint of *sufficient* contrast in order to compute the best 50 N-vowel systems for some values of N. The frequency of occurrence has been computed for each vowel independently and it turned out that there is a tendency towards more contrast as N increases.

Both the L&L and the B&L-model do not take into account the articulation of vowels. The implementation of a simple articulatory model like the n-tube model (where $n \leq 4$) already accounts for the most prominent results obtained for much larger values of n (cf. [6, 7]). In [6] a method is proposed for the prediction of modal N-vowel systems (i.e. the collection of most occurrent systems for each N). The method is based on the assumption that modal N-vowel systems (denoted as MVS_N) are hierarchically ordered:

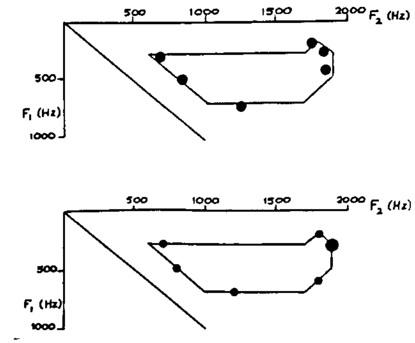


Figure 1: Vowel system prediction; (a) shows the (normalized) modal 6-vowel system; (b) shows the predicted system.

$$\mathrm{MVS}_{N+1} = \mathrm{MVS}_N \cup NV$$

where NV is the set of one vowel that is the most contrastive with all vowels in MVS_N. The boundary of the discretized 2D vowel space is determined by matching to normalized vowel data ([16]). 'Repelling forces' have been defined in the articulatory space A as well as in the vowel space F:

$$d_A(v_1, v_2) \;>\; A/\sqrt{N} \qquad (2)$$
$$d_F(v_1, v_2) \;>\; B/\sqrt{N} \qquad (3)$$

where A and B denote constants, and N the number of vowels. If $\mathrm{MVS}_3 = \{a, i, u\}$, it turns out that the logarithmic vowel space (especially up to $N = 6$) gives the best results (cf. fig. 1). It seems that the implementation of articulatory constraints is mainly important for the definition of the vowel space *boundary*, although the constraints may be used to model non-modal systems that contain more 'interior' vowels.

In [7], a vowel system model is proposed that is based on maximal acoustic contrast together with a minimal articulatory effort criterion. The vowel system quality parameter Q is defined as

$$Q = D_A^2 + S \cdot (D_F - 1)^2$$

where D_A is the total articulatory system effort, D_F is the total perceptual system discrimination, and S a slack variable as used in optimization problems (S being a large positive number). D_F is computed by means of the intervowel confusion probability between two vowels

$$p(v_1, v_2) = \exp(-\alpha \cdot d_F(v_1, v_2))$$

where α is a scaling factor. Also this model, which is more fundamentally based upon probability arguments, is able to explain the main properties of vowel systems.

Vowel system models may be further elaborated by implementing submodels that describe in a more sophisticated manner the non-uniformity of the articulation-to-acoustics relation and the perceptual contrast. Studies in the 3D formant space, performed in [9] and [21], show the great dependency of the resulting model systems on variations in parameters controlling the perceptual distances between vowels.

The paper of [13] suggest a refinement of the metric used for the measurement of the perceptual contrast between nearby vowels. One of their results show that the best metric for nearby vowels is the 2D Euclidean metric after bark transformation of F_1 and F_2. Another important goal is the refinement of the overall articulation-to-acoustics relation. The Quantal Theory (QT; [22, 23]) gives us some qualitative insight into the non-uniformity of this mapping. Although the name QT is rather misleading as the relation is *continuous*, it shows quite clear its message that the acoustic change per 'unit' of articulatory change is nor uniform over the entire formant space, nor isotropic in each point of the space. As the anisotropy is greater towards the /u/ and /i/ edge of the vowel space, QT might help to increase the goodness-of-fit of vowel system models with respect to high vowels.

Acknowledgements

This study has been supported by the Dutch Organization for the Advancement of Pure Scientific Research NWO (project 300-161-030).

5 REFERENCES

[1] ATAL B.S., CHANG, J.J., MATHEWS, M.V., and TUKEY, J.W. (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. J. Acoust. Soc. Am. 63. p. 1535 - 1555.

[2] BADIN, P., PERRIER, P., BOË, L.-J., and ABRY, C. (1990). Vocalic nomograms: Acoustic and articulatory considerations upon formant convergences. J. Acoust. Soc. Am. 87. p. 1290-1300.

[3] BLADON, R.A.W. & LINDBLOM, B. (1982). Modeling the judgment of vowel quality differences. J. Acoust. Soc. Am. 69. pp. 1414-1422.

[4] BOË, L.J., and PERRIER, P. (1990). Comments on 'Distinctive regions and modes: A new theory of speech production', by M. Mrayati, R. Carré, and B. Guérin. Speech Communication 9. p. 217 - 230.

[5] BONDER, L.J. (1983). The *n*-tube formula and some of its consequences. Acustica 52, p. 216 - 226.

[6] BONDER, L.J. (1986). A prediction method for modal n-vowel systems. Procs. Inst. Phon. Scs. Amsterdam, vol. 10. p. 73-90.

[7] TEN BOSCH, L.F.M., BONDER, L.J., and POLS, L.C.W. (1987). Static and dynamic structure of vowel systems. Procs. 11th Intern. Congress Phon. Scs., vol. 1. p. 235-238.

[8] TEN BOSCH, L.F.M. and POLS, L.C.W. (1989). On the necessity of quantal assumptions. Questions to the Quantal Theory. Journal of Phonetics 17. p. 63 - 70.

[9] TEN BOSCH, L.F.M. (1991). On the structure of vowel systems. An extended dispersion model. PhD-thesis (in preparation). University of Amsterdam, The Netherlands.

[10] FANT, G. (1960). Acoustic Theory of Speech Production. Mouton & Co., 's-Gravenhage.

[11] HOLMES, J.N. (1973). Influence of glottal waveform on the naturalness of speech from a parallel formant synthesizer. IEEE Trans. Audio-Electroacoust. AU-21. pp. 298-305.

[12] KARAL, F.C. (1953). The analogous acoustical impedance for discontinuities and constrictions of circular cross section. J. Acoust. Soc. Am. 25. p. 327 - 334.

[13] KEWLEY-PORT, D., and ATAL, B. (1989). Perceptual differences between vowels located in a limited phonetic space. J. Acoust. Soc. Am. 85. p. 1726-1740.

[14] LADEFOGED, P. and LINDAU, M. (1988). Modeling articulatory-acoustic relations. UCLA Working Papers 70. p. 32 - 40.

[15] LILJENCRANTS, J. and LINDBLOM, B. (1972). Numerical simulation of vowel quality systems: the role of perceptual contrast. Language 48, p. 839 - 862.

[16] LINDBLOM, B. (1986). Phonetic universals in vowel systems. In: Experimental Phonology (J. Ohala and J. Jaeger, eds.). Academic Press, Orlando, Florida. p. 13 - 44.

[17] LINDBLOM, B. (1990). On the notion of "possible speech sound". Journal of Phonetics 18. p. 135 - 152.

[18] MAEDA, S. (1979). An articulatory model of the tongue based on an statistical analysis. J. Acoust. Soc. Am. Suppl. 1. Vol. 65, S 22.

[19] MRAYATI, M., CARRÉ, R., and GUÉRIN, B. (1988). Distinctive regions and modes: A new theory of speech production. Speech Communication 7. p. 257 - 286.

[20] PERRIER, P., BOË, L.J., MAJID, R., and GUÉRIN, B. (1985). Modélisation articulatoire du conduit vocal: exploration et exploitation. In: Proceedings of the 14emes Journées d'Etudes sur la Parole (Groupement des Acousticiens de Langue Française, Paris). p. 55 - 58.

[21] SCHWARTZ, J.L., BOË, L.J., PERRIER, P., GUÉRIN, and ESCUDIER, P. (1989). Perceptual contrast and stability in vowel systems: a 3-D simulation study. Proceedings of the Eurospeech Conference, Paris. p. 63-66.

[22] STEVENS, K.N. (1972). The quantal nature of speech: evidence from articulatory-acoustic data. In: Human communication: A unified view (E. David and P. Denes, eds.). McGraw-Hill, New York. p. 51 - 66.

[23] STEVENS, K.N. (1989). On the quantal nature of speech. Journal of Phonetics 17. p. 3 - 45.

NOISE AND APERIODICITY IN THE GLOTTAL SOURCE: A STUDY OF SINGER VOICES

Perry R. Cook (PRC@CCRMA.STANFORD.EDU)

Center for Computer Research in Music and Acoustics (CCRMA), Stanford Univ.

ABSTRACT

Methods for extracting, analyzing, and visualizing the time-domain nature of aperiodicities in the glottal source are presented. A study of noise in the voice was conducted using 12 trained singers. Average Normalized Noise Power (NNP) was computed as a function of pitch and volume. The time domain nature of the noise signal is investigated and compared to theoretical predictions from the principles of fluid dynamics.

1. INTRODUCTION

The quasi-periodic oscillations of the glottis exhibit small period-to-period deviations in the waveform [8], much of which is brought about by bursts of noise in the oscillator itself. This paper investigates aperiodicities in steady state sung tones, using new techniques for extraction, visualization, and quantification.

2. EXTRACTION TECHNIQUES

To study the aperiodic component of a quasi-periodic signal, some method must be used to identify and separate the periodic and aperiodic components. The two methods used in this study operate in the time domain, and yield slightly different results because of the definition of periodicity each assumes. The first involves using a least-squares periodic predictor to yield an error signal representing the component of the signal which is not periodic. The second method uses period similarity processing [10], with the added improvement of sinc-interpolated sampling rate conversion [9]. Each period of a steady state vowel is compared to all others and resampled to yield a least-squares difference. All such resampled periods are then averaged as vectors to yield a prototype period, which contains no noise component in the limit as the number of periods approaches infinity. The prototype is subtracted from each period to yield the period residuals. This method has the advantage that the time domain connection between the signal and residual periods is preserved. Figure 1 shows the waveforms of a male singing the vowel /ʌ/ (bug) at 100 Hz., the periodic component, and the residual.

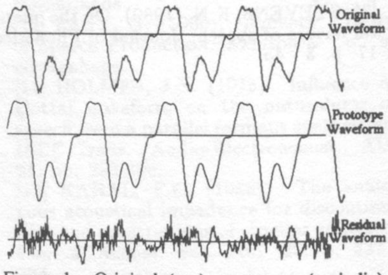


Figure 1. Original (top), prototype (periodic), and amplified residual of male vowel /ʌ/ sung at 100 Hz.

3. VISUALIZATION TECHNIQUES

Period synchronous methods of analyzing the extracted noise residual signals were used. The techniques involve identifying the periods of the original signal (already accomplished if extraction is done by the period similarity method). Identification of the periods involves detection of some time domain feature using a method such as low-pass filtering and zero-crossing detection. Each detected period yields a pointer into the time-domain residual signal, and thus the noise 'periods' can be subdivided (into 6 sections for this study) to inspect the behavior.

In the period-synchronous noise power analysis method, the noise power (sum of squares) is computed in each of the sub-sections. These powers can then be plotted in three dimensions, with height representing power, one axis representing the period number, and one axis representing the position within the period. Inspection of this noise period power surface will show clear ridges and valleys (running in the direction of the period number axis) if the signal contains pulsed noise. The duty cycle can be deduced from the width of the ridges and valleys, and the dynamic range of the noise can be deduced from the ratio of the heights of the ridges and valleys. Averaging across the power surface in the direction of the ridges and valleys yields an estimate of the average noise power at particular times within a typical period. Figure 2 shows the noise power surface of the vowel of Figure 1. The placement of the ridges and valleys shows that noise is more likely to occur as the glottal folds open and close, and less likely when the glottis is completely closed or completely open.

Another analysis technique involves performing Discrete Fourier Transforms (DFTs) on each of the residual period subdivisions. Each DFT can be used to compute a power spectrum, and the power spectra corresponding to a particular period position can be averaged across all periods, yielding an

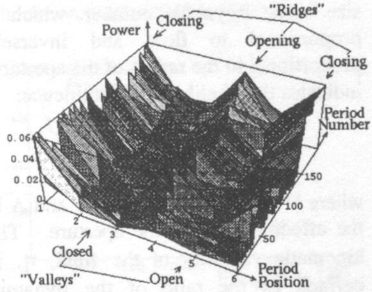


Figure 2. Noise period power surface of male vowel /ʌ/ sung at 100 Hz. Opening and closing phases of the glottis are noted.

estimate of the power spectrum of the noise at that position within a typical period. These spectra can be plotted in 3D as intensity (height) versus period position and frequency. Figure 3 shows the period spectral surface of the male vocal tone of Figure 1. The spectral energy clearly shifts toward higher frequencies at the instant where the glottal folds are opening.

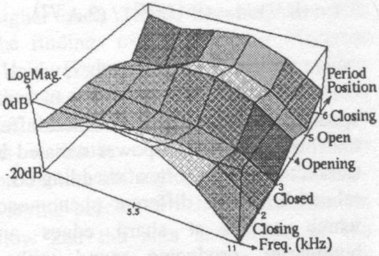


Figure 3. Noise period spectral surface of male vowel /ʌ/ sung at 100 Hz. Opening and closing phases of the glottis are noted.

4. GLOTTAL FLUID DYNAMICS

The passage of air at sufficient velocity through an aperture causes turbulent streaming and noise is generated. The flow is zero when the time varying aperture is closed, and the turbulence ceases if the aperture opens sufficiently or the flow decreases. The basic fluid dynamic equations quantifying turbulent jet formation and noise radiation are expressed in terms of flow and aperture size. The Reynolds number, which is proportional to flow and inversely proportional to the radius of the aperture, indicates the likelihood of turbulence:

$$Re = (2U) / (\eta (A \pi)^{1/2})$$

where U is the volumetric flow and A is the effective area of the aperture. The kinematic viscosity of the fluid, η, is defined as the ratio of the dynamic viscosity to the density, and is about 0.15 cm^2 per second for dry air. Turbulent streaming is likely if the Reynolds number is greater than a critical quantity, Re_{crit}, which is about 1,000 for a rectangular slit. If turbulence is present, noise is generated with a power proportional to V^8, or proportional to $(U/A)^8$ as expressed in terms of volume flow and area. The center frequency of the principal peak in the spectrum of the turbulent noise is given by:

$$f = (S\ V) / d = (S\ U\pi^{1/2}) / (2\ A^{3/2})$$

where S is the Strouhal number, which is 0.15 for the center frequency of noise spectral density. Tube resonances affect the formation of, and power radiated by turbulent jets. Vortex shedding is a related but quite different phenomenon which occurs at sharp edges and boundaries, producing sound with a power which depends on lower powers of the flow-to-area ratio. Hirschberg [3] gives power relationships of $(U/A)^4$ and $(U/A)^6$ for turbulent sound radiation in a tube, and vortex dipole sound radiation in a tube, respectively.

Figure 4 shows the characteristics of a typical cycle of oscillation of the glottal folds. Views a) and b) are drawings of the superior and cross sectional views of the glottal folds. Graphs c) and d) show the effective area and volumetric flow (flow glottograph). Assuming a maximum flow of 300 cm^3 per second, and maximum dimensions of 2.6 mm by 11 mm for the vocal folds, the Reynold's number, radiated power, and center frequency of the radiated noise power are graphed as e) f) and g) of Figure 4.

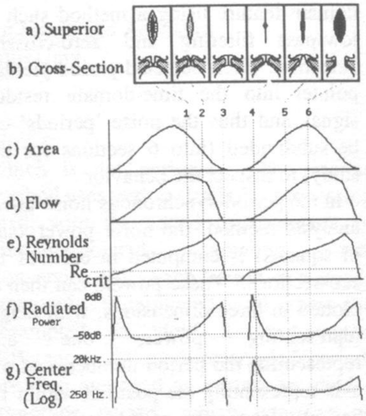


Figure 4. Phases of glottal oscillation, with area, flow, Reynolds number, radiated noise power, and noise spectrum center frequency.

The simple analysis of Figure 4 assumes that turbulence is instantly born when the dimensions and flow quantity are suitable, and the disturbance dies as quickly. A more detailed analysis of the behavior of pulsed turbulence was done by Kingston [6]. These studies investigated the effects of turbulent jets in tubes driven by pulsating sources of flow. The ratio of normalized pulsation frequency, Ω, to the Reynolds number was identified as an important measure of turbulent behavior. Ω is defined by:

$$\Omega = r^2\omega / \eta = 2AF_0 / \eta$$

where r is the radius of the tube, A is the tube cross-sectional area, and F_0 is the frequency of phonation in Hz.

Given the large frequency range of the singing voice, and allowing for large deviations in tube cross-sectional area depending on the vowel, Ω can range from 100 at 50 Hz. in an /μ/ vowel, to 10^5 at 2000 Hz. in an /a/ vowel.

Three regimes of pulse-turbulence interaction were observed by Kingston, corresponding to high, medium, and low ratios. For low pulsation frequencies (Ω / Re < 0.04), the flow is quasi-steady and follows the behavior indicated by the analysis of Figure 4. For high pulsation frequencies (Ω / Re > 0.1), the turbulence is steady and independent of flow pulsations. For intermediate frequencies, the relation between turbulence and pulsation is complex, and is characterized by vortex resonance phenomena.

The average value of the Reynolds number from Figure 4 is 2750. Assuming a minimum vocal tract tube area of 0.15 cm^2, the transition region from pulse-turbulence interaction to non-interaction lies between 55 and 140 Hz. The maximum Reynolds number is 5860, yielding a transition region bounded by 120 and 300 Hz. From these calculations, pulsed turbulence is possible at phonation frequencies below 200 Hz. Even allowing for large deviations in the assumed parameters of flow and glottal area, it is still expected that low notes sung by bass singers might exhibit pulses, or perhaps dual pulses. As simulated by Iijima, Miki, and Nagai [5], periodic vortex shedding is also expected in the glottal region.

5. NOISE IN SINGER VOICES

Twelve highly trained singers (Bel Canto) were asked to produce sung tones at three dynamic levels without vibrato on the the neutral vowel /ʌ/ across their entire comfortable singing range.

5.1. Average Noise

Figure 5 shows the average Normalized Noise Power (NNP) of all singers as a function of phonation pitch. The least-squares fit through the points of an exponential function of the form $k_1 * f^{k_2}$ is $59 * f^{-1.2}$, indicating a relationship between noise power and phonation pitch which is close to 1/f.

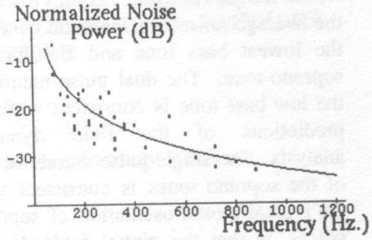


Figure 5. Normalized Noise Power (NNP) of the sung tones of 12 singers, graphed as a function of phonation pitch.

Given the fluid dynamic predictions that radiated noise power varies as a high power of flow, it may seem contradictory that NNP in singer voices was largely independent of dynamic level, and inversely proportional to frequency. A study of airflow in singer voices [7] found that flow increases slightly with both increasing pitch and loudness, but often airflow decreases in higher tones. This is also consistent with the findings of Cavagna and Margaria [1]. Higher tones are very often produced with a more 'pressed' voice, and the overall glottal resistance changes as a result. The nature of noise production in the glottis is that of a time-varying process which is dependent on flow and the area and shape of the aperture, so it is likely that any increase in flow is offset by changes in the time-varying area function. In the falsetto register there is a direct relationship between phonation frequency and flow [4], so there is a likelihood of higher noise power for increasing frequency in this range. All of the male test subjects showed an increase in noise power when entering the falsetto register, and most of the falsetto tones exhibited an increase in noise power with increasing frequency.

5.2. Pulsed Noise

The noise period power surfaces were

calculated for all singers, and averages were taken across the data at six points within the glottal cycle. Figure 6 shows the average noise characteristic curve of the lowest bass tone and the highest soprano tone. The dual pulse nature of the low bass tone is consistent with the predictions of the fluid dynamic analysis. The single pulse nature of two of the soprano tones is consistent with the typical glottal oscillation of soprano voices, in that the glottal folds do not close entirely. This is easily seen by the fact that the principal peak in the noise curve occurs when the glottal folds are in the 'closed' position.

All bass subjects exhibited dual pulse behavior in the low register, and on some tones in the high register. The alto and tenor subjects exhibited a shift from dual pulse at the low register to single pulse noise behavior in the high register. The Noise Dynamic Range (NDR), defined as the ratio of the highest noise power sample to the lowest, decreased weakly with pitch from about 10 dB in the lowest tones to 1 dB in the highest. No clear behavior as related to dynamic level was evident in any of the voices studied.

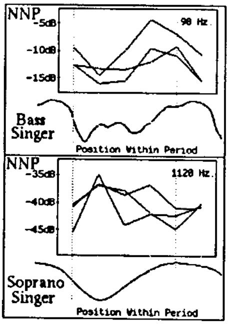


Figure 6. Normalized Noise Power (NNP) as function of period position for lowest bass and highest soprano tones, for three dynamic levels.

10. REFERENCES

[1] Cavagna, G. A. and R. Margaria (1968), "Airflow Rates and Efficiency Changes During Phonation," *Annals of the New York Academy of Sciences*, 155, 152-164.

[2] Cook, P. R. (1991), "Identification and Control of Parameters in an Articulatory Vocal Tract Model, With Applications to the Synthesis of Singing," Ph.D. Dissertation, Dept. of Electrical Engineering, Stanford Univ.

[3] Hirschberg, A. (1990), "Some Fluid Dynamics Aspects of Speech," *Fourth Colloq. Signaalanalyse en Spraak, COLSAS*, Instituut voor Perceptie Onderzoek, Eindhoven.

[4] Isshiki, N. (1965), "Regulatory Mechanism of Voice Intensity Variation," *J. Sp. Hearing Res.*, 7, 17-29.

[5] Iijima, H., N. Miki, and N. Nagai (1988), "Viscous Flow Analyses of the Glottal Model Using a Finite Element Method," *2nd Mtg. ASA & ASJ*, NN10.

[6] Kingston, G. C. (1975), "Experimental and Theoretical Studies of Pulsating Turbulent Flow," Ph.D. Dissertation, Clarkson College of Technology.

[7] Rubin, H. J., M. LeCover and W. Vennard (1967), "Vocal Intensity, Subglottic Pressure and Air Flow Relationships in Singers," *Folia Phon.*, 19, 393-413.

[8] Schumacher, R. T. and C. D. Chafe, (1990) "Detection of Aperiodicity in Nearly Periodic Signals," *Proc. IEEE Conf. ASSP*, Albuquerque, NM.

[9] Smith, J. O. and P. Gossett (1984), "A Flexible Sampling-Rate Conversion Method," *Proc. IEEE Conf. ASSP*, San Diego, CA, March.

[10] Yumoto, E., Y. Sasaki, and H. Okamura (1984), "Harmonics-to-Noise Ratio and Psychophysical Measurements of the Degree of Hoarseness," *J. Sp. Hearing Res.*, 27, No. 2-6.

THE MEASUREMENT OF THE ACOUSTIC TRANSFER FUNCTION AND THE AREA FUNCTION OF THE VOCAL TRACT : METHODS AND LIMITATIONS

B. Guérin

Institut de la Communication Parlée, Grenoble, France

ABSTRACT

In order to gain a greater insight into speech production phenomena, a better knowledge of both vocal tract acoustic transfer function and area function is needed. In these two kinds of investigation, many techniques can be used. We are going to analyse the different methods allowing us to obtain the correct results and their limits.

1. MEASUREMENTS OF VOCAL TRACT TRANSFER FUNCTION

Classical techniques for direct investigation of the acoustic characteristics of the vocal tract are based on the transcutaneous excitation of the tract near the glottis. The first measurements have been reported by Van den Berg [3]. Fant [10] later adopted a similar method. Well know results have been given by Fujimura & Lindquist [12] and this improved method is called sweep-tone measurement. In these different experiments, they used a pure tone signal swept in frequency as excitation. The source can be a small loudspeaker or a high-quality moving-coil-type electromagnetic transducer. In a session of data acquisition, a microphone picked out the sound at the mouth opening. The subject held the intended articulation as constant as possible during sweeping. A sweeping from 100 to 5 000 Hz took about 8,5 s. The major advantage of this method is the continuous frequency response curves that can be obtained. But three main disadvantages remain :

- The complex and particular experimental set up.
- A too long measurement time.
- No phonation is allowed during the process, but auditory feedback may be necessary for precise articulation configuration.

In order to resolve some of these problems, an improved method has been proposed by Castelli & Badin [5] [6] : the vocal tract was excited with white noise, which allows for auditory control. The output signal picked up by the microphone is directly sent to headphones : thus the subject can "hear" the configuration he is articulating. The signal is at the same time digitally recorded and further processed by averaging FFT spectra over a long period of time to produce good transfer functions. But the time measurement is still too big (around 10 s) and no phonation is possible.

A last improvement is proposed by Djeradi et al [9]. A technique developped earlier for acoustic room characterization has been adapted. The method is based on the impulse response of the vocal tract when it is excited throught the skin near the larynx. It is very difficult to produce a good impulse for excitation, so we use a pseudo-random sequence as input signal. The computation of the cross-correlation between the excitation and the output signal allows us phonation during the measurment process. The transfer function is obtained directly through the FFT of crosscorrelation. Finally, the measurement time is as small as about 100 ms, and the frequency resolution on the final result is of 10 Hz. As shown on figure 1, we can measure the transfer function with closed glottis condition and in phonation condition. With this method, we can consider measurement of the transfer function when the vocal tract is slowly moving, for example during vowel-vowel transition.

In conclusion, it is clear we can now measure the transfer function of the vocal tract with the following features :

- measurement time : about 100 ms
- accuracy in the frequency domain : about 10 Hz
- feasable measurement during phonation
- possibility to follow the variation of the transfer function during slow transitions

2. MEASUREMENTS OF THE AREA FUNCTION

The determination of articulatory parameters such as area functions or other representations of vocal tract shape is a long standing problem in speech research. This problem can be viewed accordingly in two ways : 1/ By direct measurement of the area using cineradiography, magnetic resonance imaging, ultrasonic scan, or X-ray microbeam. 2/ as an inverse problem from natural signal or output signal of external excitation. We are going to recall these different techniques and their limitations.

2.1. Direct measurement of vocal tract area

- X-ray photography an cineradiograph.

The first X-ray studies took place around 1925-1930. But the first well know were those from a Russian made by Fant in 1951. We obtained a mid-lateral saggital view of the vocal cavity. From the photography, we must first draw the outline of the vocal tract and after using a grid for example, we can measure the saggital distance of each cross section in order to define the centre line and the length of the vocal tract. Many errors occur during the processing of this kind of date :

- The boundary of the vocal tract is not always easy to define, in particular near the larynx.
- The choice of the method of outline to cross section magnitude conversion may give different values, especially on the total length of the vocal tract.

But a last operation must be done : the saggital distance → area cross - section conversion. Indeed the cross section shape of the vocal tract is very different according to position along the centre-line and even with the value of saggital distance. Usually, we use the Heinz & Stevens models where $A = \alpha d^{\beta}$. Computed tomography allowed us to obtain data of these cross-section shapes and a set of coefficients α and β were determined for several specific regions in the vocal tract [31]. Considering all the difficulties of this method, X-ray measurement, now improved with the cineradiographic records, gives very good information about the size and the shape of the vocal tract [17] [1] [4]. However, its use is restricted by the great deal of time and tedious work of processing involved.

Magnetic Resonance Imaging (MRI) [11] is a technique which allows the same kind of study as the X-ray method. Contrary to X-ray imaging, there are no severe problems owing to dosage limitations, but, for the moment, relatively long acquisition time make it impossible to investigate many speech sounds.

X-ray microbeam and ultrasonic scan often allow only the definition of a part of the vocal tract, usually the dorsal surface of the tongue [30] [26] [41].

2.2. Estimation of the area function from the speech signal

Considering that it is difficult to obtain the cross-sectional area function of the vocal tract from direct measurements, many researchers have tried to estimate it from acoustic data.

The first investigations were based on the inverse problem : determination of the vocal tract shape from transfer function or formant frequencies [32] [18] [20] [13] [2] [29]. But, as recalled by Mermelstein, Schroeder, Heinz and Sondhi [35], it is well known that the transfer function of a lossless vocal tract (much less that of a tract with loss) does not uniquely specify it's area function. An alternative is to try to resolve the non uniqueness by imposing "reasonable" constraints on the tongue and other articulators. But nothing allows us to judge if the result is the actual area function and not just any function which will produce the same sound.

Another approach was proposed, notably Wakita, Nakajima et al [44] [45] [46] [27], who estimated area functions from speech signals. The method is based on linear prediction analysis. If corrections on formant frequencies and bandwiths are made in order to compensate for the differences in boundary conditions and losses between the linear prediction model and actual speech production, reasonably results are obtained. Several problems remain in both of these approaches :

- the area function can be determined only to within an unknown factor,
- the vocal tract length is not directly available from the speech signal or the transfer function.

Wakita, Paige and Kirlin [28] [19] proposed algorithms to estimate this length and the various estimators give an accuracy of better 5 %. The approach which seems the most promising was proposed first by P. Mermelstein, and consists of considering the impedance function of the vocal tract measured at the lips. It has been shown that this characteristic allows the definition of a unique relation between the input impedance at the lips and the cross-section area function of the vocal tract. Sondhi has studied different aspects of the determination of area function [36] [37] [38] [39] [43] in depth. He shows that under plane wave approximation for a lossless vocal tract, as well as for tracts with certain types of distributed losses, we can reconstruct the area function with good concordance between calculated results and actual shape. Furthermore, J.P. Lefèvre, R. Descout, et al [8] [42] [22] used measurements at the lips of response to an impulse acoustic pressure wave for determining the vocal tract area function. This area function was obtained either by deconvolution, or by successive approximations of a modeled vocal tract, the search being made for the constrictions by decreasing order. In all cases, 20 to 40 measurements per second can be obtained.

In order to both improve the accuracy of the reconstructed area function and increase the frequency of measurement, P. Milenkovic [24] [25] has proposed a novel aspect of the procedure based on acoustic pulse reflection of an excitation at the lips. The results show that the period can be as short as 5 ms but only if articulatory is not too big.

These methods should be improved in different ways but two constraints remain : 1/ all of these need an impedance tube which is "connected" to the lips by a flexible coupling. There is bound to be some kind of unnaturalness in the articulatory movements. 2/ all the measurements must be made without phonation. This point could be neglected in the case of vocalic configurations, but it is very important for the acoustic characteristics of sounds like fricatives.

3. CONCLUSION

Different methods used for measuring the acoustic transfer function give accurate results. The determination of area functions is more difficult. Here different techniques are used and these give natural and verifiable results, using the driving point impulse response. Furthermore, much current research aims to obtain the area functions from the speech signal in order to achieve an efficient coding for transmission [21] [33] [34] and neural networks should also offer a new line of investigation [40].

Two ways should be explored (or continued) for increasing the quality of results : 1/ a better time model of the vocal tract, 2/ a mixed-method allowing the simultaneous measurement of transfer function and area function. Each of these studies have a specific application, firstly when we use a speech signal as input, and secondly when we use a synthetic external excitation.

REFERENCES

[1] AUTESSERRE, D., ROSSI, M., SARRAT, P., GIRAUD, G., VISQUIS, R., DEMANGE, R., CHEVROT, L. (1979), "Exploration radiologique de l'orpharynx, de l'hypopharynx et du larynx en phonation", *Proc. Séminaire Larynx et Parole (GRENOBLE)*, 45-74.
[2] BECKMANN, D.A., WOLD, D.C., MONTAGUE, J.C. (1983), "A nominuasive acoustic method using frequency perturbations and computer-generated vocal-tract shapes", *J.S.H.R., 26*, 304-314.
[3] VAN DEN BERG, S. (1955), "Transmission of the vocal cavities", *J.A.S.A., 27*, 161-168.
[4] BOTHOREL, A. (1984), "Apport de la radiocinématographie à la recherche phonétique. Phonétique instrumentale et linguistique", *Journées d'Etudes du 15 mai 1984*, 55-88.
[5] CASTELLI, E., BADIN, P. (1988), "Vocal tract transfer functions measurements with white noise excitation. Application to the naso-pharyngeal tract", *7th FASE Symposium (EDINBURGH)*, 415-422.
[6] CASTELLI, E., PERRIER, P., BADIN, P. (1989), "Acoustic considerations upon the low nasal formant based on nasopharyngeal tract transfer function measurements", *Proc. EUROSPEECH (PARIS), 2*, 412-415.
[7] CHARPENTIER, F. (1982), "Application of an optimization technique to the inversion of an articulatory speech production model", *Proc. ICASSP*, 1984-1987.
[8] DESCOUT, R., TOUSIGNANT, B., LECOURS, H. (1976), "Vocal tract area function measurements : two

time-domain methods", *Proc. ICASSP, 75*,78.

[9] DJERADI, A., GUERIN, B., BADIN, P., PERRIER, P. (1991), "Measurement of the acoustic transfer function of the vocal tract : a fast and accurate method", A paraître dans *Journal of Phonetics*.

[10] FANT, G. (1962), "Formant Bandwith Data", *STL Q.P.S.R., 1*, 1-2.

[11] FOLDVIK, A.K., HUSBY, O., KUAERNESS, J. (1988), "Magnetic resonnance imaging", *7th FASE Symposium (EDINBURGH)*, 423-428.

[12] FUJIMURA, O., LINDQVIST, J. (1970), "Sweep-tone measurements of vocal-tract characteristics", *J.A.S.A., Vol. 49, 2*, 541-558.

[13] FUJISAKI, H., OBATA, S., TAZAKI, R. (1971), "Estimation of vocal tract area function from poles of this transfer function", *Annual Report of the Engineering Research Institute, University of Tokyo, 30*, 81-88.

[14] GOPINATH, B., SONDHI, M.M. (1970), "Determination of the shape of the human vocal tract from acoustical measurements", *B.S.T.J., 49*, 1195-1214.

[15] GENIN, J. (1977), "Extraction de paramètres du conduit vocal à l'aide d'une méthode d'optimisation appliquée à un modèle mathématique", *Proc. Articulatory Modeling Symposium (GRENOBLE)*, 41-64.

[16] HEINZ, J. (1967), "Perturbation functions for the determination of vocal tract area functions from vocal-tract eigenvalues", *STL Q.P.S.R., 1*, 1-14.

[17] JOHANSSON, C., SUNDBERG, J., WILBRAND, H., YTTERBERGH, C. (1987), "From sagittal distance to area. A study of transverse, cross-sectional area in the pharynx by means of computed tomography", *Phonetica, 2*.

[18] KADOKAWA, Y., SUZUKI, J. (1969), "A simple calculation method for the vocal-tract configuration from the first three formant frequencies", *Electronics and Communications in Japan, Vol. 52-C, 9*, 123-120.

[19] KIRLIN, R.L. (1978), "A Posteriori Estimation of vocal tract length", *I.E.E.E. Trans. ASSP, Vol. 26, 6*,571-574.

[20] LADEFOGED, P., HARSHMAN, R., GOLDSTEIN, L., RICE, L. (1978), "Generating vocal tract shapes from formant frequencies", *J.A.S.A., Vol. 64, 4*, 1027-1035.

[21] LARAR, J.N., SHROETER, J., SONDHI, M.M. (1988), "Vector quantization of the articulatory space", I.E.E.E. *Trans. ASSP, Vol. 36, 12*, 1812-1818.

[22] LONCHAMP, F., ZERLING, J.P., LEFEVRE, J.P. (1983), "Estimating vocal tract area functions : A progress report", *Proc. Xth ICPhs (UTRECHT)*, 277-283.

[23] MERMELSTEIN, P. (1967), "Determination of the vocal-tract shape from measured formant frequencies", *J.A.S.A., Vol. 41, 5*, 1283-1294.

[24] MILENKOVIC, P. (1987), "Acoustic tube reconstruction from noncausal excitation", *I.E.E.E. Trans. ASSP, Vol. 35, 8*, 1089-1100.

[25] MILENKOVIC, P. (1984), "Vocal tract area funtions from two-point acoustic measurements with formant frequency constraints", *I.E.E.E. Trans. ASSP, Vol. 32, 4*, 1122-1135.

[26] MINIFIE, F.D., KELSEY, C.A., ZAGZEBSKI, J.A. (1971), "Ultrasonic scans of the dorsal surface of the tongue", *J.A.S.A., Vol. 49, 6*, 1857-1860.

[27] NAKAJIMA, T., OMURA, H., TANAKA, K., ISHIZAKI, S. (1974), "Estimation of vocal tract area functions by adaptative inverse filtering methods and identification of articulatory model", *Proc. Speech Communication Seminar (STOCKHOLM)*, 11-20.

[28] PAIGE, A. (1970), "Calculation of vocal tract length", *I.E.E.E. Trans. AU,*

Vol. 18, 3, 268-270.
[29] PAIGE, A. (1970), "Computation of vocal tract area functions", *I.E.E.E. Trans. AU, Vol. 18, 1*, 7-18.
[30] PERKELL, J.S., OKA, D.K. (1980), "Use an alternating magnetic field device to track midsagittal plane movements of multiple points inside the vocal tract", *99th Meeting of the A.S.A.*.
[31] PERRIER, P., BOË, L.J. (1987), "Passage de la coupe sagittale à la fonction d'aire", *16èmes J.E.P. (HAMMAMET)*, 128-131.
[32] SCHROEDER, M.R. (1967), "Determination of the geometry of the human vocal tract by acoustic measurements", *J.A.S.A., Vol. 41, 1*, 1002-1010.
[33] SCHROETER, J., LARAR, J.N., SONDHI, M.M. (1988), "Multi-frame approach for parameter estimation of a physiological model of speech production", *Proc. ICASSP*, 83-86.
[34] SCHROETER, J., SONDHI, M.M. (1989), "Dynamic Programming search of articulatory codebooks", *Proc. ICASSP*, 588-591.
[35] SONDHI, M.M., GODINATH, B. (1971), "Determination of vocal-tract shape from impulse response at the lips", *J.A.S.A., Vol. 49, 6*, 1867-1873.
[36] SONDHI, M.M. (1977), "Estimation of vocal-tract areas : the need for acoustical measurements", *Proc. Articulatory modeling symposium (GRENOBLE)*, 77-87.
[37] SONDHI, M.M. (1979), "Estimation of vocal-tract areas : the need for acoustical measurements", *I.E.E.E. Trans. ASSP, Vol. 27, 3*, 268-273.
[38] SONDHI, M.M. (1979), "Two acoustical inverse problems in speech and hearing", *Proc. Conf. Math. Methods Appl. Scattering Theory (WASHINGTION DC)*, 290-300.
[39] SONDHI, M.M., RESNICK, J.R. (1983), "The inverse problem for the vocal tract : numerical methods, acoustical experiment and speech synthesis", *J.A.S.A., Vol. 73, 3*, 985-1002.
[40] SOQUET, A., SAERENS, M., JOSPA, P. (1990), "Acoustic-articulatory inversion based on a neural controller of a vocal-tract model", *Proc. of the ESCA Workshop on Speech Synthesis (AUTRANS)*, 71-74.
[41] THOMPSON, M.A., ROBL, P.E. (1982), "X-ray microbeams for speech research", *Nuclear Instruments and Methods, 193*, 257-259.
[42] TOUSIGNANT, B., LEFEVRE, J.P., LECOURS, M. (1979), "Speech synthesis from vocal tract area function acoustical measurements", *Proc. ICASSP*, 921-924.
[43] VEMULA, N.R., ELLIOTT, D.L., ENGEBRETSON, A.M. (1982), "Estimation of vocal tract shape from input/output measurements", *Proc. ICASSP*, 927-931.
[44] WAKITA, H. (1983), "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveform", *I.E.E.E. Trans. AU, Vol. 21, 5*, 417-427.
[45] WAKITA, H., GRAY, A.H. (1975), "Numerical estimation of the lip impedance and vocal tract area functions", *I.E.E.E. Trans. ASSP, Vol. 23, 6*, 574-580.
[46] WAKITA, H. (1977), "Some considerations for the determination of area functions from acoustic date", *Proc. Articulatory Modeling Symposium (GRENOBLE)*, 89-96.

ESTIMATION OF VOCAL TRACT SHAPE USING NEURAL NETWORKS

K.Shirai, T.Kobayashi and M.Yagyu*

Waseda University, Tokyo, JAPAN
* Hosei University, Tokyo, JAPAN

ABSTRACT

This paper discusses an application of neural networks (NN) to the problem of estimating a vocal tract shape from speech waves.

The experimental results show that the difference in estimated articulatory parameter values between the conventional method (MM) and NN is only 3 % of value range on average. For a few data, big differences arise between MM and NN, but this is due to mis-estimation in MM rather than NN. The percentage of mis-estimates in NN is less than 70 % of MM. By introducing recurrent nodes, the value is reduced to be 50 %. In this case, the spectral error is improved by 5 % .

1. INTRODUCTION

Coarticulation compensation and speaker adaptation, which are two major difficulties in speech recognition, might be considered most clearly and fundamentally from the view point of the speech production mechanism. On the basis of this idea, the model matching (MM) method was proposed to extract articulatory parameters, which represent a vocal tract shape, from speech waves [1] and the estimated parameters were used for speech recognition [2,3,4]. These parameters are effective for the above two problems but the conventional estimation method has some problems: one is calculation cost and the other is instability of estimation. Since this method is constructed on the basis of hill climbing methods, it requires many iterations to converge and sometimes finds only local minimum. For a speech recognition system, a faster and more stable estimation method is desired. To prove these problems, we are trying to apply neural networks. In the previous papers, we tried to utilize a simple four-layer feed-forward neural network [6,7]. In this paper, we apply a recurrent neural network to the problem.

2. ARTICULATORY MODEL

The total configuration of the mod' and the characteristics of the articulator parameters are shown in Fig.1 and Tab 1.

The first five parameters (X_{T1}, X_{T2}, X_J, X_L, X_G) determine the shape of th oral and pharyngeal cavity, and the nasa ization parameter X_N describes the cros sectional area of the velopharyngeal part

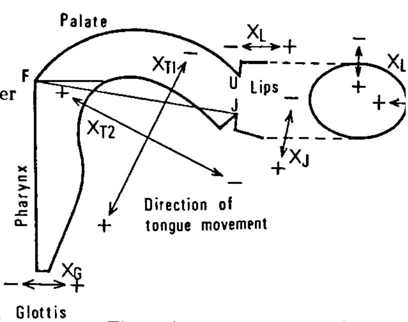


Fig.1 Articulatory model.

Table 1 Qualitative characteristics of the articulatory parameters.

parameter	organ	+	−
X_{T1}	tongue	back	front
X_{T2}	tongue	high	low
X_J	jaw	open	close
X_L	lip	rounded	unrounded
X_G	glottis	open	closed
X_N	velum	open	closed

3. MODEL MATCHING METHOD

The acoustic feature of speech waves (y) which are generated by the model (x) can be expressed by a nonlinear function of the articulatory parameters ($y = h(x)$). Therefore, the estimation problem is to solve a nonlinear function. The conventional technique for this problem is a nonlinear optimization method called the model matching method(MM). In this method, model parameters are iteratively changed to optimize a certain criterion function.

Let y_s be acoustic parameters measured from the speech wave after glottal and radiation characteristics are removed. Then, the estimate x of the articulatory parameters is obtained so as to minimize the following cost function.

$$J(x) = (y_s - h(x))^t P(y_s - h(x)) + x^t Q x + (x - x_0)^t R(x - x_0)$$

where P, Q, and R are the weight matrix, and x_0 is the estimate at the previous frame. This problem is solved by the following iterative form

$$x^{i+1} = x^i + \lambda\left(\frac{\partial h(x^i)}{\partial x^i}^t P \frac{\partial h(x^i)}{\partial x^i} + Q + R\right)^{-1} \left(\frac{\partial h(x^i)}{\partial x^i}^t P(y_s - h(x^i)) - Qx^i - R(x^i - x_0)\right)$$

4. ESTIMATION OF ARTICULATORY PARAMETERS USING NEURAL NETWORK

In our experiment, a four-layer feed-forward network and a four-layer recurrent network are adopted. Figure 2 shows the architecture of the recurrent network.

Weight coefficients are determined as follows: Firstly, vowels are selected from training dataset. Then, using MM, articulatory parameters are estimated for all frames in these data (including glides). 12th order LPC cepstral coefficients are also calculated for the same data and cepstrum - articulatory parameter pairs are prepared. Finally, applying backpropagation to these data pairs, the weights of the network are determined.

All vowel frames of phone balanced 216 tokens in the ATR word database(speaker ID: MAU) except data whose articulatory parameters are different from average estimate of corresponding vowel by more than 20 in mahalanobis distance measure are used for the training.

To estimate articulatory parameters from speech waves, cepstral coefficients are calculated by LPC analysis, and then, these data are input to the neural network.

5. EXPERIMENT

The evaluation test is performed using the vowel data in 5200 tokens in the ATR word database.

Table 2 shows the difference in estimated articulatory parameter values between MM and NN/RNN.

The difference is only 0.1 on average (3 % of value range). It can be seen that the neural networks work well to estimate articulatory parameters.

Figure 3 shows the estimated articulatory parameters and spectra. The solid line denotes the articulatory parameters obtained by MM and the dashed line denotes that by NN. Data is /niou/.

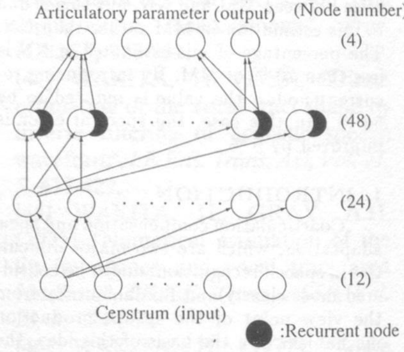


Fig.2 Structure of the neural network for the articulatory parameter estimation.

Table 2 Average and standard deviation of estimation error.

	Traning data average	Traning data std.dev.	Test data average	Test data std.dev
NN	0.08	0.12	0.11	0.13
RNN	0.08	0.12	0.13	0.14

(value range : [-1.5,1.5])

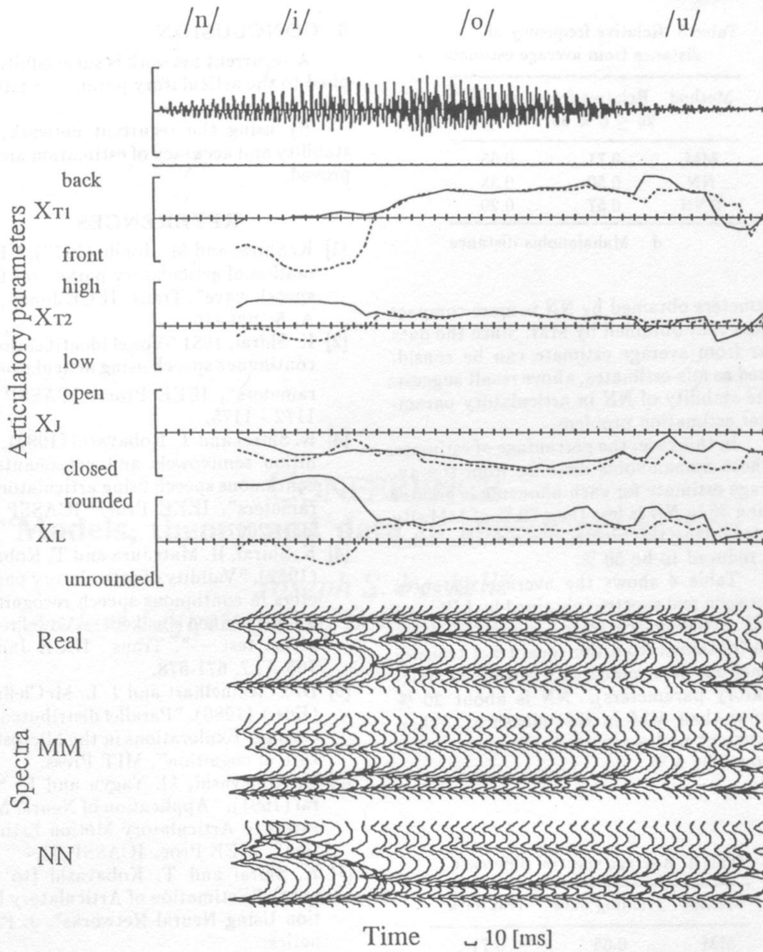


Fig.3 Movements of articulatory parameters obtained by the model matching method ( ) and the neural network ( ).

In the part of /ou/, articulatory parameters estimated with NN is almost equal to that estimated with MM. As for /i/, the big difference can be seen between MM and NN. The contour of the spectra (formant structure etc.) obtained by NN is more similar to the real one than that by MM. Parameters (X_{T1}, X_{T2}, X_J, X_L) should be (front, high, close, unrounded) in /i/ sound. The estimated articulatory parameters using NN satisfy this term but those by MM do not. Estimated articulatory parameters using NN can be regarded as more appropriate for /i/.

As we can see in this example, big differences arise between MM and NN for a few data. However, this result does not mean the mis-estimation in NN.

Table 3 shows the distribution of the estimated articulatory parameters. The values in the table denote the relative frequencies of the distances from the average estimate of each vowel.

The distribution of the articulatory pa-

Table 3 Relative frequency of distance from average estimate.

Method	Relative frequency [%] $25 \leq d < 36$	$36 \leq d$
MM	0.71	0.55
NN	0.50	0.38
RNN	0.57	0.29

d : Mahalanobis distance

rameters obtained by NN is more compact than that obtained by MM. Since the data far from average estimate can be considered as mis-estimates, above result suggests the stability of NN in articulatory parameter estimation problem.

In this case, the percentage of estimates whose mahalanobis distance from the average estimate for each phoneme is greater than 36 in NN is less than 70 % of MM. By introducing the recurrent nodes, the value is reduced to be 50 %.

Table 4 shows the average distances between real spectra (obtained by LPC analysis of the speech wave) and model spectra (obtained from the vocal tract transfer functions determined with estimated articulatory parameters). NN is about 30 % worse than MM in this measure. When recurrent nodes are used, the value is improved by 5 %.

Table 4 Average spectral difference.

Method	Training data	Test data
MM	0.65	0.71
NN	0.87	0.94
RNN	0.79	0.89

Thus, it is proved that the NN method is not suitable for the strict articulatory parameter estimation but has little risk of making serious errors. Moreover, this accuracy and stability are improved when recurrent nodes are introduced. This little risk is due to the characteristics of neural networks as associative memories. The strong constraints among articulatory parameters are embedded in the network structure on the training process. So the unnatural combination of parameters can be automatically excluded.

6. CONCLUSION

A recurrent network is successfully applied to the articulatory parameter estimation problem.

By using the recurrent network, the stability and accuracy of estimation are improved.

REFERENCES

[1] K. Shirai and M. Honda (1978), " Estimation of articulatory parameter from speech wave", Trans. IECE Japan, J61-A, 5, 409-416.

[2] K. Shirai, 1981 "Vowel identification in continuous speech using articulatory parameters", IEEE Proc. ICASSP 81, 1172 - 1175.

[3] K. Shirai and T. Kobayashi (1982), "Recognition semivowels and consonants in continuous speech using articulatory parameters", IEEE Proc. ICASSP 82, 2004-2007.

[4] K. Shirai, H. Matsuura and T. Kobayashi (1982), "Validity of articulatory parameters in continuous speech recognition for unspecified speakers — Vowel recognition test —", Trans. IECE Japan, J65-A, 7, 671-678.

[5] D. E. Rumelhart and J. L. McClelland (Eds.) (1986), "Parallel distributed processing: Explorations in the Microstructure of cognition", MIT Press.

[6] T. Kobayashi, M. Yagyu and K. Shirai (1991), "Application of Neural Networks to Articulatory Motion Estimation", IEEE Proc. ICASSP 91.

[7] K. Shirai and T. Kobayashi (to appear), "Estimation of Articulatory Mortion Using Neural Networks", J. Phonetics.

MODELS, THEORY AND DATA IN SPEECH PRODUCTION

Joseph S. Perkell

Massachusetts Institute of Technology,
Cambridge, Massachusetts, U.S.A.

ABSTRACT
This paper discusses modeling of the transformation from a linguistic-like input to a sound output in speech production. Such modeling can serve: 1) to formalize the expression of theoretical overviews and 2) as organizing frameworks for focussed programs of experimentation. As an example, a "task dynamic" production model is cited. The model incorporates underlying phonological primitives that consist of "abstract articulatory gestures", and it has been used in an initial attempt to interpret the relation between phonologically-based hypotheses and experimental data. Several issues that arise in such work are discussed, and suggestions are made for an alternative modeling approach.

1. INTRODUCTION
The fundamental motivation of production modeling is to account for the transformation that takes place from an underlying discrete linguistic representation through articulatory movements to a sound output. Global models attempt to account for most of or all of the transformation [cf. 6,30], and more detailed models attempt to account for specific parts of the transformation, such as sound production [cf. 37,8,32,15] or articulatory-to-acoustic relations (via area functions) [7,10,22]. Some other modeling work can be considered somehow to span these two categories [cf. 31,33,35,21].

In this paper I will focus on the contribution of global modeling to a basic understanding of the overall speech production process. In this kind of model, discrete linguistic representations of utterances serve as inputs to a controller which operates on a peripheral apparatus (in control theory terminology, a "plant") which produces sound. If global modeling is to inform us eventually about the nature of the actual input and control mechanisms for speech production, presumably it must incorporate an accurate model of the plant. Thus, in the long run, global models should include specific information about relevant aspects of production such as anatomy, biomechanics, aerodynamics, sound generation and articulatory-to-acoustic relations, all of which exert constraints on the role of the controller. One of the points of this paper will be that global production modeling should also consider *interactions between production and perception* (and lexical access), because perceptual mechanisms also have an influence on the control of speech production (and on sound patterns of languages).

At this point, however, not enough is known about the properties of peripheral production mechanisms and perceptual constraints to account for them comprehensively in a global production model. As a result, the peripheral components of such a model have to be represented mostly by abstractions that cannot be related directly to important constraints on speech production, and its input and controller cannot realistically represent hypotheses about the actual form of the underlying input and control mechanisms.

Given this situation, current global models have two potentially important contributions to make. One is as a means of forcing discipline on the formulation of theories of speech production. To the extent that such theory is incorporated into an implemented model, it has to be stated in precise terms. Some of the examples I will discuss below have already made this kind of contribution. Another equally-important contribution of a global model is as a link between theory and data, in effect, as an organizing framework for a focussed program of experimentation on strategies of speech production. In this arena, much less has been done up to now. The main reasons for this shortcoming are 1) the enormous amount of work involved, and 2) until recently, the lack of adequate tools, not only for efficient model development, but also for gathering and analyzing the right kinds of data in the most useful manner. Work along these lines is just beginning, and I will discuss an example to illustrate an important kind of contribution that global modeling may make in the near future. In the long run, we can hope that an iterative cycle of model development and related experimentation may inform us about the interesting but currently untestable principles that are incorporated into global models. In addition, as work advances on specific peripheral mechanisms, global models will become increasingly realistic and we will gain a much better understanding of relations among peripheral constraints, control strategies and sound patterns.

2. BACKGROUND
Twenty years ago at the VIIth International Congress of Phonetic Sciences, Bjorn Lindblom presented a paper entitled "Numerical Models in the Study of Speech Production and Speech Perception: Some Phonological Implications" [20]. In the following year, Ken Stevens published "The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data" [33]. Both of these seminal papers described the use of production modeling, including articulatory-to-acoustic relations, in combination with hypotheses about speech perception, to predict canonical articulatory targets. The targets were related to ideas about the phonological structure of language. Consistent with those ideas, the targets were essentially static, discrete and invariant in nature. Those particular production models and the related phonology had little to say about variability, timing and articulatory movement. The gap between such essentially static and discrete models on one hand and experimental observations of continuous and variable articulatory movement on the other led to suggestions that it might be more fruitful to study articulatory movements and basic physiological mechanisms without being constrained by the limitations inherent in "static" linguistic models [cf. 23].

Additional objections to such models [9] argued that timing in speech production should be a consequence of intrinsic properties of underlying units, instead of being specified extrinsically, as a separate component of the production process [cf. 25]. Such ideas about intrinsic timing have been a source of inspiration for work on a production model at Haskins Laboratories. The model is an influential component of a long-term, ongoing attempt to account for articulatory timing, kinematics and systematically-conditioned variation in speech production [cf. 29,30] in a way which is synergistic with an evolving phonological theory [3] and a theory of speech perception [19]. For this reason, and also because it is arguably the most well-developed effort of its kind, I will examine the Haskins Model (HM) as a way of illustrating some of the points raised above. I will also refer to other modeling efforts, but those references cannot do justice to the large amount of work that usually goes into production models.

My approach will be the following. I will briefly describe the HM and some initial experimental work which has been guided by the model. Then I will mention several issues raised by this work. Finally I will propose an alternative modeling approach that may have promise for the future.

3. PRODUCTION MODELING AND EXPERIMENTATION AT HASKINS LABORATORIES

3.1 The production model
Saltzman and Munhall [30] describe a "dynamical approach to gestural patterning in speech production" with which they "attempt to reconcile the linguistic hypothesis that speech involves an underlying sequencing of abstract,

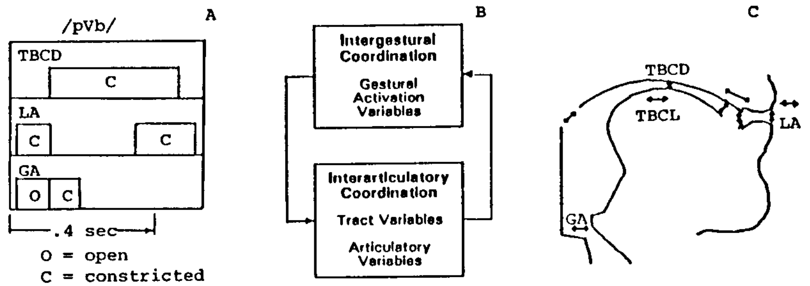


Figure 1. Components of the Haskins Laboratories production model. A: An input gestural score showing gestural activation intervals for the tract variables "tongue body constriction degree" (TBCD), "lip aperture" (LA) and "glottal aperture" (GA). B: The two-level dynamical model. C: The vocal-tract outline of the articulatory synthesizer. The arrows indicate tract variable coordinates; some are labeled as in part A, with the addition of "tongue body constriction location" (TBCL).

discrete, context-independent units, with the empirical observation of continuous, context-dependent interleaving of articulatory movements."

The fundamental, invariant unit in this approach is the abstract gesture.[1] Combinations of abstract gestures underlie phonetic segments, so in a rough sense, abstract gestures can be thought of as having a role similar to that of phonetic features. They characterize what can be done with the vocal tract, in combination, to produce speech sounds, but unlike traditional features, they are characterized by intrinsic dynamics.

Figure 1 is a schematic diagram of the major components of the HM. A "task dynamic" model (part B of Fig. 1 - [29]) is the controller for an articulatory synthesizer, which serves as the plant (part C - [27]) and produces an acoustic output. In the current stage of model development, an utterance-specific "gestural score" (part A - [3]) provides the input to the task-dynamic model in the form of a time-varying invocation of abstract gestures (each represented by a shaded rectangular pulse in one of the rows in part A of the figure). The gestural score is generated according to rules of Browman and Goldstein's Articulatory Phonology [2]. The functional sophistication and mathematical complexity of the task-dynamic model preclude a concise explanation that is also comprehensive, so the following description is necessarily oversimplified (see [12]).

In the task dynamic model (Fig. 1, Part B), there are two interacting levels. At the higher, "intergestural" level, abstract gesture combinations are specified from information in the gestural score, so they will appropriately influence vocal tract movements during the utterance. The lower, "interarticulator" level contains two sets of coordinates. The formation and release of linguistically-significant *vocal-tract constrictions*, such as lip aperture, tongue dorsum and blade constrictions (as well as constriction locations) are specified in a *tract variable coordinate system*. Articulatory movement is generated by modeling the influence of each discrete abstract gesture in the tract variable coordinate system as a time-invariant second order system (characterized as a *point attractor)*, with a characteristic stiffness, damping and equilibrium point. Thus, gestural activations determine the relative timing characteristics and evolving parameter values of a dynamical system expressed in terms of tract variable coordinates.

The tract-variable specification is transformed into motions of *model articulators* such as the lips, jaw and tongue body in the (midsagittal-plane) space of

1. Confusion can arise from use of the term "gesture" to denote an abstraction. To avoid such confusion, I will use the term "abstract gesture" to denote the abstraction, and "movement" to denote physical or simulated articulatory movement.

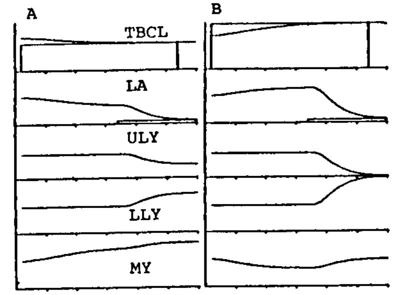


Figure 2. Example gestural activation levels and tract-variable and model-articulator trajectories for the utterances /Xib/ (A) and /Xaeb/ (B) (X = schwa).

the articulatory synthesizer, i.e., the *model articulator coordinate system.* The transformation (mapping) from tract variable to model articulator coordinates is strictly kinematic; it accounts in part for two facts: the motions of more than one articulator can be influenced by a single tract variable, and some articulators can be influenced by more than one tract variable [cf. 25].

All dynamical properties of this system reside in the controller, and biomechanical properties of the vocal tract are not represented directly. (The term "stiffness" refers to a characteristic of the abstract point attractor, and not to the actual biomechanical properties of the musculoskeletal apparatus.) In the formative stages of model development, the importance of considering biomechanics and aerodynamics is acknowledged, but bypassed for practical reasons.

In order to begin to understand some functional characteristics of the model, consider the very simple examples shown in Fig. 2. It illustrates gestural activation levels and tract-variable and model-articulator trajectories for the utterances /Xib/ (A) and /Xaeb/ (B) (X = schwa). In each panel, the row labeled TBCL contains an activation pulse specifying (target) level and duration along with a parameter value trajectory for the tract variable (abstract gesture) "tongue body constriction location"; the row labeled LA contains the same kind of information for the tract variable "lip aperture"; and rows labeled ULY, LLY and MY contain articulatory trajectories for the vertical upper lip, lower lip and mandible positions, respectively. The horizontal axis represents time. The TBCL and LA activation pulses overlap in time and have the same durations for both utterances. For the two utterances, the LA pulse has the same magnitude, indicating bilabial closure; but the different vowels (/i/ and /ae/) invoke different levels of TBCL activation. The TBCL trajectories evolve according to the activation levels and second order dynamic responses of their corresponding abstract gestures. The LA tract variable is defined with respect to the positions of the lips and jaw. Even though LA is not activated during the vowels, it has a changing trajectory which differs during the vowel portions of the two utterances because of the active influence of the vowels on the jaw (via the jaw-tongue synergy). The LA trajectories move toward closure when that abstract gesture is activated, but the rate and magnitude of movement differs, depending on the vowel-specific starting point at the onset of LA activation. For the two utterances, the pairs of ULY and LLY trajectories have the same overall shape, but different vowel-dependent rates of change and end points: since the vowel /ae/ is more open, greater lip movement is required to reach closure. The JY trajectories differ in response to the overlapping influences of the two different pairs of tract variable trajectories, since the mandible positioning is affected by both lip and tongue body positioning.

This example illustrates some (but not all) of several important characteristics of the task-dynamic model. The model accounts for: coarticulation (as "coproduction" of sequences of (partly) overlapping abstract gesture complexes), overlapping influences of multiple abstract gestures and tract variables on movements of individual articulators (as a result of "blending" of abstract gestures), and movement of articulators when they are not under active control (as governed by articulator-specific "neutral attractors" - see also, section 3.2 below).

Relative timing of the activation of the set of abstract gestures for each speech segment and the timing of sequencing of segments is currently specified extrinsically by the input gestural score, so the goal of accounting for intrinsic timing has not yet been reached. Future development of the model will incorporate additional layers of intrinsic dynamics, implemented in the form of neural networks [30,16] for inter-gestural timing within segments and timing among successive segments in an utterance.

3.2 Use of the Haskins Model to interpret experimental data

Browman and Goldstein [3] have proposed an articulatory phonology in which (abstract phonological) articulatory gestures are the "atoms out of which phonological structures are formed"; phonological structures are hierarchical "constellations" of gestures; and phonological regularities can be captured by representing constellations of gestures in gestural scores, which can be generated by rule. Some aspects of abstract gestural representations (i.e., those specifying different articulator sets) are "categorically distinct", that is, each set defines a separate phonological category. Other aspects (e.g., location and degree of constriction, stiffness) are not distinct in this way; they are hypothetically determined by relations among production, acoustics and perception as suggested by some of the above-mentioned modeling work [33,21]. It is claimed that information in the gestural score identifies particular lexical entries; phenomena such as non-canonical pronunciations in fluent contexts, segment deletions, insertions, assimilations, etc. can be characterized by orderly modifications of the gestural score. The abstract gestures of articulatory phonology are the same as those of the task-dynamic model, so control of the model with gestural scores can be used to test "phonologically-based" hypotheses.

Browman and Goldstein [4] have used the task-dynamic model to help interpret data from an experiment on the production of the vowel schwa, motivated by the observation that schwa assumes the quality of neighboring vowels. They wanted to investigate two alternative hypotheses: 1) schwa has a specific target which is coproduced with a neighboring stressed vowel, or 2) schwa is completely unspecified for tongue position. Movements of points on the lower lip, jaw, and the blade, mid and rear of the tongue dorsum were measured for one subject using the x-ray microbeam. Utterances were of the form /pV_1pX'pV_2pX/ (where X = schwa). Analysis of tongue point displacement data suggested that the V_1-V_2 trajectory was influenced by an independent schwa target, especially as evidenced by a decrease in tongue height during the schwa when V_1 and V_2 were both the vowel /i/.

The experiment was replicated in simulations with the task-dynamic model, using several different control strategies, observing the simulated articulations and performing listening tests of the acoustic output from the simulations. The control strategy that produced the most convincing result was one in which an active gesture for the medial schwa completely overlapped the gesture for V_2 and control regimes for V_1 and V_2 didn't overlap, as proposed previously [2]. The failures of alternative schemes were instructive, particularly one in which there was no active schwa gesture, but instead a gap with no vowel gesture specified between the end of V_1 and beginning of V_2. During that interval tongue motion was due to relaxation of the tongue body to its neutral position, as well as jaw motion, called into play by the bilabial gestures for /p/. The result was a decrease in tongue height during the schwa when V_1 was the same as V_2. The decrease agreed with the x-ray data for when both vowels were /i/, but disagreed when the vowels were /a/: with /a/, the simulated tongue height decreased during the schwa, but it increased (toward a neutral configuration) in the x-ray data.

It would be possible to offer alternative interpretations of the data; an apparently "successful" simulation cannot "prove" the hypotheses of Browman and Goldstein or "validate" the modeling approach. The main point of this example is rather that it illustrates how production modeling can serve as a means of focusing experimentation. I suggest that if such efforts with the HM can progress in a productive and appropriate fashion, a large, coherent body of experimental

data will result and we will have learned more about speech production than from an equivalent amount of less-well-guided research. However, trying to do this kind of modeling work raises a number of issues; some of those issues may be specific to the Haskins approach and some of them are inherent to any similar modeling effort.

4. MODELING ISSUES AND CHALLENGES

Before considering issues raised by the Haskins work, it should be noted again that their approach is unique in its comprehensiveness and the extent to which it has been implemented and tested.

4.1 The underlying theory: how valid are its basic assumptions?

As Browman and Goldstein progress, presumably they will be conducting more listening tests to investigate alternative control schemes in the production modeling. What will they be asking the listeners to do in these tests? I suggest that Browman and Goldstein will have to deal increasingly with the issue of what are appropriate acoustic and perceptual criteria, perhaps along the lines suggested by work on lexical access. It seems to me, largely on intuitive grounds, that not enough emphasis is given to acoustic and perceptual mechanisms in an approach which places such strong emphasis on gestures. As will be discussed further below, this issue is particularly important for features with prominent acoustic correlates that result from abrupt transitions at moments of vocal tract closure and release. Such features may not be efficiently characterized in terms of abstract gestures.

It remains to be determined how Articulatory Phonology does as a phonology. As one alternative, Halle and Stevens [11] discuss a system (derived from [5,28,24]) which is also hierarchical, but has as its primitive elements more traditionally-defined features. Those features play a role in speakers' "knowledge of the language", in that they capture phonological distinctions and transformations. The features are described as falling into two categories, articulator-bound (i.e. executed by particular articulators - "labial" by the lips, "coronal" by the tongue blade) and articulator-free ("continuant", "sonorant" and "syllabic" which have robust acoustic correlates and are not tied to the action of any one articulator). This division is somewhat similar to Browman and Goldstein's division of abstract gestural primitives into those that are and are not "categorically distinct". The hierarchy of primitives in both points of view is anatomically-based. However, in contrast to abstract gestural primitives, the more traditional features discussed by Halle and Stevens are based as strongly in acoustics and perception as they are in production. One of the challenges to Articulatory Phonology will be to determine how well it can account for phonological regularities and processes such as assimilation, in comparison to other systems. A major challenge to proponents of any system is to account for the relationship between underlying representations and kinematic articulatory behavior; by its nature, Articulatory Phonology claims to incorporate that relationship.

4.2 Toward more realistic models: consonant production, aerodynamics and biomechanics

For justifiable practical reasons, the HM does not yet try to account for aerodynamical and biomechanical properties of speech production. However, a modeling approach which is based on concentrating all dynamical behavior in the controller may be difficult to adapt in the future to incorporate biomechanics and aerodynamical factors, which are critical for the simulation of consonant (and voice) production. Accounting for consonant production in the task-dynamic model and in Articulatory Phonology are likely to present some of the most formidable challenges for this work.

Regardless of the choice of modeling approach, incorporating aerodynamics and especially biomechanics is very difficult, because so little is understood about those mechanisms. Scully's [31] synthesis work, which includes aerodynamic factors, implies a large increase in number of details that have to be specified and major problems with specifying parameters at moments of consonant closure and release. The aerodynamic and resulting acoustic events which occur at those critical moments may be especially important as cues for perception and lexical access, particularly in signaling temporal landmarks and as correlates of certain distinctive features [cf. 34]. When future models do incorporate biomechanical and aerodynamical constraints, the role of the controller in such models probably will be more representative of the type of control actually performed by the central nervous system. Presumably, that control will reflect: 1) the strategies speakers use in producing perceptually-important temporal landmarks, and 2) whether timing is extrinsic, intrinsic or due to a combination of factors which takes listeners' needs into account while conforming to internal constraints (also see Section 4.4).

4.3 Implications of incorporating intrinsic timing and introducing neural networks into production models

Saltzman and Munhall [30] observe that connectionist models can embody the knowledge constraining the performance of serial activity, including coarticulatory patterning. Jordan's recurrent network model (see [16]) can be used to define a time-invariant dynamical system with an intrinsic time scale that spans a sequence, and it has been used to simulate sequencing and coarticulation with a feature-like input. As mentioned above, future work with the HM will attempt to use connectionist models to account for inter-gestural temporal coordination within segments (for properties such as voicing onset time) and for timing of sequences of segments. This endeavor will also be extremely challenging, especially when it attempts to account for the complex timing relationships that have been observed in acoustic measurements as well as other perceptually-salient acoustic characteristics of the resulting signal. The future incorporation of intrinsic timing also raises questions about the ultimate role of currently-hypothesized representations of phonological regularity in the form of gestural scores.

4.4 Using models to evaluate data and vice versa

In the results of Browman and Goldstein [4], there are large differences between the "most successful" model output and the articulatory data. Some of these differences are unavoidable, because the model does not attempt to take into account a number of kinds of intra- and

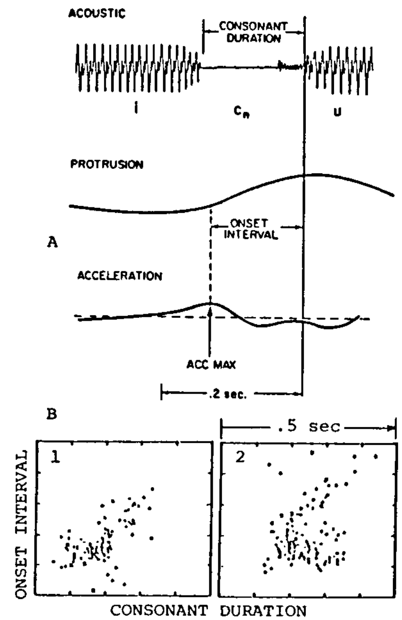


Figure 3. A: Schematic illustration of acoustic, lip protrusion and acceleration signals vs. time, illustrating two measurements, "consonant duration" and "onset interval". B: Plots of the illustrated measures for two subjects.

inter-speaker variation that are evident in actual production data. Given other important priorities, it is probably not appropriate to attempt to account for such variation in the modeling at this point. However, it will be challenging to interpret variable data with respect to a model that does not incorporate variability.

Scully's [31] synthesis of a short utterance as produced by two different speakers reveals a number of inter-speaker differences in articulatory, aerodynamic and acoustic parameters. Fig. 3 shows another kind of inter-speaker difference from an experiment we conducted on timing of upper lip protrusion movements for the vowel /u/ [26]. Part A is a schematic illustration of acoustic, lip protrusion and acceleration signals vs. time for an utterance /iC_nu/, illustrating two measurements, "consonant duration" and "onset interval". The experiment used utterances such as "lee coot" and "leaked coot" to vary consonant duration. Part B shows plots of onset interval vs. consonant duration for two subjects. For Subject 1, there was a predominantly linear relation of the two measures. For Subject 2, there was much more scatter in the data. A positive linear relation in the data can be interpreted as reflecting a constraint for the movement to begin around the time of the acoustic offset of the preceding /i/, as suggested by a look-ahead model [cf. 13]. The data containing more scatter can be interpreted as evidence for the fluctuating influence of the preceding constraint, in competition with a constraint for the lip protrusion movement to have constant kinematics, as hypothesized in the task-dynamic framework. Thus, different subjects can exhibit movement patterns that support rather different interpretations.

Such examples point out additional major challenges to production modeling: developing appropriate and objective performance metrics and model optimization procedures, as well as efficient methods for gathering, analyzing and interpreting the most useful kinds of production data. If global models are to be used effectively to evaluate experimental data in the long run, there has to be a dramatic increase in the sophistication and amount of work that compares model output with production data.

5. POSSIBLE FUTURE DIRECTIONS

5.1 Parallel-distributed-processing models (neural networks)

Jordan's work with neural networks appears to be particularly promising in its capability to account for some of the most important characteristics of speech production, while at the same time offering intuitively sound approaches to some of the difficult problems mentioned above (see [16]; also [1]). This work, consisting of two lines of research, seeks to integrate solutions to the problems of "excess degrees of freedom", serial order and learning.

As already mentioned, one component of this modeling has been used to simulate *sequencing* and *context sensitivity* (i.e., coarticulation) of speech movements with a feature-like input. This simulation is accomplished by representing actions as points in a target space which correspond to regions in an articulatory space. A trajectory is found which passes through the regions in articulatory space so that values of articulatory degrees of freedom change minimally over time. This constraint represents an interaction between the serial nature of the task and the existence of excess degrees of freedom.

The problem of using acoustic information to constrain articulatory trajectories is addressed through the mechanism of a *forward model*, which represents the second component of Jordan's work. A forward model is a learned internal model of the transformation from articulatory space to (acoustic) target space. Once the forward model has been learned, it can be used to convert acoustic errors backward into articulatory errors. Thus the system can learn to perform articulatory trajectories on the basis of iterative attempts to achieve specified sequences of acoustic targets [17].

5.2 Toward an alternative global model

The preceding material leads me to suggest that there is a need for an alternative, comprehensive production model. Such a model and the primitive elements it uses for specifying utterances should, in a balanced fashion, take into account as many aspects of the speech communication process as we think are important for a further understanding of speech production. Those aspects include: the nature of phonological regularities, control of the production mechanism, peripheral constraints on articulation and sound generation, articulatory-to-acoustic relations, relations between acoustics and perception, and mechanisms of lexical access. Clearly, such an effort has to have a long range perspective, but I believe enough of the elements exist to begin approaching the problem.

From my point of view, the most rational primitive elements are features which can be convincingly motivated in phonology and, in general, have correlates in production, perception and acoustics. As hypothesized by Stevens [34], collections of such feature specifications could serve as lexical representations. In those representations, manner features, which have robust acoustic correlates (but are articulator free), specify temporal landmarks. Each landmark is generated by the action of a primary articulator. At the time specified by each of the landmarks, one or more secondary articulations must be coordinated to produce cues corresponding to additional feature specifications (also see [14]). Such representations could serve as the input goals for a controller, as developed by Jordan.

Optimally, the controller would operate on a realistic model of the peripheral production mechanism, which would have correct anatomical, biomechanical and aerodynamic properties. Since we currently cannot build such a model, an articulatory model like the one developed by Maeda [22] could be used in articulatory synthesis. Maeda's model has the advantage of being based on statistical analyses of the articulations of individual speakers. In order to circumvent problems of inter-speaker variation and obtain more realistic synthetic area functions, it would be helpful to obtain enough articulatory data on a few speakers to specify individual versions of such a model and then use the same speakers in subsequent experimental and model-based tests of hypothesized control mechanisms.

Until we know much more about the anatomical, biomechanical and aerodynamical detail necessary for the realistic synthesis of consonants, it may be worth considering an alternative to articulatory modeling in the generation of synthetic utterances for use in perceptual tests: the time-varying articulatory positions generated at the controller output could be used to specify the parameters of a high-quality terminal analog synthesizer [36,18].

If such a modeling effort could be realized, it would be possible to "close the loop" as is being done at Haskins Laboratories, and examine the perceptual consequences of hypothesized production mechanisms in comparison with actual production data. This approach would face many of problems outlined above, but it would be based on a balanced perspective which may be more representative of the speech communication process as a whole.

6. CONCLUSION

Before global modeling of speech production can provide us with real insight about the control of speech production, it will have to come to grips with a number of extremely difficult problems as mentioned above. Undoubtedly, the solutions to many of those problems will receive major contributions from modeling work on a variety of detailed mechanisms. Those mechanisms range from interactions among aerodynamics and biomechanics in the production of transient acoustic cues, to signal processing and feature extraction in the auditory system. In the meantime, work with global models should continue to stimulate our thinking about theoretical issues and serve as organizing frameworks for focussed programs of experimentation.

ACKNOWLEDGEMENTS

I am grateful to Cathe Browman, Nick Clements, Michel Jackson, Michael Jordan, Elliot Saltzman, Ken Stevens and Mario Svirsky for their helpful comments. Caroline Smith, Cathe Browman, Louis Goldstein and Elliot Saltzman kindly generated the plots used in Fig. 2. Preparation of this manuscript was supported by N.I.H. Grant No. CD00075.

REFERENCES

[1] BAILLY, G. (1990). Robotics in speech production: Motor control theory, Proceedings of the Tutorial Day on Speech Synthesis, Autrans, France.

[2] BROWMAN, C.P. & GOLDSTEIN, L. (1986). "Towards an articulatory phonology", *Phonology Handbook, 3,* 219-252.

[3] BROWMAN, C.P. & GOLDSTEIN, L. (1989a). "Articulatory gestures as phonological units", *Phonology, 6,* 201-251.

[4] BROWMAN, C.P. & GOLDSTEIN, L. (1989b). "'Targetless' schwa: an articulatory analysis", presented at the Second Conference on Laboratory Phonology, Edinburgh.

[5] CLEMENTS, G.N. (1985). "The geometry of phonological features", *Phonology Yearbook, 2,* 223-250.

[6] COKER, C.H. (1976). "A model of articulatory dynamics and control", *Proc. IEEE, 64,* 452-460.

[7] FANT, G. (1980). "The relationships between area functions and the acoustic signal", *Phonetica, 37,* 55-86.

[8] FANT, G., LILJENCRANTS, J. & LIN, Q. (1985). "A four-parameter model of glottal flow", STL-QPSR, 4/1985, Stockholm, 1-13.

[9] FOWLER, C.A. (1980). "Coarticulation and theories of extrinsic timing control", *J. Phonetics, 8,* 113-133.

[10] FUJIMURA, O. & KAKITA, Y. (1979). "Remarks on quantitative description of lingual articulation", in B. Lindblom & S. Ohman (eds.), *Frontiers of Speech Communication Research,* Academic Press, London.

[11] HALLE, M. & STEVENS, K.N. (1990). "Knowledge of language and the sounds of speech", presented at the Symposium on Music, Language, Speech and Brain, Stockholm.

[12] HAWKINS, S. (in press). "An introduction to task dynamics", in D.R. Ladd & G.J. Docherty (eds.), Proceedings of the Second Conference on Laboratory Phonology, Cambridge University Press.

[13] HENKE, W.L. (1967). "Preliminaries to speech synthesis based on an articulatory model", Proceedings of the 1967 IEEE Boston Speech Conference, 170-177.

[14] HUFFMAN, M. (manuscript). "Articulatory landmarks: Constraining timing in phonetic implementation".

[15] ISHIZAKA, K. & FLANAGAN, J.L. (1972). "Synthesis of voiced sounds from a two-mass model of the vocal cords", *Bell System Tech. J., 51,* 1233-1268.

[16] JORDAN, M.I. & ROSENBAUM, D.A. (1989). "Action", in D.A. Posner (ed.), *Foundations of Cognitive Science,* M.I.T. Press, Cambridge, MA, 727-767.

[17] JORDAN, M.I. & RUMMELHART, D.E. (1990). "Forward models: Supervised learning with a distal teacher", *Cognitive Science.*

[18] KLATT, D.H. (1980). "Software for a cascade/parallel formant synthesizer", *J. Acoust Soc. Am., 67,* 971-995.

[19] LIBERMAN, A.M. & MATTINGLY, I.G. (1985). "The motor theory of speech perception revised", *Cognition, 21,* 1-36.

[20] LINDBLOM, B.E.F. (1971). "Numerical models in the study of speech production and speech perception: Some phonological implications", Proceedings of the VII International Congress of Phonetic Sciences, 71-73.

[21] LINDBLOM, B.E.F. & SUNDBERG, J.E.F. (1971). "Acoustical consequences of lip, tongue, jaw and larynx movements", *J. Acoust. Soc. Am., 50,* 166-1179

[22] MAEDA, S. (1990). "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", in W.J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling,* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[23] MACNEILAGE, P. (1972). "Speech physiology", in J.H. Gilbert (ed.), *Speech and Cortical Functioning,* Academic Press, New York.

[24] MCCARTHY, J.J. (1988). "Feature geometry and dependency: A review", *Phonetica, 45,* 84-108.

[25] PERKELL, J.S. (1980). "Phonetic features and the physiology of speech production", in B. Butterworth (ed.), *Language Production,* Academic Press, London.

[26] PERKELL, J.S. & MATTHIES, M.L. (manuscript). "Temporal measures of labial coarticulation for the vowel /u/".

[27] RUBIN, P.E., BAER, T. & MERMELSTEIN, P. (1981). "An articulatory synthesizer for perceptual research", *J. Acoust Soc. Am., 70,* 321-328.

[28] SAGEY, E. (1986). "The representation of features in nonlinear phonology". Ph.D. dissertation, Massachusetts Institute of Technology.

[29] SALTZMAN, E.L. (1986). "Task dynamic coordination of the speech articulators: A preliminary model", *Experimental Brain Res., Ser 15,* 129-144.

[30] SALTZMAN, E.L. & MUNHALL, K.G. (1989). "A dynamical approach to gestural patterning in speech production", *Ecological Psychology, 1,* 333-382.

[31] SCULLY, C. (1990). "Articulatory synthesis", in W.J. Hardcastle & A. Marchal (eds.), *Speech Production and Speech Modeling,* Kluwer Academic Publishers, Dordrecht, The Netherlands.

[32] SHADLE, C.H. (1986). "Models of turbulent noise sources in the vocal tract", *Proc. Inst. of Acoustics, 18,* 213-220.

[33] STEVENS, K.N. (1972). "On the quantal nature of speech: evidence from articulatory-acoustic data", in P.B. Denes & E.E. David (eds.), *Human Communication, a Unified View,* McGraw-Hill.

[34] STEVENS, K.N. (1988). "Phonetic features and lexical access", presented at the Second Symposium on Advanced Man-Machine Interface Through Spoken Language, Hawaii.

[35] STEVENS, K.N. (1989). "On the quantal nature of speech", *J. Phonetics, 17,* 3-45.

[36] STEVENS, K.N. & BICKLEY, C. (in press). "Constraints among parameters simplify control of Klatt formant synthesizer", *J. Phonetics.*

[37] TITZE, I.R. (1984). "Parameterization of the glottal area, glottal flow and vocal fold contact area", *J. Acoust Soc. Am., 75,* 570-580.

THE REPRESENTATION IN MODELS OF WHAT SPEAKERS KNOW

Celia Scully

University of Leeds
Leeds LS2 9JT, U.K.

ABSTRACT

The need to include in models of speech production all the physical processes and the interactions between them is discussed. The role of trial and error with auditory monitoring in learning to achieve the goals of speech production is emphasised, for models as well as for real speakers.

1. INTRODUCTION

Perkell's paper focusses on global models which map from discrete linguistic units through to a synthetic acoustic signal. The need to take the mapping further still, to speech perception and lexical access, is emphasised. This is an extremely important aspect of production modelling. Separation of control and "plant" is part of the philosophy of some global models. It may be interesting to consider where the line should be drawn between these two components. Finally, I should like to consider the nature of the problems of including biomechanical systems in a composite model and consider some priorities in the tasks confronting us.

2. MAPPING ALL THE WAY THROUGH TO PERCEPTION

Real speech production is goal-directed sensori-motor behaviour. The goal is a sequence of auditory patterns, broadly enough defined to accommodate variations such a cross-speaker differences; there are many different combinations of acoustic patterns that are acceptable as a specified goal. The goal is a different one, depending upon the context and style of the speech. Speakers learn to make auditory patterns during speech acquisition, and skilled (normal adult) speakers have already learned and stored the appropriate neural patterns for force and time for achieving each auditory goal.

The actions are the means to the end, not the goal itself. Extensive learning can be expected to give the speaker, like the musician or the games player, a huge repertoire of patterns of activity. The linguistic structures posited as inputs to a global model of speech production can be considered as a (partial) description of the auditory goal itself. Acquisition of speech must surely include, and indeed rely heavily upon, trial and error with monitoring, by auditory combined with other kinds of sensory feedback.

It must be acknowledged that the neural processes which permit a speaker the flexibility to speak in many different styles are mysterious. Perhaps it is premature to try to

address that question at the present time? After all, the organisation of motor activity in the Central Nervous System (CNS) is more difficult to study than perception; and speech production must surely be one of the most complex of all motor skills.

Data and modelling for the transform from acoustics to perception constitute an enormously important research area in its own right. I do not believe that it is central to the immediate tasks confronting us in speech production research, however. The perfect mapping device exists close at hand - that is, ordinary speakers of the language concerned. They can tell the modeller whether the auditory goals have been achieved or not. The fact that we do not understand this mapping need not worry us unduly. It is not even essential to acoustically segment the synthetic speech signal or to relate particular portions of the soundwave to individual linguistic elements as part of this assessment procedure. There is, in any case, no real justification for labelling a particular acoustic event, such as the onset of voicing, as a boundary between two linguistic units such as a consonant and a vowel.

Formal perception tests on every speech or speech-like sequence generated by a model or analysed from real speech would be an excessively demanding counsel of perfection. But it is perfectly feasible for one or more native speaker-listeners to check all recordings auditorily, and to confirm that the output has indeed achieved the auditory goal set.

The role of auditory feedback in trial and error improvement of performance of real speech cannot be over-stated. Speakers with a significant hearing loss cannot be expected to learn the appropriate actions. The matter for comment here is not that they get so many aspects of the speech wrong, but that any at all are right. It is no coincidence that the early attempts at speech with a composite model of speech production can be very abnormal and indeed resemble deaf speech. So a global or other model which generates sounds should not be judged a failure after its first few attempts. It must learn through experience (and theory too, where that is available) the right combinations of gestures needed. It is a vehicle for the actions which achieve the goals set, not a model of the goals themselves.

3. SEPARATION OF CONTROL AND PLANT

The name "model" is given to two rather different kinds of endeavour. A model may be a simulation of some physical processes or it may mean a hypothesised form of organisation and control, as are the conflicting comb and chain models which state how actions relate to phoneme-type linguistic units. It seems clear that both these last two are over-simplified views of the CNS. But, even assuming more plausible models of control by the CNS, I am not sure that we are able to say at present that some aspects of speech production are the control, while other aspects are the controlled system, the plant.

The signals which activate muscles originate in the CNS where the inputs to the speech producing system also reside, but perhaps the former of these neural processes could be considered to be one of the stages of speech production which together constitute the plant, simulated by the composite model. Conceptually, the following processes may be separated and listed in order: neural, muscular, articulatory, aerodynamic, acoustic sources, acoustic filtering, and radiation of a soundwave. But we know that these separate stages interact with each other. From the proposition that skilled speakers have already learned the neural patterns needed to achieve a specific auditory goal, it follows that speakers have knowledge about all these stages and their interactions, as applied to their own speech production at least.

Many speakers are not consciously aware that their oral pressure rises a little, but only to one or two cm H_2O above atmospheric pressure, during the production of nasal and approximant consonants. But the CNS of almost any speaker knows that certain force-time combinations resulting in particular coordinations for the vocal tract closure and the velum lowering and raising are unsuccessful for the production of nasal consonants. One reason for this is that oral air pressure rises too high, and an intrusive plosive is perceived. The auditory goal specifies the right sound patterns in the right order, with no extra sounds and none omitted. So aerodynamic effects are not simply the consequences of linguistically-determined neural commands - they influence the form of the neural patterns in the first place.

This example is intended to illustrate the proposition that the neuromuscular, mechanical, aerodynamic and acoustic stages all combine to determine the form of the neural signals which cause the actions. The properties, mechanisms and constraints of all these stages are crucially important both in setting limits for the force-time plan and in generating the rich, complex details of structure and multiple acoustic cues for perception in the speech signal.

4. TIMING AND ARTICULATORY EVENTS

Much research is directed towards the timing patterns of speech. I agree very much with the view that durations of acoustic segments are not simply imposed on linguistic structures. It has been argued for many years, and the principle has been made explicit in models, that one of the strongest constraints in speech production is the time required for an individual solid structure such as the tongue or the vocal folds to be accelerated and then decelerated and so perform an articulatory transition, where that gesture is essential for the achievement of the auditory goal. This seems to act as a very important factor, interacting with the control of auditory length and prominence for vowels and for some consonants in determining speech segment durations (see, for example,[3], [8], [9], [10]). As Perkell points out, timing of actions, or of the application of forces to achieve movement, must still be specified.

Inter-articulator coordination is of the essence in speech production. Elsewhere I have suggested that some time intervals between articulatory events might perhaps be preserved across a change of speaking rate, while some actions might be dispensed with [9]. The Haskins modelling is associated with the suggestion that inter-articulator timing may be expressed as a constant phase within postulated cycles of movement. There does not seem to be any obvious basic

principle, comparable to a criterion for speech production such as minimisation of work done, to favour one view or the other. Here, a model which attempts to sketch all the physical processes which constitute the mapping, either from muscle forces and their timing or from articulation, to the acoustic signal, can be neutral. It can, as Perkell says, serve as a means of focussing experiments.

5. THE BIOMECHANICS OF SPEECH PRODUCTION: COMPONENTS OF MODELS

We cannot wait for the time when the properties of all the systems are thoroughly understood; we need to incorporate them now in composite or global models of speech production. This means that we need to try to capture, qualitatively at least, some of the observed behaviour. This is, of course, a very hard task, but it must be undertaken. It may be better tactics not to focus too much effort on the input and control mechanisms, but instead take on the more modest but still very ambitious task of trying to describe, and, where possible, explain the behaviour of natural speech in all its aspects, from neuromuscular processes right through to the radiation of soundwaves. Like scientists in other disciplines, we should be content to advance our understanding little by little, piecemeal.

One severe limitation on progress is the lack of widespread availability of the advanced techniques which have been developed now for studying natural speech production. In view of the difficulties of the task, the small numbers of researchers and the limited funds available, perhaps it would be in our interests to pool our skills and resources by organising ourselves into quite large collaborative groups, with travelling speakers.

An individual researcher or group should pose the questions which interest them; we cannot and need not all aim at a global model. But it will always be important to keep sight of the implications for subsequent processes, and, especially, for the output acoustic and auditory patterns.

There is a need for a phenomenological approach to modelling at present, but better true physical models need to be developed also. One source of frustration in modelling is that the basic mathematical and physical theory for many of the processes have been so little developed.Take, for example, the conditions controlling the presence or absence and the spectral properties of turbulence noise sources. In a well established science recognised as having practical importance, surely the pioneering work of Stevens [12] would have been followed up by armies of researchers? Work on the problems of turbulence noise in jets has, I am sure, received plenty of attention in the intervening years. There is more cause for optimism now, as regards this particular example [11],[2].

6. THE INDIVIDUAL PROCESSES: INTERACTIONS AND SOME QUESTIONS

6.1. Neural Signals and Muscle Length Changes

The neural signals to the muscles interact with the muscle length changes. The timing patterns chosen can exploit these interactions to maximise force output, by ensuring that each muscle of a reciprocally acting pair is stretched prior to its innervation, a principle, which seems to be applied by fish to their swimming muscles [1] and in speech also [4]. A model of the innervation of a reciprocally acting pair of muscles, as for fast head movements [7] could perhaps generate matches to and suggest explanations for the coordination of electromyographic traces seen in speech.

6.2. Articulation

Assuming that the notion of an articulatory event is useful, the choice of particular moments as candidates is by no means self evident. Gestures, especially the middle, high velocity portions of transitions, may well be more important than the end point reached, as the thing controlled.

The positive correlation between maximum velocity near the middle of the transition and the distance transversed seems to be a strong constraint. In other non-speech tasks, subjects could not easily be made to bypass this relationship [6]. Some speech data show duration, as well as peak velocity, correlated with distance tranversed (Keller, [5] 343-364). These findings suggest the possibility that transitions large and small, fast and slow, may have kinematic similarity, with the dimensionless number $\mathbf{v^2/al}$ kept constant, where **v**, **a**, and **l** are characteristic velocities, accelerations and lengths respectively. Distance, velocity and duration might perhaps be considered as dependent variables, with muscle force as the controlling variable.

6.3. Aerodynamics

It is not really so difficult to include aerodynamic processes in models of speech production, although present representations of the processes are highly simplified. The use of packages such as the NAG (Numerical Algorithms Group) routines for numerical solution of simultaneous differential equations is to be recommended. Should the respiratory control be considered as a nett expiratory force (Ohala, [5] 23-53)? In cardiovascular studies the question has been posed as to whether the heart is a flow source or a pressure source (Pedley, personal communication). This question needs to be considered further for respiratory control in speech.

Some quite basic parameters, such as the volume enclosed in the vocal tract cavity, need to be measured, or at least estimated. Magnetic resonance and ultrasonic imaging offer hope for this as well as for the difficult task of improving the mapping from mid-sagittal views of the vocal tract onto area functions.

6.4. Acoustic sources

Individual articulatory events do not lead directly to sound patterns. It is important to consider the total effect of the movements of *all* the structures involved, including larynx and subglottal respiratory actions as well as those which shape the supraglottal vocal tract. Furthermore, articulatory geometry and aerodynamics interact to generate acoustic sources. Thus, for example, the moment of the onset of voicing for a vowel following a consonant is not a direct reflection of any one articulatory event; it depends on the interaction of at least three factors: the air pressure drop across the glottis, the state of adduction of the vocal folds, and the stiffness and effective mass of the vocal folds. So the preceding consonant will influence the moment of voice onset, for example if it requires the vocal folds to be abducted.

The need to specify articulatory parameter values at the moments of closure and release for consonants (Scully, [5] 151-186) should not be seen as a major problem. This is precisely the interest attached to the modelling. If the model generates acoustic sources in a way which approximates, however roughly, the processes of real speech, it can help to demonstrate how and why particular combinations of actions and coordination are chosen by speakers. By perturbing the timing and other aspects of the simulated articulation, modelling can investigate the limits within which the auditory goal is attained, and the covariations in multiple acoustic pattern features associated with the variability found in natural speech.

7. CONCLUSIONS

It is true that there are severe difficulties in obtaining real speech data, especially for the larynx, and that too much use has to be made of analysis-by-synthesis at present; but that is not a reason for avoiding the problems of modelling all the biomechanics of speech production. Apart from improved theory and more data, there is a real need to ease the burden for the modeller, for example by the development of code books and the use of graphical, file handling and mathematical techniques, so as to reduce the long auditory feedback loop, which puts the experimenter into almost the same situation as a hearing-impaired speaker.

8. ACKNOWLEDGEMENT

I am grateful to Esther Grabe-Georges for her help in the preparation of the manuscript.

9. REFERENCES

[1] ALTRINGHAM, J.D. & JOHNSTON, I.A. (1990) "Modelling muscle power output in a swimming fish", *J.Exp.Biol.,148*, 395-402.

[2] BADIN, P. (1989) "Acoustics of voiceless fricatives: production theory and data", *STL-QPSR 3/1989*, 33-55.

[3] COKER, C.H. (1976) "A model of articulatory dynamics and control", *Proc.IEEE, 64*, 452-460.

[4] GRACCO, V.L. (1988) "Timing factors in the coordination of speech movement", *J.of Neuro-science, 8*, 4628-4639.

[5] HARDCASTLE, W.J. & MARCHAL, A., Eds., (1990) *"Speech Production and Speech Modelling"*, Kluwer, Dordrecht.

[6] MILNER, T.E. (1986) "Controlling velocity in rapid movements", *J.of Motor Behaviour, 18*, 147-161.

[7] RAMOS, C.F. & STARK, L.W. (1987) "Simulation studies of descending and reflex control of fast movements", *J.of Motor Behaviour, 19*, 38-61.

[8] SCULLY, C. (1976) "A synthesizer study of aerodynamic factors in speech segment durations", in FANT, G., Ed. *"Progress in Speech Communication" Vol.2*, 227-234, Wiley, Stockholm.

[9] SCULLY, C. (1987) "Linguistic units and units of speech production", *Speech Comm., 6*, 77-142.

[10] SCULLY, C. & ALLWOOD, E. (1985) "Production and perception of an articulatory continuum for fricatives of English", *Speech Comm., 4*, 237-245.

[11] SHADLE, C.H. (1991) "The effect of geometry on source mechanisms of fricative consonants", *J.of Phonetics, in press.*

[12] STEVENS, K.N. (1971) "Airflow and turbulence noise for fricative and stop consonants: static considerations", *J.Acoust.Soc Amer., 50*, 1180-1192.

LINKS BETWEEN SPEECH PRODUCTION AND ACOUSTICS: A SKETCH

Shinji Maeda

Département SIGNAL, CNRS URA-820
Ecole Nationale Supérieure des Télécommunications
46, rue Barrault, 75634 Paris Cedex 13, France

ABSTRACT

This paper reviews some of issues related to speech production. It is not intended to give an overview. Rather our attention is focused to particular issues related, directly or indirectly, to the topics dealt by these papers presented in this symposium. We discuss the specification of speech at different levels in the production chain and the mapping relations between those descriptions.

1. INTRODUCTION

It is our basic assumption that linguistic messages, segmental or suprasegmental, and extralinguistic messages, such as speaker's emotion, are coded into acoustic signals of speech through the phonatory and articulatory processes. These messages are decoded by listeners, establishing speech communication. Presumably, the linguistic messages are represented symbolically by phonological forms within linguistic structures of a particular language. One of the major research goals has always been to break the code, *i.e.*, to relate the abstract symbolic representation in the linguistic description of speech to observable physical signals. We cannot expect a simple mapping relation from the symbolic description to observable physical characteristics, articulatory or acoustic, of the speech however. When we speak, a series of processes at different levels, cognitive, physiological, mechanical, and acoustic, is involved in speech production. The process at each level is a complex phenomenon. Probably, the processes do not operate in a simple sequential manner. Rather they interact in complex ways through both forward and backward pathways. The influences or constraints exercised at the different levels shapes "real-time" characteristics of speech as well as the phonetic structures of speech. These observations give us a good reason for studying the description of speech at different levels of speech production and the mapping relations at different successive levels.

Primarily due to technical difficulty in the acquisition of articulatory data, however, the majority of experimental works has dealt with acoustic data. Working with acoustic signals alone was justified by saying that "we speak to be heard" and that the signals are the intermediate media between the speaker and the listener. Presumably, the acoustic signals contain all information specifying the linguistic description of utterances. The simplest hypothesis can be that the acoustic signals contain context independent invariant features related to phonological forms, despite widely variable aspects of articulatory gestures for realizing the same phoneme in different phonetic contexts. In certain cases, the reality might be more complicated than that. When the portion of signal corresponding to the reduced initial /ka/ is extracted out from /kaka/ for example, what we perceived is /kɛ/ [18]. Only when an onset portion of the second vowel is added to the signal, listeners

perceive the vowel /a/, indicating that the context is necessary for the vowel to be perceived as the intended vowel /a/ instead of /ɛ/. It is necessary, therefore, to understand how the characteristics of acoustic signals vary as a function of the contexts. We have one more reason to study speech on the articulatory levels where the context effects actually occur. It is quite possible that speech perception can be better explained with reference to the description at a level other than acoustics [6]. Although the emphasis was put on the articulatory studies, this does not mean that the acoustic studies are considered as less important. Contrarily, the articulatory studies become more meaningful when they are carried out with the acoustics as the reference. What is really important is to understand how processes at different levels are linked together to shape speech sounds.

In this paper, we do not intend to provide a comprehensive overview of various issues on speech production (for such a review, see [2], for example). We shall focus our attention to the description of speech at different levels of speech production processes, and to the mapping relations between those descriptions.

2. PHONATION

Aspects of the sound generation inside the vocal tract, which is essentially an areodynamic phenomenon, are often put aside, when the segmental characteristics are at issue. The source-filter theory established by Fant [5] has a profound influence in our way of seeing the acoustic characteristics of speech signals. Segmental attributes are often related solely to the filter, *i.e.*, acoustic characteristics of the vocal tract which are the function of its geometrical configuration. An articulatory gesture in speech production is not only to create acoustic effects. It is also to create aerodynamic effects. In order to generate fricative sounds, the constriction is narrowed voluntarily so as to increase particle velocity high enough for the formation of turbulence. In order to produce stop sounds, the constriction has to be closed momentarily to increase the air pressure inside the cavity behind the closure for the generation of consonantal release. Probably, time-varying spectral characteristics of the source sounds contribute considerably to the identity of certain consonants. It is nearly impossible to isolate the spectral contribution of sources from the acoustic signals, however. We need to observe directly the aerodynamic phenomena inside the vocal tract, which is also nearly impossible at present. An alternative is to carry out such measurements on a mechanical model of the vocal tract [14] or to observe in detail the time-varying tract configuration and estimate the aerodynamics by calculations.

The paper presented by Drs. KIRITANI, HIROSE, and IMAGAWA describes a novel instrumentation, a high speed video recorder in connection with an endoscope or a fiberscope, to observe the vocal-fold vibration. The instrument is already useful for clinical applications. Probably it is possible to derive quantitative data, for example time-varying glottal area, from captured video images. In that case, the perspective is an exciting one for speech research. Such data can be used to test theoretical works, for example the classical two-mass model of the vocal-folds [1]. In the past, suprasegmental aspects of speech were studied primarily with fundamental frequency variation along utterances. With glottal area data, we may have an access to evaluate the time variation of glottal pulse shapes, and thus of glottal spectra, which might contribute to signal lexical stress, tones, emphasis, and emotion. Such instrument would be very useful also for observing fast labial release gestures. The flow variation just after the release, which is critical for the spectral characteristics of the burst onset, can be estimated from the observed time variation of the lip opening area [8].

3. TRACT LEVEL

At frequencies below 3 kHz, the sound waves propagate only at the directions of the tract length. The cross-sectional area varying along its length, *i.e.*, an area function, therefore, characterizes the acoustics. The vocal tract is represented by a piece of a straight tube with one end closed (corresponding to the glottis) and the other end opened (to the lips). An arbitrary tube shape, i.e., cross-sectional area variation along its length, must be constrained to account for geometrical shapes of the human vocal tract. Often the shape is parameterized including, in particular, the constriction locations and their degree, that excise a strong influence on the spectral shapes [*e.g.*, 5, 15]. The parameterization of the tract shapes allowed researchers to investigate the mapping relation from the tract shape to the acoustics.

As early as 1955, Fant has studied the relationships between tract shapes and the acoustics in terms of formant patterns, now classical "nomograms", using a four-tube tract model [5]. Stevens noticed that the formant sensitivity to a small variation in the constriction location, specified by the distance from the glottis to constriction point, is not uniform along the vocal tract. Rather there are locations where the sensitivity is relatively high (formant patterns are unstable against a small variation in the constriction location), and other locations where the sensitivity is low, (formants are relatively stable). This observation has lead Stevens to the proposition of "quantal nature of speech" [16]. The concept is generalized to "quantal theory" [17]. Although the theory is still under controversy (*e.g.*, [11]), it seems to explain well the formation of speech sounds in a number of cases.

In their paper, Drs. MRAYATI and CARRE propose another type of model of the vocal tract area function. Instead of constructing a tube model to characterize the tract shapes, as described above, the tube is divided into specific "regions" having fixed lengths solely on acoustic consideration. The division into regions is strictly based on the formant sensitivities to the localized variation of the cross-sectional area. Region lengths remain fixed and only region areas are varied to specify a vocal tract configuration. This model effectively exploits physical properties, manifested on the sensitivity functions, which ensures the maximum modulation of formant frequencies with respect to the neutral uniform tube configuration. One might question how accurately such model can describe observed static or time-varying vocal tract shapes. Even though the model is capable of producing any observed temporal pattern of the formant frequencies, with appropriate variations of the region areas, it could be still merely an equivalent representation of the original formant variations. Nevertheless, it is quite appealing to ask whether or not humans also exploit such physical properties in speech production. If that is the case, they influence upon the formation of sound pattern of speech.

The specification of speech in terms of the tract configuration plays an important function in formulating a scheme for the articulatory control as described later. It serves for the specification of articulation goals. The search of articulatory correlates, for example, of vowel height, such as mandibular height or tongue dorsum height, constantly fails in the past, due to a large observed variabilities of these positions. Contrarily, the acoustic specification of vowels, typically with F1, turned out be much more consistent with the phonological notion of height. Wood has demonstrated, however, that the tract configuration, essentially in terms of the constriction location and the degree of the constriction, results in more coherent relation with vowel height [19]. This is not so surprise in the sense that the area function and the acoustic are tightly related by laws of physics, *i.e.*, the tract configuration determines the acoustic. (The inverse is not necessarily true.) It is reasonable, therefore, to set up an articulatory goal in terms of the tract shape,

which is equivalent to the acoustic goal. This equivalence, however, is meaningful only in the specification of the static configurations, such as articulatory goals.

In the description of the dynamic aspects of speech, the specifications at the tract level and at the acoustic level become distinctively different. The articulatory process involves both spatial and temporal organization. For example, a vowel-consonant-vowel (VCV) sequence is produced by a global vowel articulation of the tongue dorsum plus spatially more localized consonantal gestures with participation of the lips, tongue tip *etc.* [12]. For a longer sequence, such as VCVC.., consonantal gestures can be characterized explicitly in terms of the places and timing relations of the sequential consonantal gestures. The multidimensional organization of speech would appear on the single dimensional temporal pattern of the corresponding acoustic signals. One of difficulties in describing context dependent variability of the acoustic characteristics, such as coarticulation, is due to the fact that the specification of speech at articulatory level, which is inherently multidimensional, is mapped or linked to the single dimensional acoustic space.

Although the spatiotemporal organization of speech might be described at the tract level, there exists an inherent limitation in its effectiveness. In reality, the tract shapes are determined by the states of the individual articulatory organs. The spatiotemporal organization means actually temporal patterns and their inter-timing (phasing) relations of the individual articulators. In the cases where successive articulatory goals are the spatially localized and, say, anatomically separated, for example a sequence of gestures involving the labial, velic, tongue tip, or laryngeal manoeuvres, the description of the spatiotemporal organization at the tract level might correspond in a simple way to that at the articulatory level. In other cases where gestures simultaneously involves the activation of multiple of the articulatory organs, such as a sequence of vowels and consonants including velars, the correspondence can become complex, say, tangled up. Strictly speaking, even localized labial gesture involves the participation of not only the intrinsic labial manoeuvres but also the mandible movements. Therefore, even when the organization appears to be simple and clean at the articulatory level, the same organization can appear to be quite messy at the tract level, and probably worse at the acoustic level. These arguments direct us towards the specification of speech at the articulatory level, especially when the spatiotemporal organization is in question.

4. ARTICULATORY LEVEL

The specification at the articulatory level is characterized by its relatively large variability, as already mentioned before. If the observed variability is random and unexplainable, the articulatory description of speech would be of little value. The question might be raised is what factors cause such variability. One of the causes is motor equivalence. It has been evidenced that the variability of the composite product is significantly smaller than that of individual articulatory movements. For example, the variability in the lip aperture is smaller than that of the lip or of the jaw position [4]. From "bite-block" experiments, Lindblom, Lubker, and Gay concluded that speech production is compensatory in nature [7]. A deviation of the mandibular position, due to an external as the bite block or an internal organizational cause, can be compensated by the readjustment of the other articulators to restore the acoustic characteristics of speech. It may be noted that the motor equivalence is a mapping characteristic between the articulatory and tract levels, whereas the "compensation" is that between the articulatory and acoustic levels.

The paper presented by Dr. WOOD in this symposium adds more evidence to the compensatory phenomenon. From a model experiment, he has demonstrated almost perfect acoustic (F1-F2) restoration in the mandible-labial coordination for the rounded vowels, such as /u/ and /o/. A sound change can occur between two proximate vowels, such as /u - o/, as the consequence of a single gesture, for example, the jaw opening or closing movement. Such sound change can be prevented, if necessary, by maneuvering the other articulators. We have shown also, in a similar model experiment, that such acoustic compensation occurs in the mandible-tongue dorsum coordination for unrounded vowels [9]. Moreover, at least for a limited case, such compensatory relation in the articulatory space can be specified by a simple proportional (or linear) relation between the positions of the paired articulators [10]. This implies that the mapping between the articulatory level and the acoustic level, by-passing the tract level, can be described in simple manner and that the acoustic goals can be specified directly at the articulatory level as relative potions of the specific paired articulators.

In order to model the observed spatio-temporal coordination, we must assume a function that controls the different articulators to achieve a series of targets, and that operates at levels higher than those for controlling the movements of the individual organs. Such a function is called "motor programming". Unfortunately, there is no means to observe directly the motor programming in operation. Electromyographic measurements are at present the best we can do for observing physiological patterns at levels higher than the articulatory level. Still they are limited only to the observation of individual muscle activities. It is important, therefore, to accumulate experimental data at the lower levels and to characterize their behaviors as precise as possible, in order to formulate the most plausible scheme for the articulatory control. The formulation depends on what we observed at lower levels. As mentioned earlier, the specification of speech at the tract level has a tight relation with that at acoustic level and exhibits relatively small variability. Thus in control scheme, such as a simulated feedback [7] or a task dynamic [13], the articulatory goals are specified at the tract level. It is also foreseeable to postulate another scheme, at least for vowels, in which the goals are specified directly at articulatory level, in terms of not absolute positions but of relative position among the individual articulators. In any case, any control scheme must take into account for such links among articulatory, tract, and acoustic levels.

Compensation is, probably, not only the factor involving the articulatory organization. The paper presented by Drs. ABRY and LALLOUACHE deals with the anticipatory articulation in the lip rounding gesture. The individual articulators tend to anticipate for production of upcoming string of segments, unless the anticipatory movement causes a sound change. Henke has explicitly implemented anticipation into the control scheme in his dynamic articulatory model, as "a lookahead operator" [3]. Alternative models for anticipation, as "time-locked", "hybrid", were proposed to explain observed data. The two authors demonstrated that the rounding anticipation could not be predicted by none of the three models and suggested that unpredictable data were due to prosodic effects which were not controlled in their data acquisition separated by two sessions. If this is the case, the segmental patterns in articulation are also influenced by the suprasegmantal factors, such as accents, grouping of words, *etc.*. This implies that the motor programming has to handle both segmental and suprasegmental requirement to issue appropriate commands to the individual articulators. If anticipation depends target positions specified for individual articulators, which also depend how compensation is employed, then

intricate calculations involving both anticipation and compensation are required for the motor programming function. There is reason to believe, therefore, that the articulatory organization and thus speech production process is indeed a complex phenomenon.

REFERENCES

[1] Flanagan, J.L., Ishizaka, K., and Shipley, K.L. (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *The Bell System Technical Journal,* **54**(3) 475 - 506.

[2] Fujimura, O. (1990). Methods and goals of speech production research. *Language and Speech,* **33**(3), 195 - 258.

[3] Henke, W.L. (1966). *Dynamic Articulatory Model of Speech Production Using Computer Simulation.* PhD thesis, Department of Electrical Engineering, MIT.

[4] Hughes, O.M. and Abbs, J.H. (1976). Labio-mandibular coordination in the production of speech: Implication for the operation of motor equivalence. *Phonetica,* **33**, 199 - 221.

[5] Fant, G. (1960). *Acoustic theory of speech production.* The Hague: Mouton.

[6] Liberman, A.M. and Mattingly, I.G. (1985). The motor theory of speech perception revised. *Haskins Laboratories: Status Report on Speech Research,* **SR-82/83**, 63 - 93.

[7] Lindblom, B., Lubker, J., and Gay, T. (1979). Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation. *Journal of Phonetics,* **7**, 147-161.

[8] Maeda, S. (1987). On the generation of sound in stop consonants. *Speech Communication Group Working papers,* Research Laboratory of Electronics, MIT, **5**, 1 - 14.

[9] Maeda, S. (1990). Compensatory articulation during speech: evidence from the analysis and synthesis of vocal tract shapes using an articulatory model. In W.J. Hardcastle & A. Marchal, (eds.), *Speech Production and Speech Modeling,* pp. 131-149. Kluwer Academic Publishers.

[10] Maeda, S. (1991). On articulatory and acoustic variabilities. (to be appeared on *Journal of Phonetics.*)

[11] Ohala, J.J. (ed.), Theme issue on the quantal nature of speech. *Journal of Phonetics,* **17** (1989).

[12] Öhman, S.E.G. (1967). Numerical model of coarticulation. *J. Acoust. Soc. Am.,* **41**, 310 - 320.

[13] Saltzman, E.L. and Munhall, K.G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology,* **1**(4), 333 - 382.

[14] Shadle, C. (1985). *The acoustics of fricative consonants.* PhD thesis, Department of Electrical Engineering and Computer Science, MIT.

[15] Stevens, K.N. and House, A.H. (1955). Development of a Quantitative Description of Vowel Articulation. *J. Acoust. Soc. Am.,* **27**(3), 484 - 493.

[16] Stevens, K.N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In E.E. Devid and P.B. Denes (eds.), *Human Communication: A Unified View,* 51 - 66. New York: McGraw Hill.

[17] Stevens, K.N. (1989). On the quantal nature of speech. *Journal of Phonetics,* **17**, 3 - 45.

[18] Vaissière, J. (1987). Effect of phonetic context and timing of the F-pattern of the vowels in continuous speech. *Proceedings of the 11th Congress of Phonetic Sciences,* Tallin, Estonie, 43 - 46.

[19] Wood, S. (1979). A radiographic analysis of constriction locations for vowels. *Journal of Phonetics,* **7**, 25 - 43.

VOCAL CORD VIBRATION AND VOICE SOURCE CHARACTERISTICS --OBSERVATIONS BY A HIGH-SPEED DIGITAL IMAGE RECORDING--

S. Kiritani, H. Hirose and H.Imagawa

Res. Inst. Logopedics and Phoniatrics,
University of Tokyo, Tokyo, Japan.

ABSTRACT

By using a high-speed digital image recording system, the relationship between vocal vibration and voice source characteristics has been investigated. Fiberscope was used to observe vocal cord vibrations during running speech. Pattern of vocal cord vibration at the onset and offset of the consonants were analysed. Solid endoscope system was used for observing sustained phonation of pathological voices. Many cases of rough voice show asymmetric and/or asynchronous movements between the right and left vocal cords, and between the anterior and posterior parts of the vocal cords. These movements appear to be related the periodical fluctuations in the vibratory pattern.

1. INTRODUCTION

For the study of the voice source characteristics, it is essential to record the vocal cord vibration simultaneously with the speech signal and to analyze the relationship between the pattern of the vocal cord vibration and the acoustic characteristics of speech signal. Observation of the vocal cord vibration has generally been performed by using a high-speed motion picture. However, that method requires special equipment and is not suited for flexible data collection.

In order to facilitate high-speed recording of vocal cord vibration, a special high-speed digital image recording system was developed by the present authors. The system consists of an image sensor and a digital image memory combined with the solid endoscope or the fiberscope. The system is small and compact and, thus, enables flexible data collection.

2. SOLID ENDOSCOPE SYSTEM

Fig.1 shows a block diagram of the system. The system consists of an oblique-angled solid endoscope, a camera body containing an image sensor and an image processor. The output video signal from the image sensor is fed into the image processor through a high-speed A/D converter. Stored images are then displayed on a CRT monitor as a slow motion picture, etc.

In order to obtain a brighter image, a new model of the solid endoscope was constructed. The diameter of the scope was larger than that of the

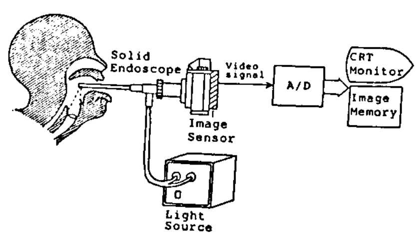


Fig.1 Blockdiagram of the solid endoscope system.

ordinary scope and contains two separate bundles of light guide. Light sources are the two 250W halogen lamps. Number of the picture elements in the image sensor is 100x100 and the sampling rate is 10MHz. In order to realize a high frame rate, a special scan method was devised in which only the selected scan lines were sampled. When 37 scan lines are sampled out of 100 scan lines, the frame rate is 2000 per second. The image memory is 1MByte and can store the image data for the period of about 200msecond.

3. FIBERSCOPE SYSTEM

In order to perform observation of the vocal cord vibrations during consonants in running speech, a high-speed image recording system using a fiberscope was also developed. A special fiberscope was also constructed the diameter of which was slightly larger than that of the ordinary scope. At the same time, a CCD image sensor was employed in this system, because the image by the fiberscope is darker than that by the solid endoscope. The sensitivity of the CCD image sensor is generally higher than that of the MOS image sensor which was used in the solid endoscope system. The light source is a 300W xenon lamp. A frame rate of 2000 per second was achieved with the picture element of 112x32.

Because the main objective of the fiberscope system is the observation of the running speech, it is necessary to record high-speed images during utterances as long as several seconds. Thus, a special stand-alone digital video system was constructed. The system consists of 64MByte image memory which can store the image data of 6.5 second duration under the frame rate of 2000 per second. The system generates standard NTSC video images. It can be operated as an ordinary video tape recorder, and is equipped with following operation modes; PLAY, FAST-FORWARD, REWIND, SLOW-MOTION, STILL, REPEAT. A personal computer can be connected to the system and the image data of the selected part of the utterance were sent to the computer for the later data analysis.

4. VOCAL CORD VIBRATION FOR CONSONANTS

By using a fiberscope system, vocal cord vibrations during the production of the intervocalic consonants were observed.

Figure 2(a) shows vocal cord vibrations during the release of word initial /p/ in the utterance /i: pepe desu/. Up to about 20th frame in the figure, the glottis remains open and there is no appreciable change in the glottal opening. At about 20th frame, narrowing of the glottis starts and the vocal cords start to vibrate. In the first cycle of vibration, glottal closure is incomplete. However, in the next cycle, closure of the glottis becomes complete. It can be seen in the figure that in the subsequent cycles, duration of the closed phase becomes longer. It should be noted that, due to this change in the duration of the closed phase, the interval between successive closing points of vocal cord vibration is shorter during consonant release than during the following vowel. This phenomenon contributes, at least in part, to the higher pitch frequency in the post-consonantal period. It is also noted that the start of the clear excitation in the speech wave corresponds to the appearance of the closed phase in vocal cord vibration. Contrary to this, during the implosion of consonant, the excitation pattern of the speech wave decays even when vocal cord vibrations still maintains complete closure. It appears that the decay of the excitation pattern at this phase is due to the formation of the closure in the vocal tract.

Figure 2(b) shows vocal cord vibration for the production of /b/ (word medial /b/ in /i: bebe desu/). Vocal cord vibration during the consonant closure and during the preceding vowel are compared. There is no appreciable difference in the pattern of vocal cord vibration during vowel /e/ and consonant /b/. In the present images, it is difficult to find any apparent indication that the glottal constriction is looser for /b/ than for the vowel. The closed phase is even longer for /b/.

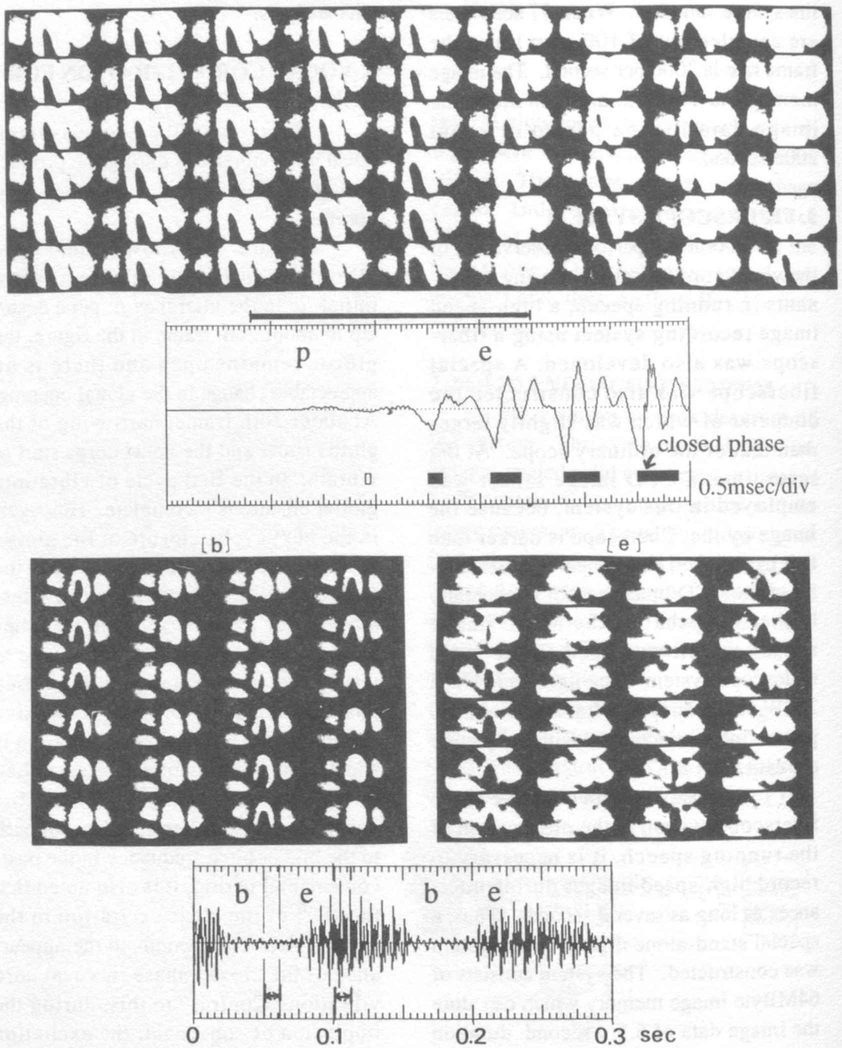


Fig.2 Vocal cord vibration in the production of the stop consonants.
(a) /i: pepe desu/ (b) /i: bebe desu/ Time: top to bottom, left to right.

5. VOCAL CORD VIBRATION IN ROUGH VOICES

The solid endoscope was applied to the observation of pathological vocal cord vibration in rough voices. Generally, rough voices have cycle to cycle fluctuations in speech waveform. However, many cases of rough voices show characteristic pattern of fluctuations (i.e., not purely random fluctuations). Fig. 3(a) shows a case of vocal fold cyst. The speech wave shows alternation of the period of strong excitation and the period of weak, noisy excitation. Glottal images in the figure reveal that there is a timing difference between the movements of the anterior and the posterior part of the vocal cord. The pattern of this timing difference varies from cycle to cycle. It appears that, in the period B, anterior part of the glottis starts to open as soon as the posterior part gets closed.

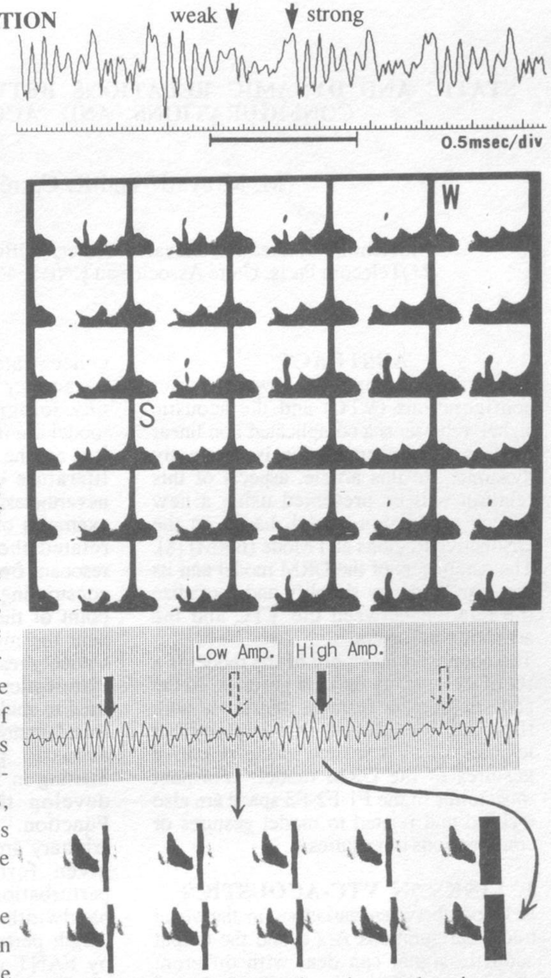


Fig. 3 Vocal cord vibrations in rough voices.
(a) Vocal cord cyst. (b) Recurrent nerve palsy.

Fig. 3(b) shows a case of recurrent nerve palsy. In this case, the left and the right vocal cords show the difference in the vibratory frequency. During the preiod of high-amplitude waveform, the movements of the left and right vocal cords are in phase. In the following cycles, the left vocal cord gradually gets behind the right vocal cord and the movements of the vocal cords become out of phase. The vocal cords moves almost in pararell and the closed phase disappears. This state corresponds to the periods of the low-amplitude waveform. When the phase difference becomes greater than certain threshold, the vocal cords appear to pull each other and the movements again become in phase.

These examples demonstrate that the present system is useful for the analysis of the production of the pathological voices.

1) S.Kiritani, K.Honda, H.Imagawa and H.Hirose: Proc. ICASSP, Tokyo, 1633 (1986).

STATIC AND DYNAMIC RELATIONS BETWEEN VOCAL TRACT CONFIGURATIONS AND ACOUSTICS

M. Mrayati* and R. Carré**

*)Scientific Studies and Research Center, POBox 4470, Damascus
**)Telecom Paris, Unité Associée au CNRS, 46 rue Barrault, Paris

ABSTRACT

The relation between vocal tract configurations (VTC) and the acoustic signal represents a complicated non linear process. This phenomenon is inherently dynamic. In this article, aspects of this relation will be presented using a new speech production model, based on the Distinctive Regions and Mode (DRM) [8]. The parameters of the DRM model and its command strategy simplify and formalize the relation between the VTC and the acoustic parameters.

The command inventory of the model is a set of dynamicaly defined gestures. In the acoustic domain formant transitions with time are formalized into a primitive lexicon which is related to the primitive gestures of the DRM model. Formant trajectories in the F1-F2-F3 space are also stylized and related to model gestures or combinations of gestures.

1. LINKS IN VTC-ACOUSTICS

Relations between variations in the vocal tract configurations A(x,t) and the output acoustic signal can deal with different parameters of this signal. Examples of acoustic parameters of the signal are: (1) formant frequencies Fn, (2) formant transitions in the time domain (FT), (3) formant bandwidth (Bn), (4) formant trajectories in the F1-F2-F3 space, (5) noise type, frequency, intensity and duration. One can trace two main schools relating to the consideration of the variations of the VTCs. The first emphasises the articulatory aspect involved in the process, and consequently, concentrates on articulatory models and parameters and then relates the acoustic output to them. The second school emphasises the acoustic tube depicting the VTC, and as such concentrates on acoustic models and parameters taking into account physiological constraints. The DRM model is a model of the second school.

It is not the object of this article to review literature on VTC-acoustic relations, nevertheless, it is helpful to mention some exemples of it. Chiba and kajiyama [4] related the increase or decrease in a resonant frequency of an acoustic tube to constricting the tube near the maximum point of the pressure standing wave, or near the maximum point of the volume velocoty respectivly.

Perturbation theory has been successfully used to study the relation between a small area-function variation and a crresponding acoustic parameter variation [10]. Starting in 1967 Fant used this theory to develop the concept of Sensitivity Function. Sensitivity functions of an arbitrary area function A(x) relate, for a given formant, small local spatial perturbations to formant frequency or bandwidth. Sensitivity functions for length perturbations are also introduced by FANT and served as a measure of "formant-cavity" affiliations [6].

Sensitivity functions can also help to define formant stability as a function of local perturbations. The DRM model is based on the concept of sensitivity function.

Several models of the vocal tract have been proposed to study the relation between VTC and formant frequencies. These models are capable of providing insight into some area-function-acoustic relations or articulatory-acoustic relations. Fant [7] elaborated a model composed of four cavities representing the vocal tract. He provided nomograms relating the five formant frequencies to the dimensions of these four cavities. These nomograms reflect the variations in the formant frequencies due to variations in place and area of the constriction.

As early as 1955 Stevens and House [11] proposed a three-parameter model for vowel production. They presented several nomograms relating acoustic parameters (F1, F2, F3) to articulatory parameters (place of articulation, degree of constriction, and lip parameter).

This article presents the use of a new model of the vocal tract to study the VTC-acoustic relations. Results on simple and formalized relations between model gestures and formants are presented.

We argue that in the research for relations between speech production and acoustics, the adopted model should not overlook important considerations such as: First, the model should be part of, and coherent with, an underlying unified concept linking the different phases involved in the speech communication process; namely: the central representation and motor control, the articulatory, the acoustic and the phonologic levels. Second, the acoustic signal is actually the output of an apparatus for the conversion of muscular energy into acoustic energy. This conversion is hypothesized to be effecient. The modeling of this apparatus and the choice of its parameters and its command strategy are crucial for the right detection and explanation of the relations between the different levels of the speech communication process. Third, speech production is inherently a dynamic process. We are forwarding the hypothesis that the DRM model incorporates these considerations [9].

2. THE DRM MODEL, GESTURES AND ACOUSTICS

For an acoustic tube, closed at one end open at the other, there exist distinctive spatial regions (R) having specific Formant Transition Behavior (FTB). These FTBs are monotonic as long as the variations of the cross sectional area (S) of the different regions (R) are within specific limits defining a mode denoted One Tract Mode (OTM) (approximately between 1 and 16 cm², if the neutral position is 4 cm²). Two other modes can also be defined depending on region cross sectional areas. These are the Transition Mode (TM) corresponding to narrow S(R) (approximately between 0.05 and 1 cm²); and the Two-Tract Mode (TTM) corresponding to practically closed S(R). These regions and modes are deduced from sign changes of sensitivity functions of the uniform tube (Figure 1). If one is interested in the first three formants, eight regions, can be distinguished.

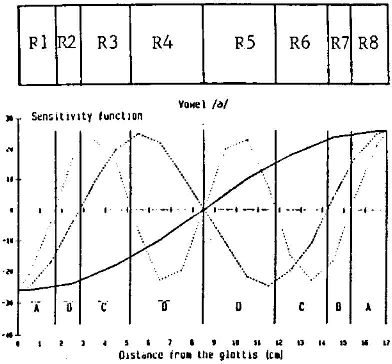


Fig1 - Eight spacial regions from the sensitivity functions for the three formants obtained from a uniform tube and the DRM model.

These regions have the following lengths respectively: L/10, L/15, 2L/15, L/5, L/5, 2L/15, L/15, L/10 (L being the total effective length of the tube). It is clear from the sensitivity functions that, for each of the regions, an increase or a decrease of its cross sectional area leads to a well defined variation sign of the three formants (as far as the OTM is concerned). This aspect distinguishes the eight possible variation signs (three formants with 2 possible variation signs for each result in 8 combinations. The four front regions are antisymetric with the four back ones. As a consequence, an increase in one of the front regions can acoustically compensate a similar increase in the corresponding back region (compensation effect) and area changes in the opposite direction involves a maxmial formant change (synergetic effect). If one considers one resonant mode only, there exists two regions. When only the first two formant are considered, four regions have to be taken into account.

It can be shown that region area changes around the neutral result in efficient simultaneous three formant modulations (efficient transition). When the cross sectional area of a region varies substantially away from the neutral, the change in formant frequency saturates. Such cases represent quantal acoustic targets (Tac) [9]. The definition of fixed spatial regions, delimited for the neutral, permits the existence, in a same model, of efficient transitions of formants on the one hand and stable targets on the other hand. The transversal command strategy for controlling the DRM model is simple, and can take advantage of the inherent synergy principle.

The comand of the model is achieved by means of what we call region-gestures S(R,t). There are eight classes of primitive region-gestures, one for each region (R1 to R8). The parameter of each gesture are the span S(R) and the tempo S(R,t). These gestures are defined around the neutral. In simple or primitive uterances, such as(əCə) one region-gesture is involved. In natural uterances serveral region-gestures are combined. Actually, the complete inventory of gestures involves other types of gestures such as the velic and the glottal ones.

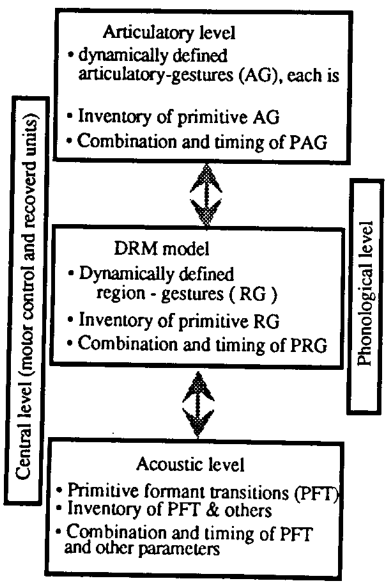


Fig 2 - The DRM model as a gestural acoustic model and its relation to acoustic and articulatory levels.

In the articulatory domain, it is postulated that to realize a region-gesture, one or a group of articulator is controlled in a coordinated manner. The units of the articulatoy gestures are dynamically defined. The articulatory gestures producing the primitive region-gestures could form an inventory of articulatory primitive gestures. These gestures are combined to form natural uterances.

In the acoustic domain, the units corresponding to the primitive region-gestures are specific patterns of formant transitions in addition to other units such as noise patterns. Figure 2, shows a schematic representation of this concept, the central representation level and the phonological level could be added, but it is out of the scope of this study.

It is supposed that in each of these domains, the set of units form a primitive inventory. These primitive units are combined or organized in the temporal domain, taking into acount anatomical, acoustical, phonological and perceptual constraints.

In the following paragraphs we shall concentrate on the relation between the region- gestures and the formant statics and dynamics in the acoustic domain.

3. STATIC RELATIONS

In this paragraph, we point out certain static VTC-acoustics relations dèmonstrated by the DRM model. It can be shown that any area-function A(x) representing a VTC can be modeled using the DRM model and consequently mapped into a point in the vocalic space F1-F2-F3. It can also be demonstrated that the model can produce a vocalic space larger than any other known model of comparable limits on A(x). This VTC-acoustic property of the DRM model is due to the pseudo-orthogonality of its region-gestures [8] [2]. Thanks to the symetrical aspeats of the model, it can incorporate the influance of all parts of the vocal tract. One consequence of this is that what is important in some vowel production is not only the dègree of constriction, but also the opening of the cavity which is at distance from the mid-point equal to that of the constriction.

It has been shown that the model incorporates inherently quantal acoustic tragets Tac [9]. For exemple it is shown, that when the VTC defined by the model becomes close to that of a cardinal vowel, one formant at least becomes quasi-stable, i.e. $dFn/dS(R) = 0$ (see figure 1 in [9]). This interesting VTC-acoustic relation deserves extensive investigation.

4. DYNAMIC RELATIONS

The speech production process is inherenty dynamic and non linear. In this paragraph, we shall treat briefly two aspects of the OTM mode of the DRM model. The first one is the relation between the primitive region-gestures (PRG) and the primitive formant transition behavior (PFT). These gestures are combined in natural utterances to produce all possiple formant patterns. The second one is the relation between region-gestures and corresponding trajectories is the formant space F1-F2-F3. Particularly, trajectories of vowel-vowel transitions (V1-V2) are formalized and related to region-gestures.

4.1 Relations between PRG and PFT

Starting from a neutral tube as a reference (schwa or /ə/), we define a simple PRG to be the complete or partial closure of a region. We have, eight such PRGs, one for each region. Knowing that the regions of the DRM model are favorite places of articulation of consonants [8], therefore the eight PRGs represent sounds of the type /əC/. For example closing region eight (R8: the lips) will lower the three formants, i.e. produces the corresponding PFT. Inversing a PRG, i.e. opening a region, will produce the inverse of the PFT (i.e. raising the three formants in our example). Combining the PRG with its inverse will simulate sequences of the type /əCə/. Figure 3 schematizes the eight PFTs corresponding to the eight PRGs. These region-gestures and their acoustic counterparts are confirmed by well known data on natural speech. Figure 4 gives examples that can be found in the literature and mainly by Delattre [5]. We added five cases for five arabic consonants having their place of articulation in the back half of the vocal tract, (complete results on particular arabic consonants are being prepared for publishing). The following remarks are worth mentioning:

(1) the vowel /ε/ is presented for natural speech because data on /ə/ is not available; (2) all the PFTs given for the front half of the vocal tract are those reported by Delattre except for /ʃ/ and /ʒ/ where we presented them non labialized, where the F3 transition is inverted to keep the gestures primitive; (3) other PRGs are defined for the DRM model and the combination of serval gestures can produce any formant pattern [2].

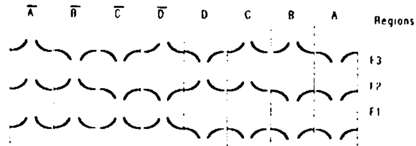


Fig 3 - The schematized eight PFTs corresponding to the eight PRGs.

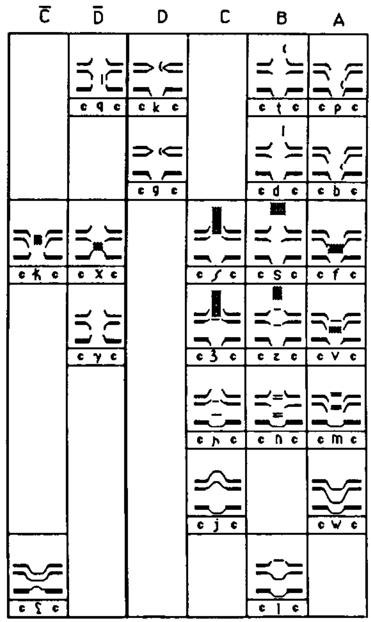


Fig 4 - /εCε/ transitions deduced from the natural speech, with corresponding regions marked.

4.2 Region-gestures and trajectories in the formant space

The mapping of Region Gestures into the vocalic space F1-F2-F3 has certain pseudo-orthogonal properties. For clarity purposes these trajectories are calculated by computer simulation for a four region model and F1-F2 only.

The model trajectories, their main axis, their spans, and the parameters translating them in the vocalic space, are analysed

[3]. Comparing such trajectories with those obtained for natural V1-V2 utterances, one can investigate and understand the rôle played by the regions of the vocal tract in producing V1-V2 sounds.

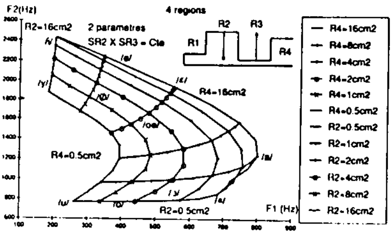


Fig 5 - Trajectories obtained in the F1-F2 plane. Four regions are taken into account. R1 is equal to 1.4 cm^2. Regin areas are varied logarithmically from 0.5 to 16 cm^2. The product of R2 area by R3 area is constant (synergetic area command).

Figure 5 shows an example of trajectories in the F1-F2 plane, corresponding to the combination of two gestures simultaneously (synergetic mouvement). Analysis of vocalic trajectory representing a natural utterance can be achieved by projecting it on such a plane. For exemple, the trajectory of a natural /ai / transition is easily compared with that produced by the model as in figure 5 and for R4 =16 cm^2.

The DRM model is capable of producing any vocalic traget and realistic trajectories by means of simple region-gestures.

5. CONCLUSION

The new DRM model, has been used to show new insight into the relation between VTCs and formant dynamics. Uterances of the form /əCə/ were analysed. The consonants /C/ were Arabic pharyngial ones. Results confirmed the formant-transition patterns predicted by the DRM model for regions R3 and R4 in the back half of the vocal tract ($\bar{C}$ and $\bar{D}$).

The model with its parameters and command strategy seems to be appropriate for the gestural task dynamic approach of the speech communication process [1]. Finally, the model has an explanatory power in relating VTCs to acoustics.

A nasal tract could be added to the model using the same DRM concepts, while noice production in the TM mode is to be investigated.

REFERENCES

[1] BROWMAN, C.P., and GOLDSTEIN, L. (1987), "Tiers in articulatory phonology, with some implications for causal speech", Haskins Laboratory, Status Report on Speech Research, SR 92, 1-30.

[2] CARRE, R., MRAYATI, M. (1990), "Articulatory-acoustic-phonetic relations and modeling, Regions and Modes", in *Speech production and speech modeling*, (W.J. Hardcastle and A.Marchal, editors), Kluwer Academic Publishers.

[3] CARRE, R., and MRAYATI, M., (1990), "Vowel-vowel trajectories and Region modeling", 2nd Seminar on Production: Models and Data, Leeds 13 - 15 May. To appear in the "Journal of Phonetics".

[4] CHIBA, T. and KAJIYAMA, M., (1941). "*The vowel. Its nature and structure*", Tokyo

[5] DELATTRE, P. (1967). "Des indices acoustiques aux traits pertinents", Proc. of the 6th ICPHS, 35-46.

[6] FANT, G., (1980), "The relations between area functions and the acoustic signal", phonetica, 73, 55-86.

[7] FANT, G., (1960). "*Acoustic theory of speech production*", Mouton, the Hague.

[8] MRAYATI, M. CARRE,R. and GUERIN,B. (1988), "Distinctive regions and modes : a new theory of speech production", Speech Communication, Vol 7, 257-286.

[9] MRAYATI, M. CARRE, R. and GUERIN, B. (1990), "Distinctive Regions and Modes: articulatory-acoustic-phonetic aspects", Speech Communication, 9, 231-238.

[10] SCHROEDER, M. (1967), "Determination of the geometry of the human vocal tract by acoustic measurments", J. Acoust. Soc. Am. 41, 1002-1010.

[11] STEVENS , K.N. , and HOUSE, A.S. (1955), "Development of a quantitative description of vowel articulation", J. Acoust. Soc. Am., 27, 484 - 493.

VOWEL GESTURES AND SPECTRA: FROM RAW DATA TO SIMULATION AND APPLICATIONS

Sidney A. J. Wood

Universities of Umeå and Lund, Sweden

ABSTRACT

This paper discusses the relationship between vowel gestures and the spectral output. Some sample model experiments on gestures analysed from x-ray films are presented, and examples are given of phonological applications (the tense vs lax contrast, Bulgarian vowel reduction, assimilations in Circassian).

1. GESTURE TO SPECTRUM

The single-cavity theory related timbre to the size of the front cavity. One interpretation (Roudet) claimed that compensations, such as trading lip rounding and tongue retraction, were unforeseeable and that articulation was therefore irrelevant [12]. On the other hand, Bell [1,73-74] postulated that defined settings of tongue height and retraction would yield predictable configurations and timbres. This theory was only discarded once phoneticians had accepted at least two formants. Paget [11, ch. 3-5] assigned them to the throat and mouth cavities respectively, and he also assumed there were other cavities for additional resonances.

Joos [6,57-59] pointed out that all resonances are modes of oscillation of the entire vocal tract and cannot be ascribed to minor side chambers. A gesture contributes to them all. Thus, there is no simple causal relationship between height and F1, or retraction and F2. The spectral consequences of a gesture depend on how it widens or narrows the vocal tract locally with respect to the standing wave of each resonance mode [3,4,5,9,15; see also 2,10]. Seen from this vantage point, Roudet's compensations and Bell's vowel configurations do not work [13,19]. Following [3,4,5,9,15], we can get a rough idea of the spectral consequences of a gesture from how it shapes the vocal tract locally in relation to the nodes and antinodes of each resonance mode. My own calculations on configurations from several languages show that the nodes and antinodes never wander very far from the theoretical locations given in [3].The relative sensitivity of each mode to the gesture can be judged from the local distributions of kinetic and potential energies. Published energy data based on Russian [4], French [8] and Arabic [16] are very similar for similar vowels, and thus mainly reflect universal acoustical properties of the vocal tract rather than language specific habits.

Published nomograms give the magnitude of a formant shift that can be attributed to a gesture, although the three-parameter models [4,14] are difficult to understand in gestural terms, as they are really models of the area function, and not of the manoeuvres that created it. Lindblom and Sundberg [7] have done a similar systematic mapping of the spectral consequences of manipulating a simulated vocal tract.

The alternative to reading nomograms is to simulate the gesture on a vocal tract profile and calculate the vocal tract resonances for successive configurations. This is the approach used for the model experiments reported here, that were designed to elucidate the spectral contributions of vowel gestures analysed from x-ray motion films of speech. The method is outlined in [17].

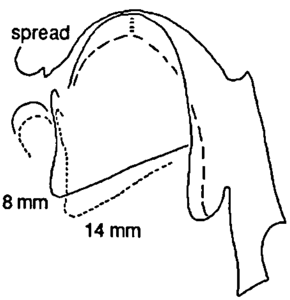
2. VOWEL GESTURES

The gestures were analysed from x-ray motion films by tracking articulator displacement between successive picture frames [17,22]. The actual gestures were as follows. (1) Four tongue body manoeuvres (with respect to the mandible) directed towards the hard palate, velum, upper pharynx or lower pharynx [16], (2) mandibular depression, (3) lip rounding with respect to the jaws, (4) tongue blade elevation or depression with respect to the tongue body, and (5) larynx depression.

3. MODEL EXPERIMENTS

3.1 Palatal vowels

The analysed x-ray profiles revealed a set of [i-ɪ]-like vowels with closer jaw openings, typically 6-9 mm, and a more open set of [e-ɛ]-like vowels, with jaw openings 9-14 mm [18,21]. They are further divided into *tense* [i-e]-like (with a bunched tongue posture) and *lax* [ɪ-ɛ]-like (with a flatter tongue). An example of a model experiment is given here in Fig. 1. The tongue was modified in 6 steps from a bunched to a flatter posture, at two jaw positions (8 and 14 mm). At 8 mm, the palatal passage ranged from 0.3 to 2 cm^2 cross section, and at 14 mm from 1 to 3.2 cm^2. This also reproduces the frequently observed phonomenon that the tongue is *higher* (the palatal passage is narrower) for [e] than [ɪ].



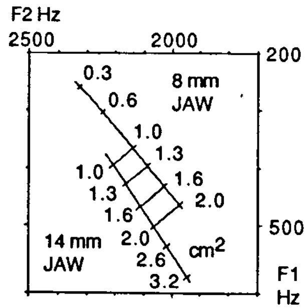
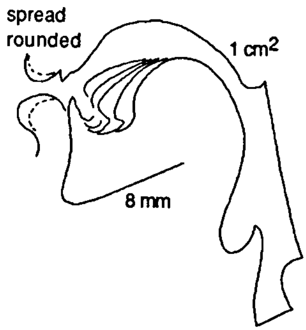


Figure 1. Left: the 6 degrees of palatal tongue bunching (relative to the mandible) at two typical jaw positions for [i-ɪ] and [e-ɛ]-like vowels; the profile illustrates the two extreme postures at the closer jaw position; the same six postures were then repeated at the more open position. Right: the resulting F1 and F2 frequencies at the 6 degrees of palatal constriction and two jaw positions.



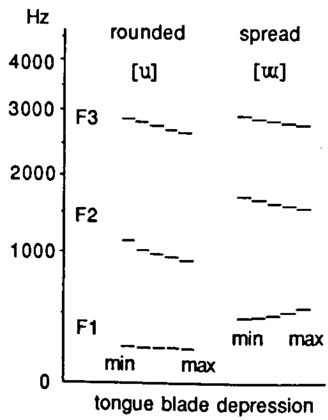


Figure 2. The effect of tongue blade position on palatovelar [u] and [ɯ]-like vowels (close jaw opening).

Other experiments [19] demonstrated the role of larynx depression in rounded palatal vowels and revealed some quantal relationships between component gestures and their spectral correlates. The experiments reported in [18] successfully reproduced the variation observed in natural speech, provided a precision of at least ±1 mm was allowed for compensatory interplay between the tongue and mandible.

3.2 Palatovelar vowels

The [u-ʊ] and [ɯ]-like vowels also have a close jaw opening ranging 6-9 mm. *Lax* [ʊ] has a less bunched tongue body (a less constricted faucial passage and less ATR), while the lips are less rounded and the larynx less depressed [21].

The example in Fig. 2 illustrates the effect of tongue blade depression on palatovelar vowels. X-ray pictures usually show how languages contrasting [u] and [y] have the tongue blade tightly depressed in [u], which increases the gravity of this vowel (lower F2). The less grave English [u] (higher F2) has traditionally been ascribed to an advanced tongue body whereas it is in fact due to tongue blade elevation.

Other experiments confirm that very efficient restoration of lip rounding is needed to compensate for mandibular variation in [u]. A similar experiment for [o] is described in the next section.

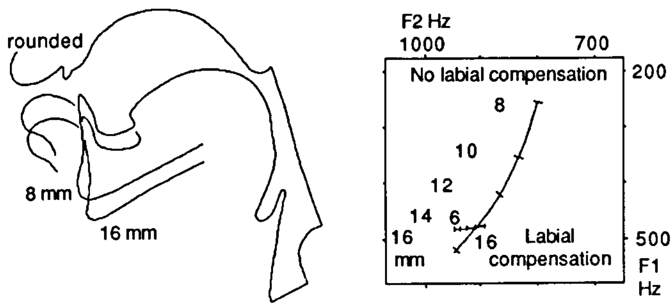


Figure 3. The effect of varying the jaw opening from 8 to 16 mm in an [o]-like vowel, both with labial compensation (the lip opening at 14 mm jaw opening was preserved, while the jaw was stepped from 16 to 6 mm), and without labial compensation.

3.3. Pharyngovelar vowels

Pharyngovelar (uvular) [o-ɔ] and [ɤ]-like vowels have an open jaw opening (9-14 mm). *Lax* [ɔ] has less lip rounding, less tongue blade depression, less ATR and less larynx depression. The example in Fig. 3 shows the success of perfect labial compensation for mandibular variation. Figure 3 also demonstrates that uncompensated mandibular variation in [o] can shift the spectrum towards an [u]-like timbre (i.e. the close, rounded palatovelar and close, rounded pharyngovelar configurations are spectrally ambiguous).

3.4. Low pharyngeal vowels

The low pharyngeal [ɑ-æ]-like vowels are usually produced with a jaw opening of 9-14 mm. The narrowed lower pharynx is widened from about 0.5 cm^2 for [ɑ], through 1.5 cm^2 for [a], to 2.5 cm^2 for [æ]. Grave [ɑ] is darkened to [ɒ] by adding lip rounding. The example at Fig. 4 illustrates the effect of varying the degree of low pharyngeal constriction, together with spread and rounded lips, all at one 14 mm jaw opening.

An interesting finding in [23] was that the "indeterminate" Bulgarian vowel /ǎ/ has an [a]-like low pharyngeal constriction but a close jaw position. This is a configuration that is precluded by the Bell model (where the mandible is explicitly disregarded and where vowels implicitly cannot be both *high* and *low*).

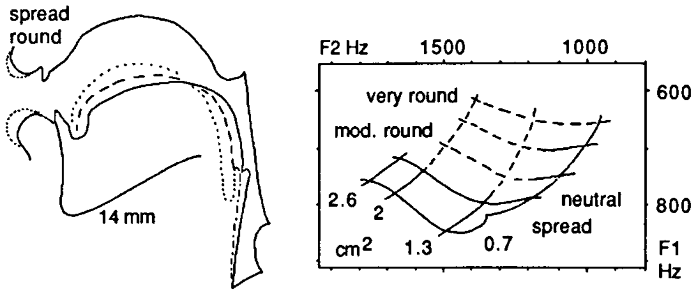


Figure 4. The effect of varying the degree of low pharyngeal constriction from 0.7 cm^2 to 2.6 cm^2 cross section, and lip rounding, at one mandible position (14 mm). The dashed areas are not generally relevant in speech.

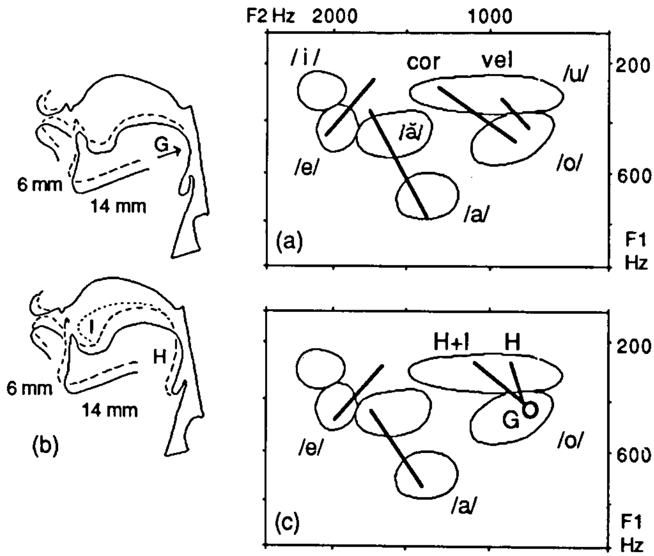


Figure 5. Simulation of vowel reduction in Bulgarian: (a) F1/F2 regressions for all /a,e,o/ (stressed and unstressed), also showing coronal and velar allophones of /o/, (b) simulated mandible variation 16-6 mm for /o/ (G perfect lingual and labial restoration, H no lingual or labial restoration, I tongue blade elevation), (c) simulated spectral reductions.

4. APPLICATIONS

The simulation of vowel gestures in model experiments is a useful tool for solving phonological problems. An example is vowel reduction in Bulgarian [23], where unstressed open /e,a,o/ merge with the reflexes of close /i,ă,u/ (Fig. 5a). The hypothesis was that this results from the speaker omitting gestures that otherwise ensure the spectral contrasts. Fig. 5b illustrates the case of /o/. If lingual and labial compensation are turned off in unstressed syllables, and the jaw opening weakened, the spectrum becomes ambiguously similar to that of [u] without the tongue articulation having to be reorga-

nised to palatovelar (cf. Fig. 3). The simulations reproduced the observed regressions (Fig. 5c). A further point is that /o/ (stressed and unstressed) was brighter (higher F2) after coronal consonants. This was simulated by having the tongue blade elevated (I in Fig. 5c).
The spectral ambiguity arising from some configurations may also offer an explanation for some anomalous vowel assimilations in the Circassian language Kabardian [20]. The close phoneme is [u] or [ɯ]-like in pharyngovelar contexts, i.e. they may well be pharyngovelar rather than palatovelar. This mirrors the reduction of Bulgarian /o/. Close pharyngovelar is one more configuration that is alien to the Bell model. Finally, there is a similar alternation in Kabardian between [o] and [ɑ] in pharyngovelar environments that ostensibly seems to involve a tongue shift in only the one case. An alternative explanation is that rounded pharyngovelars and rounded low pharyngeals can be spectrally ambiguous. This is currently being investigated.

5. REFERENCES

[1] BELL, A. M. (1867), *"Visible Speech"*, London: Simpkin, Marshall.
[2] BOE, L. J. & PERRIER, P. (1990), "Comments on 'Distinctive regions' by M. Mryati et al", *Sp. Com.* 9,217-230.
[3] CHIBA, T. & KAJIYAMA, M. (1941), *"The Vowel, its Nature and Structure"*, Tokyo: Phon. Soc. Jap.
[4] FANT, C. G. M. (1960), *"The Acoustic Theory of Speech Production"*, The Hague: Mouton.
[5] FANT, C. G. M. (1980), "The relation between area functions and the acoustic signal", *Phonetica* 37,55-86.
[6] JOOS, M. (1948), *"Acoustic Phonetics"*, suppl. *Lang.*24.
[7] LINDBLOM, B. E. F. & SUNDBERG, J. (1971), "Acoustical consequences of lip, tongue, jaw and larynx movements", *J.A.S.A.* 50,1166-1179.
[8] MRYATI, M. & CARRE, R. (1976), "Relations entre la forme du conduit vocal et les traits caractéristiques acoustiques des voyelles françaises", *Phonetica* 33,285-306.
[9] MRAYATI, M., CARRE, R. & GUERIN, B. (1988). "Distinctive regions and modes; a new theory of speech production", *Sp. Com.* 7,257-286.
[10] MRAYATI, M., CARRE, R. & GUERIN, B. (1990), "Distinctive regions and modes; a reply to Boë and Perrier", *Sp. Com.* 9, 231-238.
[11] PAGET, R. (1930), *"Human Speech"*, London: Kegan Paul, Trench, Trübner.
[12] ROUDET, L. (1911), "La classification des voyelles de M. Sweet", *Rev. de Phon.* 1,347-356.
[13] STEVENS, K. N. (1972), "The quantal nature of speech: evidence from articulatory-acoustic data", in *Human Communication, a Unified View*, ed. by E. E. David & P. B. Denes, 51-66, New York: McGraw Hill.
[14] STEVENS, K.N. & HOUSE, A. S. (1955), "Development of a quantitative description of vowel articulation", *J.A.S.A.* 27,484-495.
[15] STEVENS, K. N. & HOUSE, A. S. (1961), "An acoustical theory of vowel production and some of its implications", *J. Sp. H. R.* 4,303-320.
[16] WOOD, S. A. J. (1979), "A radiographic analysis of constriction locations for vowels", *J. Phon.* 7,25-43.
[17] WOOD, S. A. J. (1982), "X-ray and model studies of vowel articulation", *Working Papers* 23,1-41, Dept. of Linguistics, Lund University.
[18] WOOD, S. A. J. (1982), "Radiographic and model studies of the palatal vowels", *Working Papers* 23,119-155, Dept. of Linguistics, Lund University.
[19] WOOD, S. A. J. (1986), "The acoustical significance of tongue, lip and larynx maneuvers in rounded palatal vowels", *J.A.S.A.* 80,391-401.
[20] WOOD, S. A. J. (1990), "Vertical, monovocalic and other 'impossible' vowel systems; a review of the Kabardian vowels", *Working Papers* 36,191-212, Dept. of Linguistics, Lund University.
[21] WOOD, S. A. J. (1991), "A radiographic and model study of the tense vs. lax contrast in vowels", in *Phonologica 1988*, ed. by W. Dressler et al, Cambridge University Press.
[22] WOOD, S. A. J. (in press), "X-ray data on the temporal coordination of speech gestures", *J.Phon.*
[23] WOOD, S. A. J. & PETTERSSON, T. (1988), "Vowel Reduction in Bulgarian; the phonetic data and model experiments", *Fol. Ling.* 22,239-262.

AUDIBILITY AND STABILITY OF ARTICULATORY MOVEMENTS
Deciphering two experiments on anticipatory rounding in French

C. Abry
M. T. Lallouache

Institut de la Communication Parlée, CNRS URA 368
Grenoble, France

ABSTRACT

The proposal that the search for economy principles in movement control could take advantage of the potentiality that some phases and/or components of articulatory movements are *poorly* or *non-audible*, is illustrated by two experiments on anticipatory coarticulation for French rounding.

At a first guess, within speaker variability between sessions, and within session variability for the same speaker, could simply point out that poorly audible protrusion movements through five consonants [kstsk] complexes are «free» to vary, provided they reach their audible goal, in our case the vowel [y].

However, a trend to proportional stability in one session, where the eliciting technique allowed to improve prosodic pausing control, indicated that, in complex clustering tasks, prosodic mastery can counteract effectively the variability induced by complexity, ultimately preventing poorly audible movements – in spite of their increased sluggishness – to become «bumpy».

0.FOREWORD

The topics of the present paper were not directly oriented towards issues currently held in relating speech production and acoustics : they were initially dedicated to movement control, namely *anticipatory rounding* behavior. Of course, the fact that unavoidable and invaluable theories – proponents of different *control spaces* for speech production – were encountered «by-the-way» was not completely beyond our control, since we are currently interacting with colleagues on quantal experience [1] and experiments in learning gestures-from-sounds [2, 9]. So the proposal we are holding that, in search of economy principles for movement control, one could take advantage of the potentiality that some phases and/or components of articulatory movements are *poorly* or *non-audible*, this proposal will not receive here a thorough treatment, but just preliminary support from some evidences of such loose links between production and acoustics.

1.PARADIGM & EXPERIMENTS

On testing competing models of anticipatory rounding – the so-called *look-ahead* (LA), *time-locked* (TL), and *hybrid* (H) models, following the procedure set up by Perkell [11] – with French data, we had the opportunity to observe a lot of variational behavior, for the *same* speaker, between and *within* sessions (2 sessions spaced by about 6 months). Since we will focus our attention on the variability *within the same task*, examples presented here will not include those obtained by manipulating the number of consonants and the position of the juncture within [i–>y] transitions. Thus we chose deliberately the most complex case : the «mirror» sequence [...ikstsky...] in *Ces deux Sixte sculptèrent* «These two Sixte [popes] sculptured» (the classical French *sinistre structure* [4] appeared to be unpronounceable without schwa, even by Northern speakers). A mean representative of [iky] tokens was selected only as a «control» reference.

Four illustrations are shown on Figs 1 & 2, displaying upper lip protrusion time functions, with kinematic events, which were detected manually on instantaneous velocity and acceleration functions (derived from cubic spline functions fitted to raw measurements on each 20 ms video field; for more details see [10]). An *obstruence interval* was determined on the synchronized audio signal (sampled at 16KHz) by detecting [i] offset and [y] onset, corresponding respectively to the disappearance and appearance of a clear vocalic formant structure. Among parameters other than upper lip protrusion (this one is chosen here mainly for comparison with [11]), image processing enabled us to track between-lips area from front views. Additional cepstral formant tracking and sections were checked when needed.

Events of ten samples for each session are overlaid on Fig. 3. Perkell's conventions have been essentially adopted, though this presentation suffers from statistical artefact, i.e. part-whole correlations ([5] : this has been taken into account, at least in part, in further work, pers. comm.). Protrusion kinematic events are referred to [y] onset (upper plot) or [i] offset (lower plot). This mirror image is adopted simply to bring out all the possible correlations (regression slopes and intercepts being entirely redundant). Regressions traced here are only those which reached significance at $p<0.01$. Symbols are of course aligned vertically for each token. The minimum protrusion events in the *1st session* are linked by the vertical lines.

Results will be first discussed qualitatively (movement *profiles*); then quantitatively (*dates* of events).

2.VARIABILITY & STABILITY OF MOVEMENT PROFILES

During the 2nd session, all three movement *profiles* – characterized by Perkell [11] – were observed : (i) a unique ramp, with a nearly constant slope, i.e. a *one-phase* gesture (a variant of it, corresponding to a *bell shape velocity profile*, is shown on Fig. 2, upper part); (ii) no (or a weak) movement phase, followed by a rather steep start of protrusion, i.e. a *two-phase* protrusion (a preretracted variant is seen in Fig. 2, middle part); (iii) an initial ramp-like phase, followed by a steeper phase, i.e. *two phases* again (the example shown on Fig. 2, lower part, is an extreme case, since it shows *temporal overshoot* beyond the end of the vowel, *into the following [l]*; it will be discussed later).

In contrast to this large variety of profiles, the 1st session displayed almost exclusively movements of the type (ii), as shown on Fig. 1.

3.VARIABILITY OF KINEMATIC EVENTS

Inspecting Fig. 3, we must agree with Perkell [11 : 280] in rejecting all three "strong versions" of LA, TL and H models. The protrusion «beginning» (conventionally : minimum value) was not locked at the offset of the unrounded first vowel (LA; contrary to [4] for French), nor fixed relative to the onset of the rounded second one (TL; its peak acceleration neither, thus rejecting H). The only consistent fact through both sessions was that peak protrusion was locked *about* the onset of [y] (with one notable exception, see Fig. 2, lower part; its events are marked by + on Fig. 3). This means simply that no plateau-like and/or spatially overshooted anticipations were observed.

So our data exemplify all the three main types of *profiles*, but they violate all three models with respect to their predicted *dates*.

4.ELICITING TWO STRATEGIES

Such negative results have puzzled students in coarticulation for years. And for our part, we were about to give up and to come to a conclusion about variability *per se*, when we suddenly remembered (*post hoc*!) that we had used innocently two different eliciting techniques to make produce such complex consonant chains as [...kstsk...]. In the 2nd session, instruction was to repeat the sentence, prompted with a long pause : «Ces deux Sixte...sculptèrent», *as a whole*; whereas in the 1st one the subject had to repeat, when prompted, the noun phrase : «Ces deux Sixte», linking up with : «Ces deux Sixte sculptèrent». This possibility to «prime» the action could be compared with a trial approach before jumping the hurdle, allowing to size it up. In our case the effect was a better movement «chunking» (corresponding to prosodic parsing) in this 1st performance, which is visible looking at converging cues, such as overall longer obstruence intervals (Fig. 3) and less elisions of the closure phase for [t] (Fig. 1 vs Fig. 2).

Since it has been recently reemphasized that any lengthening effect on the obstruence interval – such as slow rate, stress and number of consonants – would lead to more complex patterns of protrusion movement, allowing «individual gestures [vocalic *and consonantal* protrusions] to emerge as distinct entities» [6 : 186], it is interesting to note that the greatest stability in movement profiles was obtained in the 1rst session, which brought about the longest and more carefully pronounced tokens. Testing the *proportionality* (for this procedure, cf. [8]) for the different kinematic events, it was found that, within the obstruence interval, maximum velocity was relatively stable in the 1rst session (33% ≤ protrusion lead ≤ 47%), whereas it drifted towards [y] onset in the 2nd one, as obstruence duration increased (from 70% to 4% lead).This last trend was followed by acceleration events, a behavior which corresponds, in the part-whole presentation, to the steepest regression lines on Fig. 3 (lower part).

An interpretation of the proportional behavior in the 1st session could be that *more pausing lengthens rounding anticipation* (see here [7], for evidence in silent pauses, for French and for the same speaker), hence increasing *both* phases : the first one that corresponds to a rather clear realization of [kst≠] (without a full silent pause, of course), with no, slow or just starting protrusion; and the second to the deceleration phase towards peak protrusion.

The noteworthy findings concerning with the 2nd session are that – due to the eliciting technique – pausing was not as easily controlled, which led to unsteadily junctured products. But we have no suggestion to explain the fact that this instability caused the maximum velocity event to draw nearer to [y] onset... which – as noted above – led peak protrusion to occur *right in the middle of the following [l]*, in one extreme case of temporal overshoot.

5.AUDIBILITY & STABILITY

For this acoustically critical case (Fig. 2, lower part) – where, after all, the peak protrusion stays in the domain of the syllable – we checked formant values and found, as early as the first periods of the vowel, maximum energy at about 2000Hz (this energy concentration is normally at about 3000Hz for this and other French speakers for [i] [12]). Looking at between-lips area it appeared that, about at the beginning of the vowel, it was well on the way to reach its minimum value. On the other hand, since this second impulse in narrowing began with the release of the [k], the result was that its burst kept an [i]-like coarticulation. Having not performed, at present time, systematic identification tests, we can only refer to a gating experiment done earlier on French [kstR], etc. [3], where it was found that listeners could hardly identify the following vowel until they were delivered at least half of the last [R] consonant. So whatever the perceptual effects of a possible conflict between the cues of the burst and the vowel may be, it is likely that rounding anticipation is rather *lately* perceived in such cases, at least *auditorily*, if not *visually* [7].

From a *motoric* point of view, the problem remains to explain why during certain poorly audible phases of articulatory movements (under *don't care conditions* in learning gestures-from-sounds) velocity profiles are classically bell-shaped (i) & (ii), whereas during other such phases, they are double peaked (iii) (or even resemble those of a bow (i) !).

We hope to have shown that rate is not the only factor capable of smoothing articulatory trajectories – including poorly audible ones – as it increases. In the case of a complex clustering task (hence «bumpy» in nature in its execution, and this not necessarily because it would prevent an aggregation of idiosyncratic gestures [6]), an improvement in prosodic pausing control can counteract effectively the variability induced by complexity, ultimately preventing poorly audible movements – in spite of their increased sluggishness – to become «bumpy».

• Thanks to S. Maeda for trying to improve our English... and ideas.

6.REFERENCES

[1]ABRY, C., BOË, L.-J. & SCHWARTZ, J.-L. (1990), "Plateaus, catastrophes and the structuring of vowel systems", *Journal of Phonetics*, 17, 47-54.

[2]BAILLY, G., LABOISSIÈRE, R. & SCHWARTZ, J.-L. (1991), "Formant trajectories as audible gestures : an alternative for speech synthesis", *Journal of Phonetics*, 19, 9-23.

[3]BENGUÉREL, A.-P. & ADELMAN, S. (1976), "Perception of coarticulated lip rounding", *Phonetica*, 33, 113-126.

[4]BENGUÉREL, A.-P. & COWAN, H.A. (1974), "Coarticulation of upper lip protrusion in French", *Phonetica*, 30, 41-55.

[5]BENOIT, C. (1986), "Note on the use of correlations in speech timing", *J.A.S.A.*, 80, 1846-1849.

[6]BOYCE, S.E., KRAKOW, R.A., BELL-BERTI, F. & GELFER, C. E. (1990), "Converging sources of evidence for dissecting articulatory movements into core gestures", *Journal of Phonetics*, 18, 173-188.

[7]CATHIARD, M.-A., TIBERGHIEN, G., CIROT-TSEVA, A., LALLOUACHE, M.T. & ESCUDIER, P. (1991), "Visual perception of anticipatory rounding during acoustic pauses : a cross-language study", *Proc. of the XIIth Congress of Phonetic Sciences*, Aix-en-Provence.

[8]GENTNER, D.R. (1987), "Timing of skilled motor performance : tests of the proportional duration model", *Psychological review*, 94, 255-276.

[9]LABOISSIÈRE, R. , SCHWARTZ, J.-L. & BAILLY, G. (1990), "Motor control for speech skills : a connectionist approach", in D.S. Touretzky, J.L. Elman, T.E. Sejnowski & G.E. Hinton (Eds), *Connectionist models, Proc. of the 1990 Summer School*, Morgan Kaufmann Publishers, Palo Alto, CA, 319-327.

[10]LALLOUACHE, M.T. (1990), "Un poste 'visage-parole'. Acquisition et traitement de contours labiaux", *Actes des XIIIèmes Journées d'Études sur la Parole*, Montréal, 282-291.

[11]PERKELL, J.S. (1990), "Testing theories of speech production : implications of some detailed analyses of variable articulatory data", in W.J. Hardcastle & A. Marchal (Eds), *Speech production and speech modelling*, Kluwer Academic Publishers, Dordrecht, Boston, London, 263-288.

[12]SCHWARTZ, J.-L., ABRY, C., BEAUTEMPS, D., ESCUDIER, P., MANTAKAS, M. & PERRIER, P. (1989), "Stratégies interindividuelles de production d'équivalences perceptives", in H. Méloni (Ed.), *Speaker variability and specificity*, Marseille-Luminy, 78-86.

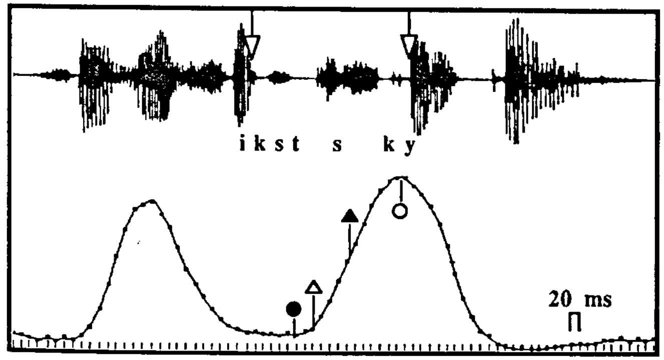


Fig. 1 .– Typical movement profile for upper lip protrusion obtained during the 1st session (cubic spline functions fitted to raw data points; filled circle: min. protrusion; open triangle: max. acceleration; filled triangle: max. velocity; open circle: peak protrusion; white arrows indicate obstruence interval).

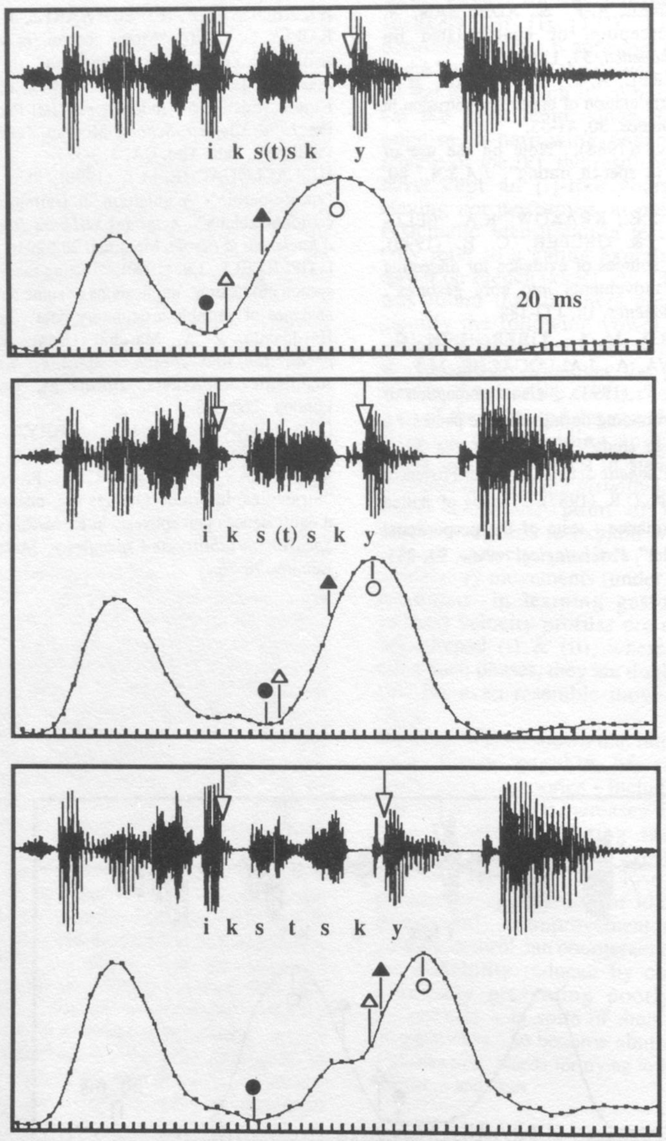


Fig. 2 .– Three types of movement profiles for upper lip protrusion obtained during the 2nd session (cubic spline functions fitted to raw data points; filled circles: min. protrusion; open triangles: max. acceleration; filled triangles: max. velocity; open circles: peak protrusion; white arrows indicate obstruence interval).

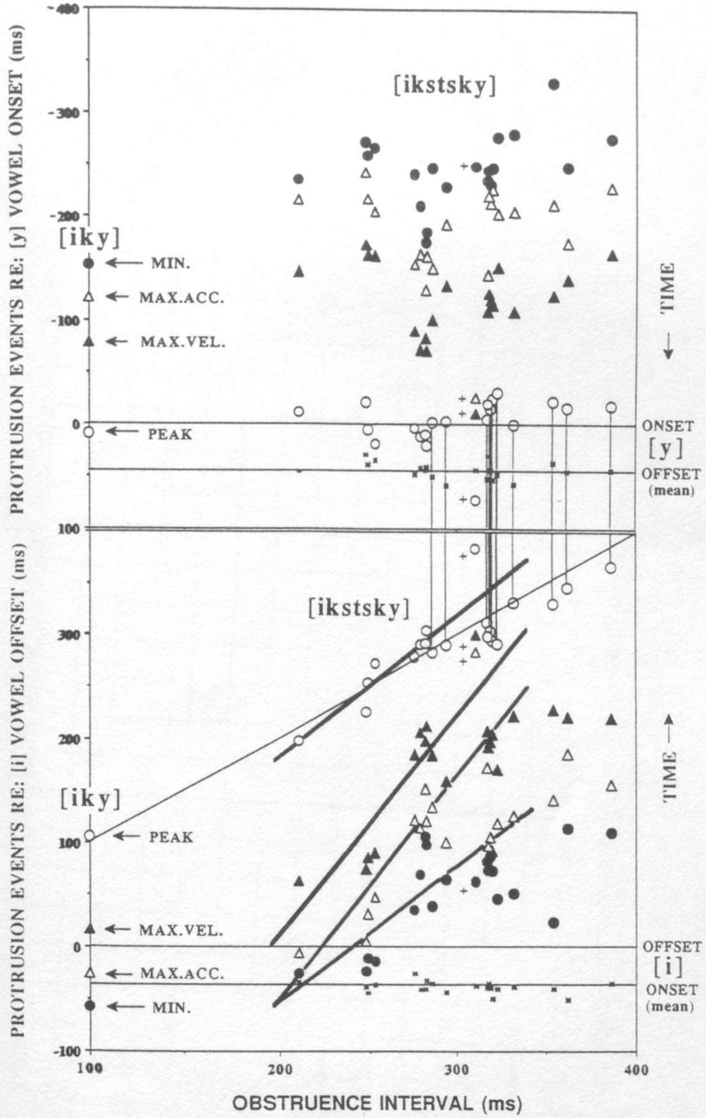


Fig. 3 .– Overlaid plot of kinematic events (same symbols as in Figs 1-2, plus cross-stars indicating [y] offsets and [i] onsets) vs obstruence interval duration. Events are referenced to [y] onset (upper part) or to [i] offset (lower part). Peak protrusion symbols of the 1st session are linked by the vertical lines. Only regression lines significant at p <0.01 are displayed (lower part, thick lines; thin oblique line : y = x) : they correspond to the 2nd session. Symbols prefixed by + are events belonging to the temporally overshooted sample edited on Fig. 2 (lower part). [iky] mean representative "control" events (for both sessions) are actually slightly outside the plot (mean obstruence interval = 96 ms). See text.

SOME OBSERVATIONS ON THE TEMPORAL ORGANISATION AND RHYTHM OF SPEECH

S.G. Nooteboom

Research Institute for Language and Speech, Utrecht University.

ABSTRACT

This paper pleads for quantitative models incorporating many interacting factors controlling the temporal organization of speech, and for testing such models in statistical studies. The paper also warns that such studies tend to obscure real regularities and should be supplemented with classical laboratory experiments.

Isochrony and stress groups are rejected as useful notions, words are proposed as important units. Evidence is shown that within words there is rhythmical alternation of unstressed syllable durations, and that vowel shortening due to change of tempo not necessarily leads to vowel reduction. There is an urgent need for studying the acoustic/phonetic characteristics that distinguish spontaneous from prepared speech.

1. INTRODUCTION

Speech is a slippery phenomenon. Physically speaking, at each moment in time the sound of speech is nothing more than a momentary air pressure perturbation. One moment it is there, the next moment it is gone. The sound of speech has only extension in time as far as we, in our role of listeners, can hold it in memory, or in as far as we, in our role of researchers, can transform it into oscillograms or spectrograms where time is transformed into space. In such registrations we observe rapid discontinuities in intensity and spectral structure, delineating fragments where changes seem to be less rapid.

Such changes or discontinuities in the sound of speech are caused by movements of the sound generating vocal organs. They delineate fragments of speech that can be associated with vowels and consonants realized by the speaker, and perceived by the listener.

Because such fragments with measurable durations can be associated with consonant and vowel realizations, we can observe that realizations of one and the same phoneme can vary tremendously in duration. Durations of realizations of one and the same vowel phoneme may vary from practically zero to many hundreds of ms.

Such variation is not random, but rather rigorously controlled by many factors and their interactions. The result is what we call the temporal organization or temporal patterning of speech.

It is a major task of phonetic research to account for temporal patterns functioning in speech communication. This is not an easy task. Part of the complexity of the problem stems from the fact that there are so many factors involved, on different levels of speech production, and that these factors often strongly interact.

In this presentation I intend to present some opinions, observations, and experimental results that may help to further the ongoing discussion of some, mainly prosodic, aspects of the temporal organization and rhythm of speech.

2. QUANTITATIVE MODELS AND THEIR LIMITATIONS

Let me begin with the following statement:

(1) The systematic effects on speech sound durations of any one particular factor can only reliably be assessed when we take the effects of many other factors into account.

This point is illustrated in Fig. 1, containing some nearly twenty years old data of mine [31]. Here we see the effect of compensatory shortening of the lexically stressed vowel of a word as a function of the number of following unstressed syllables in the word. The data show that this effect strongly interacts with vowel identity, postvocalic consonant and tempo. Interactions are found both in the absolute and in the relative durations.

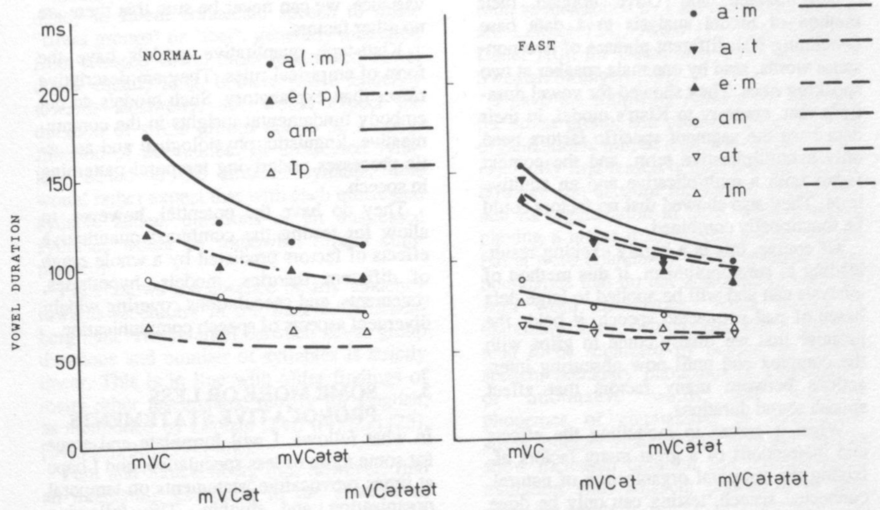


Fig.1. Vowel duration in the initial stressed syllable as a function of number of following unstressed syllables in the word. The parameter is syllable rhyme. Left: normal speech rate, right: fast speech rate (data from Nooteboom, [31]).

Now this is only one example of a great many such interactions that can be demonstrated on the basis of available data in the literature. The existence of such strong interactions can largely explain seemingly contradictory findings by different researchers. An effect that is strong in one speech tempo or one position, may virtually vanish in another speech tempo or another position. As long as we do not take into account such quantitative interactions between different factors, we will not know where to expect an effect of any one particular factor, and how big this effect will be. This remains true despite a recent and interesting demonstration that interactions become less strong, and that some interactions perhaps even vanish, when durational variations of phoneme segments are described in terms of the durational variance of the phoneme type concerned [5].

The upshot of this is that we can only reliably assess the effect of any one particular factor when we take the interactions of this factor with many other factors into account. This leads to my second statement:

(2) There is a real need for quantitative models accounting for speech sound durations, and ways of testing these models.

Interactions of the type demonstrated can be modeled by equations combining additive terms with multiplicative terms. A well known example is the empirical rule proposed by Klatt [21], which in its simple form can be written as:

$$\mathbf{DUR = k(D_{inh} - D_{min}) + D_{min}}$$

in which **DUR** is the segment duration to be calculated, **k** is a parameter describing a context effect, or any combination of such parameters, $\mathbf{D_{inh}}$ is a table value standing for the segment specific inherent duration, and $\mathbf{D_{min}}$ is a table value standing for the segment specific minimal duration.

In this model context parameters provide a multiplicative term, and segment specific parameters provide additive terms. Furthermore, context parameters are functionally combined, under the implicit assumption that the order of the joint effects of these parameters is unaffected by other factors.

Klatt's model was until recently never rigorously tested. It is the merit of Van Santen and Olive [43], that they show how to generalize models of this type mathematically, and how such models can be tested by analyzing the covariances between subarrays of a multifactorial data matrix.

Van Santen and Olive applied their method of model analysis to a data base containing 304 different phrases of two non-sense words, read by one male speaker at two speaking rates. They showed for vowel durations that, contrary to Klatt's model, in their data base the segment specific factors need only a multiplicative term, and the context factor both a multiplicative and an additive term. They also showed that no factors could be functionally combined.

Of course, this is a highly exciting result, leading to some optimism. If this method of analysis can and will be applied to large data bases of real connected speech, it holds the promise that we finally come to grips with the complex and until now obscuring interactions between many factors that affect speech sound durations.

When it comes to modelling the effects and interactions of a great many factors affecting the temporal organization of natural, connected speech, testing can only be done on data obtained in statistical studies based on extensive corpuses of connected speech. There are already a number of such studies available in the current literature. Examples are Barnwell [2], Harris and Umeda [19], Umeda [41], Crystal and House [6],[7],[8], Fant and Kruckenberg [14],[15], Fant, Nord and Kruckenberg [16],[17], and Van Santen [42].

Tuning quantitative models to such data bases has not yet been done. It will be exciting to watch the outcome of such an enterprise, and see how far research tools as provided by Van Santen and Olive will bring us. We should be aware of the fact, however, that going back and forth between quantitative models of this kind and statistical data bases has its limitations as a research tool, among other things, for the following reason:

(3) Statistical studies on corpuses of connected speech obscure real regularities: there remains a need for testing specific ideas with well controlled materials in laboratory experiments.

The point is, of course, that factors we have never thought of will not show up in such studies, except in their contribution to the remaining variance. If such factors have big effects, but are not very frequent in the data base, this contribution will be only marginal. Even if the factors investigated account for a high percentage of the overall variance, we can never be sure that there are no other factors.

Klatt-type quantitative models have the form of empirical rules. They are descriptive rather than explanatory. Such models do not embody fundamental insights in the communicative, linguistic, physiological and acoustic processes underlying temporal patterning in speech.

They do have the potential, however, to allow for testing the combined quantitative effects of factors predicted by a whole range of different theories, models, hypotheses, statements, and speculations covering widely divergent aspects of speech communication.

3. SOME MORE OR LESS PROVOCATIVE STATEMENTS

In what follows, I will formulate and argue for some more or less speculative, and I hope at times provocative, statements on temporal organization and rhythm. The following statement concerns the age honoured question of isochrony, and is perhaps less provocative now than it would have been twenty years ago:

(4) There is no tendency towards isochrony in speech production.

The absence of isochrony is illustrated by data for Swedish read aloud text, taken from Fant and Kruckenberg [15] and presented in Fig.2.

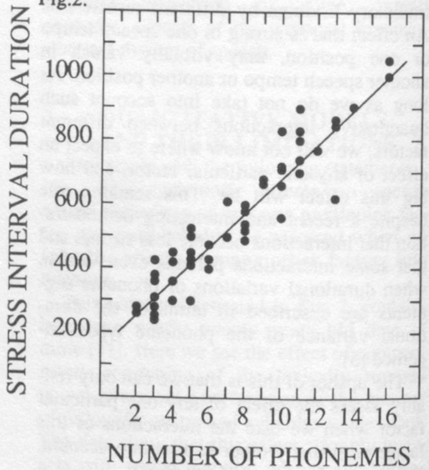


Fig. 2. Free foot duration versus number of phonemes per foot (data from Fant and Kruckenberg, [15]).

Isochrony would mean that there is a tendency in fluent connected speech to make "stress groups" or "feet", generally measured from stressed vowel onset to stressed vowel onset, equally long. If there was a tendency towards isochrony, one would expect the duration of stress groups not to be a linear function of the number of unstressed syllables added to the stressed syllable. One would rather expect that with each unstressed syllable added, durations of all unstressed syllables would be somewhat further compressed.

Fig.2 shows the absence of any tendency towards isochrony in the Fant & Kruckenberg data. The relation between stress group durations and number of syllables is strictly linear. This is in line with older findings of many other researchers, mostly for English, as mentioned by Lehiste [24]: ([4],[22],[23],[34],[35],[36],[39],[40],).

Fant and Kruckenberg did, however, find an interesting exception to the absence of isochrony. They were able to show that the duration of a stress group containing a speech pause is predictable from the number of phonemes in this stress group plus the duration of an average embedded stress group. Apparently, the moment in time a speaker continues after pausing seems to be determined by some rhythmical measure derived from the average time interval between stressed vowel onsets in the preceding stretch of speech.

Of course there remain some questions here. How important is the effect to speech perception and how sensitive are listeners for deviations from predicted durations of pause containing stress groups?

Also one would like to know whether there is any other measure to be derived from the preceding stretch of speech, from which pause duration could be predicted equally well. It is likely, for example, that average word duration is closely correlated with average stress group duration. This leads up to my following statement:

(5) Words are important units for the temporal organization of speech, stress groups are not.

This statement runs counter, for example, to the following statement by Fant and Kruckenberg: "The stress group (....) is a major constituent of durational structure. As an organizational unit of connected speech it overrides the word".

Other proponents of the stress group or foot, as mentioned by Fant and Kruckenberg, are Lehiste [24], Allen [1], Lea [22],[23], Dauer [10], for Dutch Den Os [11], and for Swedish Strangert [38].

I have the following reasons for rejecting the stress group, and promoting the word as organizational unit of temporal organization:

(a) My first reason is this: it is hard to see how we can account for speech production, and its organization in time, without words playing a major role. Speech pauses always fall at word boundaries, never at stress group boundaries that do not accidentally coincide with word boundaries. When speech is extremely slow, words, not stress groups, tend to be separated by pauses. In normal speech, boundary phonemes of emphasized or informative words, not boundary phonemes of emphasized or informative stress groups (whatever that may be), tend to show increased duration and reduced coarticulation with adjacent phonemes of surrounding words.

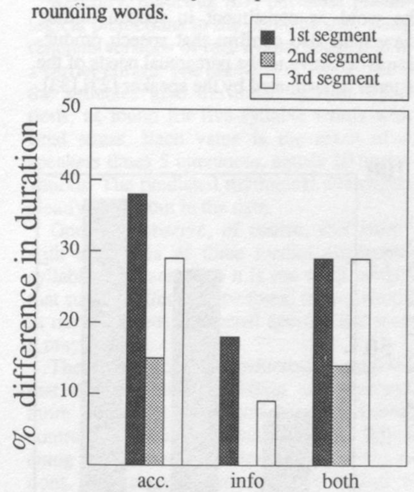


Fig. 3. Percent difference in duration between two realizations of C_1, V, and C_2 in embedded CVC words. Left: the effect of plus versus minus accent (100% is plus accent), middle: the effect of new versus given (100% is new), right: the effect of plus accent new versus minus accent given (100% is plus accent new).(Data after Eefting [13]).

An illustration of the last point is given in Fig.3, containing data from Eefting [13] for

Dutch. She, among other things, compared the temporal structure of within-sentence realizations of the same one-syllable CVC words in three different comparisons: Accented versus unaccented for informative words (i.e. words containing new information), informative versus not informative (new versus given) for unaccented words, and both accented and informative versus unaccented and not informative. The figure plots the relative differences in percent of the accented or informative values, for C_1, V, and C_2.

My interpretation of these data is that the temporal structure of these words is affected by two factors, accent and informativeness. Accent leads to increased vowel duration and some increase in prevocalic and postvocalic consonant durations, informativeness leads to increased durations of word boundary segments, in a tendency to disconnect the word somewhat from its preceding and following context. Apart from providing evidence for the word as constituent in speech timing, these data also confirm that speech production is sensitive to the perceptual needs of the listener as estimated by the speaker [27],[33].

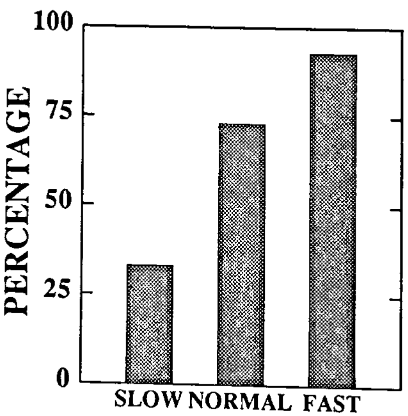


Fig. 4. Percent cases in which the clusters /p#b/ and /t#d/, containing a word boundary, are perceived as /b/ and /d/ respectively, due to assimilation and degemination, as a function of tempo. The deletion of /t/ or /d/ in each case gave an existing alternative word before the boundary (data from Menert, [24]).

Fig.4 gives another illustration of the same point. The figure shows data from Menert [29] who experimentally studied the frequency of perceptual ambiguity resulting from assimilation and degemination in /td/ and /pb/ clusters across word boundaries as a function of speech rate. Word boundaries coincided with potential phonological phrase boundaries. Such data reflect at the same time the relevance of phonological phrases and the relevance of words as constituents of timing in speech. Similar effects for stress groups have yet to be shown.

(b) Secondly, Beckman and Edwards [3], in a carefully controlled experiment examining relations between vowel duration and prosodic constituency, found that there are two different prosodic boundary effects, phrase-final lengthening and word-final lengthening. The word-final effect could not be explained in terms of isochronous intervals of some sort.

(c) Thirdly, Van Santen [42], in his earlier mentioned statistical study on a data base derived from 2,262 sentences read aloud by a single male speaker, finds considerable effects on durations of both stressed and unstressed vowels of the number of syllables and position of the stressed syllable for words, but fails to find similar effects for stress groups. This study is exceptional, because the author explicitly compares words and stress groups as potential constituents of temporal structure. In most other comparable studies, either words or stress groups are chosen as units and we can not judge which of the two explains most of the variance in the data, or whether one of the two would be superfluous.

(d) My fourth and last reason for rejecting the stress group is one of economy. We should not introduce more units than necessary to account for our data. The question is of course, whether there are durational data concerning normal fluent speech that cannot be explained without recourse to stress groups. Perhaps there are. But I know of no publications where that is convincingly shown. As long as that is the case, we should hesitate to accept the stress group as a necessary control factor of the temporal organization of speech.

Let me add two remarks here. One is that the fact that some phenomena can, pragmatically, be easily described in terms of stress groups is not sufficient evidence that stress

groups are part of the mental control of durational variation in speech. The other is that I do not wish to deny that some phenomena in speech appear to be controlled by a rhythmic principle, as exemplified by the rhythm rule in accent structures. There may be similar phenomena in durational control. But these can be accounted for without recourse to the beginnings and endings of stress groups. In this sense stress groups are dispensable, words are not.

Consequently, my next statement is concerned with the temporal pattern of words, and runs as follows:

(6) Within-word sequences of medial unstressed syllables follow a pattern of rhythmic alternation,'short'- 'long'-'short, etc.

Syllable duration depends among many other things on lexical stress, and on position in the word. There is an interesting difference between observations by phoneticians, and predictions made by word level phonologists. This difference particularly concerns temporal patterns in sequences of unstressed syllables. An example in case is shown in fig. 5, showing the abstract rhythmical pattern of a five-syllable Dutch word, with final stress. The example is taken from Kager [20].

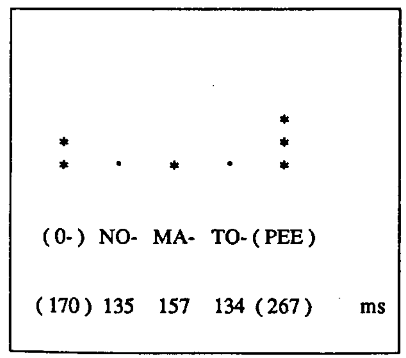


Fig. 5. Stress pattern of the Dutch word **onomatopee** *according to Kager [20], and vowel durations as measured in reiterant versions of such words by Slootweg [37].*

The number of stars above each vowel represent the relative prominence level of the syllable. Kager predicts that there is a hierarchy of reducibility, in which syllables with the lowest stress level, having no stars, are most easily reducible to schwa, etc. This agrees with intuitions about reducibility. He also predicts that in nonreduced realizations, the syllable durations follow an underlying pattern represented in the columns of stars. Similar predictions are made for English.

The pattern shown correctly predicts of course, that the lexically stressed syllable is most prominent, and has the longest syllable duration. Also the prediction that the unstressed initial syllable is somewhat longer than medial unstressed syllables is in agreement with phonetic measurements and earlier proposed empirical rules [32]. The rhythmical alternation, however, in the sequence of three unstressed medial syllables, following a 'short'-'long'-'short' pattern, is not part of commonly proposed empirical rules for temporal patterning, such as proposed for Dutch by Nooteboom [32] for Swedish by Lindblom and Rapp [25], and for English by Klatt [21].

Recently, Slootweg [37] put these phonological predictions to the phonetic test, using reiterant versions of real words, embedded in a carrier phrase. The numbers in Fig. 5 below the syllables give the mean syllable durations, as found for five-syllable words with final stress. Each value is the mean of 4 speakers times 5 utterances, equals 20 observations. The predicted rhythmical alternation clearly stands out in the data.

One may observe, of course, that words with sequences of three medial unstressed syllables are rare. Also it is not at all certain that such an effect can be found in real words in normal fluent connected speech. But then it may.

The connection with reducibility suggests that this and similar effects will become more prominent as speech becomes faster, contrary for example to compensatory shortening of stressed syllables. These observations also suggest that there is indeed a rhythmical component to speech production, albeit different from the kind that would lead to isochrony of stress groups. Note also that this kind of rhythmical pattern has no recourse to units that are orthogonal to linguistic structure, such as stress groups. The relation between vowel duration and spectral vowel reduction is the topic of my next statement:

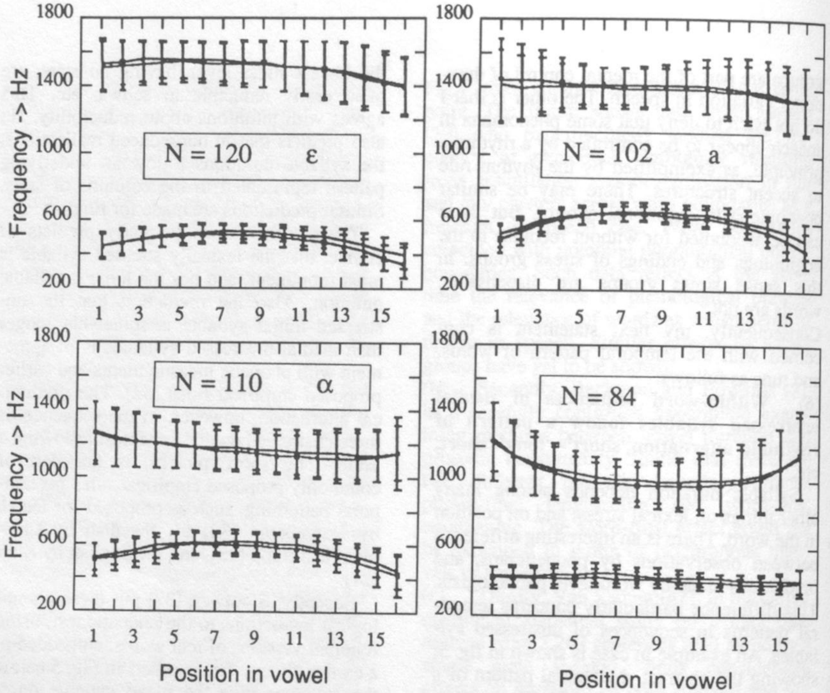


Fig. 6. Mean value and standard deviation for first (F_1) and second (F_2) formant frequency of fast and normal rate vowels. Vowels as indicated (data from Van Son and Pols [44]).

(7) Other things being equal, vowel shortening due to a higher speech tempo, does not lead to vowel reduction.

More than a quarter of a century ago, Lindblom [26] proposed the "Target-Undershoot" model, predicting that, when durations of a certain vowel get shorter and shorter, the articulatory movement keeps farther away from the vowel target, and there will be increasingly more spectral reduction. Essential to this model is that the articulatory movements do not move faster when vowel duration is decreased.

Although Lindblom confirmed his prediction with some simple nonsense utterances, ever since attempts to find further confirmation have led to controversial results [45]. Of course, in testing the model, it is important that speed of articulatory movement and target value are not different due to other factors than vowel duration alone.

Some recent results on this issue are exemplified in Fig. 6, taken from Van Son and Pols [44]. They showed in an acoustic study that if everything is kept equal except speaking rate, the speed of articulatory movement is adapted to vowel duration and there is no spectral reduction.

Their data were obtained from two versions of the same text, once read at a normal speaking rate and once at a fast speaking rate by a single highly experienced professional speaker. They paired normal and fast realizations of the same vowels in the text only when prosodic conditions were the same. Fig. 6 shows average tracks of the first two formants each time of the same vowel, sampled every millisecond, after normalization of vowel duration. Obviously, there is no vowel reduction at all. And if after normalization of vowel duration the formant movements are identical, then before normalization they were of course different, being slower in slow speech and faster in fast speech.

Of course we should be aware of the fact that in cases where durational differences depend on differences in stress level, or on position, the relation between duration and articulatory movements reflected in speed of formant changes can be quite different. Macchi, Spiegel, and Wallace [28] convincingly

showed that the effect of position in the word on vowel duration leaves formant transitions virtually intact, and lengthens and shortens specifically the steady-state portions of vowels. It is yet to be assessed how important such phenomena are for perception. A recent study by Drullman and Collier [12] showed that using reduced instead of full diphones in appropriate positions in synthetic Dutch speech, does not improve speech quality as soon as segment durations are optimized.

Admittedly, the data of Van Son and Pols shown in Fig.6 are based on text read aloud by a highly competent professional speaker. We do not know whether perhaps other speakers do behave according to Lindblom's predictions, or whether perhaps vowel reduction as a function of speech tempo does not occur in prepared speech but does in spontaneous speech.

Unfortunately, this is true of practically all effects discussed. We know preciously little about temporal organization and rhythm of spontaneous speech. It is a source of constant amazement to me that when I turn on the radio or television, and hear someone speak, I seem to need only a few syllables to determine whether I listen to spontaneous speech or prepared speech. The acoustic-phonetic correlates of this difference are unknown. There is room here for the following plea:

(8) There is an urgent need for studying the acoustic/phonetic characteristics that distinguish spontaneous speech from prepared speech.

4. CONCLUSION

I conclude this presentation with the following remarks. Somewhat oversimplifying the situation, we can say that research on temporal organization and rhythm in speech is either descriptive, or directed towards understanding the mechanisms underlying observable timing in speech.

The descriptive type of research seems to hold the promise that we may account for the combined effects of many factors of widely different origins in connected prepared speech. Such an account of necessity will have the form of empirical rules that can be very useful for speech synthesis and perhaps for speech recognition. But this approach will not tell us where the many effects and their interactions come from. It will not satisfy our scientific curiosity, nor will it lead to the detection of completely new phenomena, phenomena we have never thought of before.

In the research directed at the underlying mechanisms of timing in speech often only a single aspect or a few aspects of timing control are studied within the same theoretical framework. This may be unavoidable, given the highly complex and multifaceted nature of speech, but it is also unsatisfactory. There is, according to Osamu Fujimura, a need for "an integrated understanding of linguistic and behavioral as well as physiological and pathological processes involved in speech production" [18], and, I like to add, speech perception. Such integrated understanding will not come fast. But it will not come at all, if we do not make a conscious effort to bring together insights from different areas, and study how the predicted effects interact in actual speech production and perception.

The descriptive approach proposed by Van Santen and Olive [43] can then perhaps offer a thorough testing ground for our predictions.

5. REFERENCES

[1] Allen, G.D. (1975): "Speech rhythm: its relation to performance universals and articulatory timing", Journal of Phonetics 3, pp.75-86.

[2] Barnwell, T.P. (1971): "An algorithm for segment durations in a reading machine context", MIT-RLE Technical Report No 479.

[3] Beckman, M.E. and Edwards,J. (1987): "Lengthenings and shortenings and the nature of prosodic constituency", Paper presented at the First Laboratory Phonology Conference, Columbus, Ohio.

[4] Bolinger, D.L. (1965): "Pitch accent and sentence rhythm". In: Forms of English: Accent, Morpheme, Order, p.163ff. Cambridge Mass., Harvard Unive. Press.

[5] Campbell, N. (1990): "Evidence for a syllable-based model of soeech timing". Proceedings of the First International Congress on the Processing of Spoken Language, Acoustic Society of japan, Kobe, pp. 9-12.

6] Crystal, T.H. and House, A.S. (1982): "Segmental durations in connected-speech signals: Preliminary results".

Journal of the acoustical Society of America 72, 705-716.

[7] Crystal, T.H. and House, A.S.(1988a): "Segmental durations in connected-speech signals: Current results". Journal of the acoustical Society of America 83, 1553-1573.

[8] Crystal, T.H. and House, A.S.(1988b): "Segmental durations in connected-speech signals: Syllabic stress". Journal of the acoustical Society of America 83, 1574-1585.

[9] Crystal, T.H. and House, A.S. (1989): "Articulation rate and the duration of syllables and stress groups in connected speech", unpublished manuscript, Institute for Defense Analysis, Princeton.

[10] Dauer, R.M. (1983): ""Stress-timing and syllable-timing reanalyzed", Journal of Phonetics 11, pp.51-62.

[11] Den Os, E. (1988): "Rhythm and tempo of Dutch and Italian", Doctoral dissertation, Utrecht.

[12] Drullman, R. and Collier, R. (1990) "On the combined use of full and reduced diphones in speech synthesis", unpublished manuscript, Institute for Perception Research, Eindhoven.

[13] Eefting, W. (1990): "The effect of "information value" and "accentuation" on the duration of Dutch words, syllables, and segments", unpublished manuscript, accepted for publication. Institute for Language and Speech, Utrecht University.

[14] Fant, G. and Kruckenberg, A. (1988): "Some durational correlates of Swedish prosody", In: Proceedings of the Seventh FASE Symposium, Vol.2 (SPEECH '88), Edinburgh.

[15] Fant, G. and Kruckenberg, A. (1989): "Preliminaries to the study of Swedish prose reading style", Speech Transmission Laboratory, Quarterly Progress Report No 2/1989, pp. 1-83.

[16] Fant, G., Nord, L., and Kruckenberg, A. (1986): "Individual variations in text reading. A data-bank pilot study", Speech Transmission Laboratory, Quarterly progress 4/1986, pp. 1-7.

[17] Fant, G., Nord, L. and Kruckenberg, A. (1987): "Segmental and prosodic variabilities in connected speech. An applied data-bank study", Proceedings of XIth International Congress of Phonetic Sciences, Vol. 6, Estonian Academy of Sciences, Tallinn, USSR, pp. 102- 105.

[18] Fujimura, O. (1989): "Articulatory Perspectives of Speech Organization", to be published in the Proceedings of the Bonas Conference on Speech Production.

[19] Harris, M.S. and Umeda, N. (1974): "Effects of speaking mode on temporal factors in speech: vowel duration", Journal of the acoustical Society of America 56, 1016- 1018.

[20] Kager, R.W.J. (1989): "A metrical Theory of Stress and Destressing in English and Dutch", doctoral dissertation, Utrecht.

[21] Klatt, D.H. (1976): "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", Journal of the acoustical Society of America 59,(5), 1208-1221.

[22] Lea, W.A. (1974) "Prosodic aids to speech recognition: IV. A general strategy for prosodically-guided speech understanding", Univac Report No. PX10791, St.Paul, Minnesota, Sperry Univac, DSD.

[23] Lea, W.A. (1980) "Prosodic aids to speech recognition", In: "Trends in Speech Recognition", Prentice Hall Inc, London.

[24] Lehiste, I. (1977): "Isochrony reconsidered", Journal of Phonetics, 5, 253-263.

[25] Lindblom, B. and Rapp, K. (1973): "Some temporal regularities in spoken Swedish", Papers from the Institute of Linguistics 21, Stockholm University, Stockholm.

[26] Lindblom, B. (1963) "Spectrographic study of vowel reduction", Journal of the acoustical Society of America 35, 1773-1781.

[27] Lindblom, B.E.F. (1989): "Phonetic invariance and the adaptive nature of speech", In: Working Models of human Perception, edited by B.A.G.Elsendoorn and H.Bouma, Academic Press, London, pp. 139-173.

[28] Macchi, M.J., Spiegel, M.F., and Wallace, K.L. (1989): "Using dynamic time warping duration rules for speech synthesis", unpublished manuscript, Bell Communications Research, Morristown.

[29] Menert, L. (1989): "Perceptual ambiguity as an indicator of voice assimilation", In: "OTS Yearbook 1989", edited by P. Coopmans, B. Schouten, and W. Zonneveld, Research Institute for Language and Speech, Utrecht University.

[30] Nooteboom S.G. and Cohen, A. (1975): "Anticipation in speech production and its implication for perception", In: "Structure and Process in Speech Perception", edited by A.Cohen and S.G.Nooteboom, Springer-Verlag, Berlin, pp.124-142.

[31] Nooteboom, S.G. (1972a): "The interaction of some intrasyllable and extrasyllable factors acting on syllable nucleus duration", Institute for Perception Research Annual Progress Report 7, 30-39.

[32] Nooteboom, S.G. (1972b): "Production and Perception of Vowel Duration, a Study of durational Properties of Vowels in Dutch", Utrecht doctoral dissertation.

[33] Nooteboom, S.G. (1985): "A functional view of prosodic timing in speech", In: "Time, Mind and Behavior", edited by J.A.Michon and J.L.Jackson, Springer-Verlag, Berlin, pp.242-252.

[34] O'Connor, J.D. (1965): "The perception of time intervals", Progress Report 2, Phonetics Laboratory, University College, London, 11-15.

[35] O'Connor, J.D. (1968): "The duration of the foot in relation to the number of component sound segments", Progress Report 3, Phonetics Laboratory, University College, London, 1-6.

[36] Shen, Y. and Peterson G.G. (1962): "Isochronism in English", Occasional Papers 9, Studies in Linguistics, University of Buffalo, 1-36.

[37] Slootweg, A. (1988): "Metrical prominence and syllable duration", In: "Linguistics in the Netherlands", edited by P. Coopmans and A. Hulk, Foris Publications, Dordrecht, pp. 139-148.

[38] Strangert, E. (1985): "Swedish Speech Rhythms in a Cross-Language Perspective", Almqvist and Wiksell Int., Stockholm.

[39] Uldall, E.T. (1971): "Isochronous stresses in R.P.", In: Form and Substance: Phonetic and linguistic papers presented to Eli Fischer-Jörgensen", edited by L.L. Hammerich, R. Jakobson, E. Zwirner, Akademisk Forlag, Copenhagen, pp. 205-210.

[40] Uldall, E.T. (1972): "Relative durations of syllables in two-syllable rhythmic feet in R.P. in connected speech", Work in Progress 5, Edinburgh University Department of Linguistis, 110-111.

[41] Umeda, N. (1975): "Vowel duration in English", Journal of the acoustical Society of America 58, 434-445.

[42] Van Santen, J.P.H. (1989): "Modeling contextual effects on vowel duration. I. Description of individual factors", unpublished manuscript, ATT Bell Telephone Laboratories, Murray Hill.

[43] Van Santen J.P.H. and Olive J.P. (1990): "The analysis of contextual effects on segmental duration", Computer, Speech and Language, 4, 359-390.

[44] Van Son R.J.J.H. and Pols L.C.W. (1989): "Comparing formant movements in fast and normal rate speech", In: "Eurospeech 89, the European Conference on Speech Communication and Technology", edited by J.P.Tubach and J.J.Mariani, CEP Consultants, Edinburgh, Vol 2., pp.665-668.

[45] Van Son, R.J.J.H. (1990): "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", manuscript accepted for publication, Institute for Phonetic Sciences, University of Amsterdam.

COMMENTS ON "SOME OBSERVATIONS ON THE ORGANISATION AND RHYTHM OF SPEECH"

Fredericka Bell-Berti

Haskins Laboratories, New Haven, CT and St. John's University, Jamaica, NY, USA

ABSTRACT

The proposal that the word is the basic organizing unit of speech production is satisfying as well as being a proposal that can be supported by a substantial body of data. These comments review some of the supportive data and also raise questions about the origins of utterance-level speech-timing effects.

1. INTRODUCTION

The notion of the word as the basic organizing unit of speech production is both intriguing and intuitively satisfying. A good measure of this satisfaction derives, I think, from the idea that the primary organizational unit of speech may be the same for all languages, regardless of their status as "syllable-timed" or "stress-timed," or unspecified on that dimension. Or, in another light, since speech tempo is usually specified as a measure of syllables per unit of time, if "stress-timing" assumes that the constant duration intervals occur between stressed syllables, then, for some languages at least, "the durations of utterances are determined by syllable count, but not all syllables count" [26].

However, before embracing a model having the word as the basic unit of speech production, it would be useful to be able to answer what, on the surface, appears to be a very simple question: "What is a word?" The most widely offered definition of "word" is that it is a string of characters set off by spaces, and the orthographic conventions for representing words are language-specific. What is not considered in this definition (or attempt at definition) is the nature of a "word" for an illiterate speaker, for a listener to a foreign language, or for a child first learning to speak. For these groups of speech users and hearers, the "words" may be quite different entities than they are for accomplished literate speaker/hearers of a language. (To make this point explicit: consider the many anecdotes of the sort in which a string of segments intended to convey one set of "words" is heard as another: e.g., "pullet surprises" for "Pulitzer Prizes.") And although lack of a satisfactory definition of "word" seriously limits our ability to confirm or reject this model immediately, the study necessary to answer questions abut the nature the word ought to provide evidence useful in evaluating this model. That is, identifying the nature of the differences (if any) among the "words" of these groups of speakers should provide the linguistic framework necessary for evaluating models of speech production in which "words" play a critical organizing role.

2. SOME RELEVANT STUDIES

Dauer's [11] evidence that their is no more isochrony in English and Thai (both stress-timed languages) than in Spanish (a syllable-timed language) or Italian and Greek (both unclassified on this dimension) offers compelling support for rejection of the stress group as the temporal organizing unit of speech. It will be necessary, though, to reconcile Umeda's [28] data with the word as the organizing unit. Her data do not provide evidence that the number of syllables in a word influences the durations of the vowels in the word; rather, she found that vowel durations could be predicted from a number of phonological conditions, including whether a vowel occurred pre-pausally, whether it occurred in a stressed syllable, and whether it occurred word-finally. Thus, Umeda's data suggest that, other things being equal, syllable number is not very important--instead, one syllable is much like another. This result contrasts with that of Lehiste [22], who found a reduction in syllable duration as suffixes were added to monosyllabic stems. (We must, of course, consider the possible effects of the different tasks on the experimental outcomes: in Umeda's study, *vowel* durations were measured in extended, continuous--i.e., 10-20 minute--speech samples, whereas in Lehiste's study, *syllable* durations were measured in much shorter speech samples.) It is possible to interpret Lehiste's data in terms of the word as the organizing unit: within the word, durations are determined by phonological conditions including the number of syllables. This leads, of course, to questions of the syllable as the organizing unit of speech. However, questions of speech timing and speech rhythm necessarily require units larger than the syllable, since questions of *relative* segment duration mandate comparisons among parts of larger pieces of speech.

Returning, then to the proposal that the word is the temporal organizing unit of speech, it may be instructive to examine some physiological and acoustic data collected for other purposes, to see if and how they may support this model, or how the model must be modified in order to account for these data. To begin, such a model must account for observed differences in the organization of speech gestures in utterances produced at different speaking rates, including differences in the relative magnitude and timing relations of the articulatory gestures for successive segments. There are studies [13, 29] whose data are in conflict with the target-undershoot model [23], suggesting that although increasing speech tempo results in vowels of shorter duration, it does not result in spectrally reduced vowels. This result is taken as support for the word as the temporal organizing unit, and suggests that there is a reorganization of the word's articulatory patterns to achieve the same spectral targets as occur in slower speech. These results, however, also suggest a homogeneity among speakers in the way that they achieve changes in speech tempo, a homogeneity that is contradicted by Harris' [16] data, in which one of her three subjects showed overshoot and the other two showed undershoot of vowels in speech produced with increased tempo. In addition, the genioglossus muscle electromyographic (EMG) data accompanying Harris' acoustic data support the notion of reorganization of the magnitude of the articulatory gestures, and not simply their timing (or overlap, [23]). It is also worth noting, I think, that the relation between acoustic overshoot and undershoot and the underlying muscle potentials is not a simple one. That is, one subject's peak EMG activity was reduced for increased tempo utterances, one subject's activity was substantially increased for increased tempo utterances, and the third subject's activity was only modestly increased for increased tempo utterances.

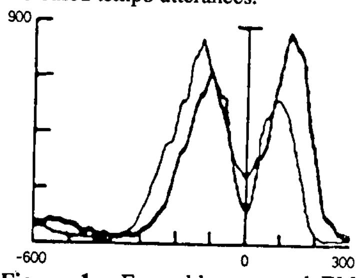


Figure 1. Ensemble-averaged EMG potentials (in.µv) from *genioglossus m.* for 16-20 tokens each of /əpípipə/ (thin line) and /əpipípə/ (heavy line). Zero, the reference point for signal alignment, represents the beginning of /p/ closure in the acoustic waveform.

A model of speech timing that takes the word, rather than the stress group, as its basic unit of organization can also be supported by the EMG and acoustic data of Bell-Berti and Harris [5]. That study reports on the effect of changing primary stress and speaking rate on the separation of lingual EMG activity associated with the production of two /i/ vowels separated by an intervening /p/ or /b/. Briefly, they reported a direct relation between the duration of the medial stop closure and the depth of the trough between the flanking vowel gestures (Fig. 1). The depth of this trough was taken to be a reflection of the change in relative timing (and hence, overlap) between the end of the pre-stop vowel and the beginning of the post-stop vowel. Similar data on the depth of the trough in labial EMG data for /u/ vowels separated by alveolar consonants have also been reported [8, 14, 15]. This view would have the durations of segmental articulatory gestures determined by the position of the segment within the word, the segmental composition of the word, the location of stressed syllables, and speaking rate, but would have the "edge" relations between the gestures for successive segments remain quite stable across substantial changes in all of these parameters (Fig. 2). That is, the timing of the beginning of a gesture for one segment will be relatively unchanged in relation to that segment's acoustic onset (or, viewed another way, the end of the gesture for the preceding

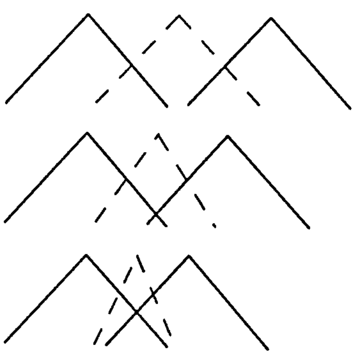


Figure 2. Schematic representation of gestures in a VCV utterance. As the medial (consonant) gesture shortens, the flanking vowel gestures move together, increasing temporal overlap. No changes in gesture magnitude are represented.

segment). This view is supported by the data from a number of studies [5, 6, 8, 9, 15]. That is, these studies and the model they support [1, 7] may be compatible with a model of speech organization having the word as its primary temporal organizing unit.

Even more directly to the point is Krakow's [20] study of the articulatory organization of syllables. She has shown that the position of a nasal consonant within a syllable determines the relative timing of maximum velum and lip displacements, with velum lowering movements generally being enhanced for syllable-final nasal consonants. In addition, however, Krakow has shown small, but stable, scaling effects on lip and velum movements for word-marginal (both syllable- and word-initial or final) nasals, compared with movements for word-medial (but syllable-initial or syllable-final) nasals. That is, although syllable position determines the basic patterns of articulator movements and interarticulator coordination, the position of a segment within a *word* does effect articulatory organization.

3. UTTERANCE-LEVEL PATTERNS

It is also obvious that there are utterance-level effects on segment durations in speech [18]. Thus, we know that segments occurring late in an utterance will be longer than those occurring earlier in an utterance--the final-lengthening effect [e.g., 24, 25]. So other questions we should address are: What are the sources of utterance-level timing patterns? and, How do 'words' fit into the larger units of language?

One possibility is that some aspects of speech timing are determined by the linguistic characteristics of an utterance (the inherent segment durations and the phonetic, semantic, and syntactic context in which a segment occurs), while other aspects of speech timing are determined by the neurological, muscular and mechanical components of the speech system. Bonnot [10] has proposed that there are two levels of speech-timing control, a motor planning level that results in Tatham's [27] "notional time," whose output timing pattern is the result of linguistic and motor programming interactions, and a motor execution level that results in Kent's [17] "clock time."

One utterance-level timing pattern that has been studied widely is the final lengthening effect [24, 25], whose importance in perceiving speech has been suggested by Klatt and Cooper [19], who have shown that listeners expect longer durations for words in phrase- and sentence-final positions. It has seemed reasonable, at least until this point, to assign final-lengthening to the motor-planning level [12]. However, some recent data [2, 3, 4] from studies of acoustic segment durations in French-speaking normal and cerebellar dysarthric subjects seem to shed a different light on the origin of the final-lengthening effect. In those studies, the final lengthening that was a characteristic of the speech of the normal adult speakers was absent in the speech of a group of ataxic dysarthric speakers whose speech was also marked by an overall slowing of speech tempo (measured as overall utterance duration). That is, the reduced speech rate of the ataxic speakers was not simply the result of a global slowing of speech; rather, the durational relations within an utterance were disrupted. These dysarthric speakers suffered from cerebellar disease, and the cerebellum is though to have a "setting" function for motor activity (possibly through muscle spindle biasing [21]). One possible interpretation of these data [2, 3,], then, is that final-lengthening originates at the motor-execution level. Alternatively, however, it may be that there is a limit on how much reduced speaking rate may be and still have final elements that are relatively slower than those occurring earlier in the utterance [4].

4. CONCLUSION

It seems, then, that substantial support can be found in the speech production literature for the word as the basic unit of speech organization. It is also reasonable to assume, though, that, in addition to basic units, we must also identify the larger linguistic units that affect speech timing (e.g., utterance-level effects), as well as the role that physiological systems may play in determining the temporal characteristics of speech. Furthermore, the development of a comprehensive model of speech timing requires that we explore the interactions between the linguistic and physiological systems involved in producing the timing patterns speech.

5. ACKNOWLEDGMENTS

The preparation of this paper was supported by NIH Grants DC-00121 and RR-05596 to the Haskins Laboratories.

6. REFERENCES

[1]BELL-BERTI, F. (1980), "A spatial-temporal model of velopharyngeal function", in N. J. Lass (Ed), *Speech and Language: Advances in basic research practice* (Vol. IV). New York: Academic Press.

[2]BELL-BERTI, F. & CHEVRIE-MULLER, C. (to appear), "Motor levels of speech timing: Evidence from studies of ataxia", in H. F. M. Peters & C. W. Starkweather (Eds.), *Speech Motor Control and Stuttering*. Amsterdam: Elsevier.

[3]BELL-BERTI, F., GELFER, C. E., BOYLE, M., & CHEVRIE-MULLER, C. (1990), "Neurological factors in speech timing", *Journal of the Acoustical Society of America, 88,* S128 (A).

[4]BELL-BERTI, F., GELFER, C. E., BOYLE, M., & CHEVRIE-MULLER, C. (1991), "Speech timing in ataxic dysarthria", *Proceedings of the XIIth International Congress of Phonetic Sciences*, Paper #614 Aix-en-Provence, France.

[5]BELL-BERTI, F., & HARRIS, K. S. (1974), "More on the motor organization of speech gestures", *Haskins Laboratories Status report on Speech Research, SR37/38,* 73-77.

[6]BELL-BERTI, F. & HARRIS, K. S. (1979), "Anticipatory coarticulation: Some implications from a study of lip rounding", *Journal of the Acoustical Society of America, 65,* 1268-1270.

[7]BELL-BERTI, F., & HARRIS, K. S. (1981), "A temporal model of speech production", *Phonetica, 38,* 9-20.

[8]Bell-Berti, F. & Harris, K. S. (1982), "Temporal patterns of coarticulation: Lip rounding", *Journal of the Acoustical Society of America, 71,* 449-454.

[9]BELL-BERTI, F. & KRAKOW, R. A. (submitted), "Anticipatory velar lowering: a coproduction account", *Journal of the Acoustical Society of America.*

[10]BONNOT, J.-F. (1989), "Timing intrinsèque et timing extrinsèque: le temps est-il une variable controlée?", *Journal d'Acoustique, 2,* 287-296.

[11]DAUER, R. M. (1983), "Stress-timing and syllable-timing reanalyzed", *Journal of Phonetics, 11*, 51-62.
[12]EDWARDS, J., BECKMAN, M. E., & FLETCHER, J. (1991). The articulatory kinematics of phrase-final lengthening. *Journal of the Acoustical Society of America, 89*, 369-382.
[13]GAY, T. (1978a), "Effect of speaking rate on vowel formant movements", *Journal of the Acoustical Society of America, 63*, 223-230.
[14]GAY, T. (1978b), "Articulatory units: Segments or Syllables?", in A. Bell and J. B. Hooper (Eds.) *Syllables and segments*. Amsterdam: North-Holland.
[15]GELFER, C. E., BELL-BERTI, F., & HARRIS, K. S. (1989), "Determining the extent of coarticulation: Effects of experimental design", *Journal of the Acoustical Society of America, 86*, 2443-2445 (L).
[16]HARRIS, K. S. (1978), "Vowel duration change and its underlying physiological mechanisms", *Language and Speech, 21*, 354-361.
[17]KENT, R. D. (1983), "The segmental organization of speech", in P. F. MacNeilage (Ed.), *The Production of Speech*, New York: Springer-Verlag, pp. 57-89.
[18]KLATT, D. (1976), "The linguistic uses of segment duration in English: Acoustic and perceptual evidence", *Journal of the Acoustical Society of America, 59*, 1208-1221.
[19]KLATT, D., & COOPER, W. E. (1975), "Perception of segment durations in sentence contexts", in A. Cohen and S. Nooteboom (Eds.), *Structure and Process in Speech Production.* Heidelberg: Springer Verlag.
[20]KRAKOW, R. A. (1989), "*The articulatory organization of syllables: A kinematic analysis of labial and velar gestures*", Unpublished Ph. D. dissertation, Yale University, New Haven, CT.
[21]LARSON, C. R., & SUTTON, D. (1978), "Effects of cerebellar lesions on monkey jaw-force control: Implications for understanding ataxic dysarthria", *Journal of Speech and Hearing Research, 21*, 309-323.
[22]LEHISTE, I. (1972), "The timing of utterances and linguistic boundaries", *Journal of the Acoustical Society of America, 51*, 2018-2024.
[23]LINDBLOM, B. E. F. (1963), "Spectrographic study of vowel reduction", *Journal of the Acoustical Society of America, 35*, 1773-1781.
[24]LINDBLOM, B. E. F., LYBERG, B., & HOLMGREN, K. (1981), *"Durational patterns of Swedish phonology: Do they reflect short-term motor memory processes?"*, Bloomington, Indiana: Indiana University Linguistics Club.
[25]LINDBLOM, B. E. F., & RAPP, K. (1973), "Some temporal regularities in spoken Swedish", *Papers in Linguistics, University of Stockholm, 21*, Stockholm.
[26]LISKER, L. (1976), "Phonetic aspects of time and timing", *Haskins Laboratories Status report on Speech Research, SR47*, 113-120.
[27]TATHAM, M. A. A. (1970), "A speech production model for synthesis-by-rule", *Ohio State University Working Papers in Linguistics, 6*. (Cited in R. D. Kent, "The segmental organization of speech", in P. F. MacNeilage (Ed.), (1983), *The Production of Speech*, New York: Springer-Verlag, pp. 57-89.
[28]UMEDA, N. (1975), "Vowel duration in American English", *Journal of the Acoustical Society of America, 58*, 434-445.
[29]VAN SON, R. J. J. H., & POLS, L. C. W. (1989), "Comparing formant movements in fast and normal rate speech", in J. P. Tubach and J. J. Mariani (Eds.), *Eurospeech 89, the European Conference on Speech Communication and Technology*, CEP Consultants, Edinburgh, Volume 2, pp. 665-668. (Cited in Nooteboom, 1991).

DURATION MODELS IN USE

R. Carlson

Dept. of Speech Communication and Music Acoustics, RIT, Stockholm, Sweden
Currently at: Spoken Language System group -LCS, MIT, Cambridge, Mass, USA

ABSTRACT
The main point in this paper is to describe how duration models actually are in use. Most obviously we find them applied in text-to-speech systems. We also find that such models are slowly introduced in speech understanding systems. We will also discuss the notion of local speech tempo and the need to connect linguistic factors to low-level models. We will also discuss speaker-dependent parameters such as vowel-consonant ratio.

1. INTRODUCTION
The paper by Sieb Nooteboom discusses several topics that have proven to be of importance for duration modelling. The difference between descriptive models and explanatory models is made clear. Furthermore, the need for studies using large speech corpora is emphasized. At the same time the author issues a warning that important details can be lost in these studies. Thus, they must be complemented by selective studies of specially collected data.

In this discussion paper we will elaborate a little more on some aspects of duration modelling that have not been completely covered by the author. We will especially argue that the picture is not that pessimistic as the reader of the paper might think. Many aspects of duration have been studied and duration models have been formulated. These models have been used in synthesis systems and also in recognition systems with some success.

2. KLATT DURATION MODEL
The work by Klatt has had much importance for the development of duration models. The notions "inherent duration" and "minimal duration" have been used by many researchers. The Klatt duration model [8] has become a standard as will be seen in this discussion paper.

At the time the Klatt model was presented, it was also perceptually evaluated in a synthesis experiment [5]. It was shown that the model actually predicted durations of equal naturalness as durations taken from a reference speaker "DK." It also performed better than a model completely based on the isochrony concept. However, the model creates a duration framework with some degree of isochrony anyway. This is mainly a result of the stress-dependent rules and cluster shortening rules.

Klatt's model was adjusted for Swedish by Carlson and Granstrom [3]. Special rules had to be formulated in order to cover the V:C/VC: variation in Swedish. The resulting rule system was tested against a Swedish speech corpus based on one speaker. The standard deviation for phoneme duration was 34 ms. The difference between measured and predicted duration had a standard deviation of 20 ms.

Testing of duration models against speech corpora is an important part of the evaluation process. When comparing the model predictions with the actual data, we found that some "well known" facts needed adjustments. The shortening rule of vowels before unvoiced stops turned out to have some restrictions. Only in stressed position could we find evidence for this rule.

3. SOME DURATION MODELS USED IN RECOGNITION

One of the first ambitious efforts to study duration in a large speech corpora was conducted by Pitrelli [12][13]. As a starting point, the Klatt model was tested against the Timit database [9]. The result was compared to a model based on a hierarchical structure. The statistical model had a better performance than the rule-based model. A total of 630 sentences spoken by 127 speakers were used in the evaluation. The statistical model was able to describe 60% of the vowel duration variance and 55% of the consonant duration variance. The resulting variance was 31 ms for vowels and 26 ms for consonants.

The model was also used as part of a recognition system. In a pilot experiment, a reduction of the error rate by 2 to 3 percent could be shown. The system initially had an error rate of around 15 percent.

Similar efforts to include complex duration models as part of recognition systems have have been made by other researchers. Riley and Ljolje [14] report a method to create a regression tree that takes input from a phone recognizer. The system was trained and tested on the special Darpa resource management task [11]. The standard deviation in the residual in the prediction of phone durations was 29 ms which compares to the overall 45 ms standard deviation of the phones themselves. No adjustment for speech rate was pursued. The improvement to the recognition was only minor with the duration model included.

These examples illustrate how duration models in speech recognition have started to attract interest. However the methods ,so far, have only made small contributions to improved performance. We will later discuss some reasons for this.

4. CORPORA-DERIVED MODELS FOR SPEECH SYNTHESIS

The dominating methods to predict duration in speech synthesis have been based on rule-driven models. However, statistical approaches have also been used. In a sequence of papers, the ATR group has reported results of duration modelling based on statistical analysis of speech corpora [16]. They achieve similar results compared to the earlier-mentioned studies. A reduction of the standard deviation from 33 ms to 21 ms has been reported.

The model developed by Pitrelli was also used to predict phone durations in a text-to-speech system. In a small listening test, the performance was shown to be comparable to the the output of the original Klatt rules.

Campbell [2] has shown that a neural network can be trained to perform as well as the Klatt rules. Other experiments based on statistical analysis have been reported from the CSTR group [1].

5. DURATION IS RELATIVE

We have discussed duration models based on rules or statistically derived models. It is interesting to note that in all these studies the phoneme durations have a standard deviation of around 40 ms. After some kind of model is applied we typically get an error with a standard deviation of 25 ms. What is the reason for this general result? It seems to be the same irrespective of approach.

We can find one possible answer in how local speech tempo is modelled, or rather disregarded. In most approaches it is assumed that the speech tempo is constant during a sentence or clause. It is also assumed that stress has a limited number of levels. A syllable can either be stressed, reduced or unstressed. These simplifications create significant problems. When comparing the duration prediction to natural speech we often find that the prediction error is a function of time [4]. This can be interpreted as a tempo change inside the phrase or the sentence. To some extent this has already been modelled by the introduction of lengthening rules for final phonemes in words, phrases and clauses. However, the rules are not taking into account the type or the function of the syllable, word or phrase. A prefix, root or suffix probably follows slightly different duration rules. A noun phrase probably follows slightly different rules compared to a prepositional phrase.

In a special study by the ATR group [7], parts of speech were included. It could be shown that the segment duration is correlated to parts of speech. The classical difference between function words and content words was clearly manifested in the results. Pronouns and auxiliary verbs were shorter than nouns and adjectives. Ordinary verbs tended to form an in-between class. Despite the striking result, it might be argued that the parts of speech label is not the primary factor for this correlation. Rather, the use of the words in different syntactic positions is the real cause. The formation of phonological words might be a helpful method in this context.

It is interesting to note that the verbs can be prosodically associated to either the preceding or the following words. Depending on the association we will get a final lengthening and a prosodic marking of one phrase boundary or the other.

The use of such duration cues has recently been tested in the context of a speech understanding system [10]. A special break index was designed to encode the possible decoupling between words. With the help of acoustic analysis this index could be predicted. This break index made it possible to significantly reduce the number of possible syntactic parses.

It is clear that the duration cues will play an important role in the future to guide the natural language processing in speech understanding systems. In a complementary manner, we can get advice on how to approach duration modelling from the natural language processing community. It is known that the distribution of possible word sequences is different depending on the syntactic function [15]. Intuitively the distribution of pronouns is a good example of this uneven spread.

6. SPEAKER-DEPENDENT DURATION

Several parameters in a duration model are speaker dependent. Speech tempo and vowel/consonant ratio are two such variables. To illustrate this point we did an analysis of the two sentences spoken by all 600 speakers in the Timit database [9]. In Figure 1 the total vowel duration divided by the sentence duration for these two sentences are plotted for each speaker. This ratio for the two sentences are clearly correlated. One explanation could be that a slower speech tempo usually is realized by an increased vowel duration rather than consonant duration. Plotting the data as a function of speech tempo did not support this hypothesis.

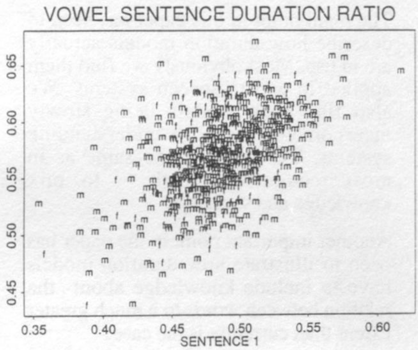


Figure 1. Vowel/sentence duration ratio for 600 speakers. Each mark represents one speaker's data for two sentences plotted along the x and y axes.

During evaluation of our duration models [4], it has become clear that the models of a speaker have to fit together. Naturally both the intonation and the duration model are closely related. However, it is also important to note that acoustic parameters like spectral shape and vocal tract dynamics in general must model the same speaker. We find in our synthesis work that it is not always possible to impose the duration structure from one speaker on a synthesis model with other parameters from another speaker.

7. CONCLUDING REMARKS

Based on the discussion above, we would thus like to modify the following two comments made in the invited lecture: - Klatt's model was until recently never rigorously tested. - Tuning quantitative models to databases has not been done.

It has been our goal to show that attempts to do such evaluations and tunings actually have been taking place.

We would like to support the comment regarding the isochrony question:
- there is no tendency towards isochrony in speech.
In a number of publications, e.g., Fant et. al. [6], it has been shown how a simple framework can correctly predict the duration of a stress interval.

The main point in this paper has been to describe how duration models actually are in use. Most obviously we find them applied in text-to-speech systems. We also find such models being slowly incorporated in speech understanding systems. The trend is the same as in most recognition work -- to mix knowledge and statistics.

Another important point in the paper has been to illustrate how duration models have to include knowledge about the relation between words to a much greater extent than currently is the case.

For a long time the progress in duration modelling has been rather slow. The last years have shown an encouraging new change. The importance of understanding the duration framework is once again starting to be put in focus.

8. REFERENCES

[1] CAMPBELL, W & ISARD, S.D. (1991):" Segment durations in a syllable frame", J. of Phonetics, forthcoming.
[2] CAMPBELL, W (1989): "Syllable-level duration determination", Proc. 'Eurospeech 89', European Conf on Speech Comm & Technology, Paris.
[3] CARLSON, R. & GRANSTROM, B. (1986): "A search for durational rules in a real-speech data base", Phonetica 43.
[4] CARLSON, R. & GRANSTROM, B. (1989): "Modelling duration for different text materials", Proc. 'Eurospeech 89', European Conf on Speech Comm & Technology, Paris.
[5] CARLSON, R., GRANSTROM, B. & KLATT, D.H. (1979): "Some notes on the perception of temporal patterns in speech", in Frontiers of Speech Communication Research, (Lindblom & Ohman, eds.), Academic Press, London.
[6] FANT, G., KRUCKENBERG, A. & NORD, L. (1990):" Studies of prosody and segmentals in text reading", in "Speech Perception, Production and Linguistic Structure", (Tohkura, Sagisaka & Vatikiotis-Bateson, eds), forthcoming.
[7] KAIKI, N., TAKEDA, K. & SAGISAKA, Y (1990):"Statistical analysis for segmental duration rules in Japanese speech synthesis", Proc. Int. Conf. on Spoken Language Processing, Kobe, Japan.
[8] KLATT, D.H. (1979): "Synthesis by rule of segmental durations in English sentences", in Frontiers of Speech Communication Research, (Lindblom & Ohman, eds.), Academic Press, London.
[9] LAMEL, L.F., KASSEL, R.H. & SENEFF, S. (1986):" Speech database development: Design and analysis of the acoustic-phonetic corpus", Proc. DARPA Speech and Natural Language Workshop, Report No SAIC-86/1546.
[10] OSTENDORF, M., PRICE, P., BEAR, J. & WIGHTMAN, C.W. (1990):" The use of relative duration in syntactic disambiguation", Proc. third DARPA Speech and Natural Language Workshop. June1990.
[11] PALLET, D.S. (1987):"Public domain speech recognition database", NBS Report, March.
[12] PITRELLI, J. & ZUE, V.(1989): "A hierarchical model for phoneme duration in American English", Proc. 'Eurospeech 89', European Conf on Speech Comm & Technology, Paris.
[13] PITRELLI, J. (1990): "Hierarchical modeling of phoneme duration: application to speech recognition", Dr theses, MIT, Cambridge, Mass, USA.
[14] RILEY, M. & LJOLJE, A. (1991): "Lexical access with a statistically-derived phonetic network", Proc. fourth DARPA Speech and Natural Language Workshop. Feb 1991.
[15] SENEFF, S. , personal communication.
[16] TAKEDA, K, SAGISAKA, Y & KUWABARA, H (1989): " On sentence-level factors governing segmental duration in Japanese", J. Acoust. Soc. Am. 86(6)

UNITS OF TEMPORAL ORGANIZATION. STRESS GROUPS VERSUS SYLLABLES AND WORDS.

Gunnar Fant

Department of Speech Communication and Music Acoustics, KTH, Box 700 14, S-100 44 STOCKHOLM, SWEDEN

Phone 46 8 790 7872, Fax 46 8 790 7854

1. INTRODUCTION

This is a contribution to the discussion of a keynote paper, "Some observations on the temporal organisation and rhythm of speech", by Sieb Nooteboom for the XIIth International Congress of Phonetic Sciences, 1991, in Aix-en-Provence. Over the years, Nooteboom has made important contributions to these aspects of speech prosody. Several of his main issues I find uncontroversal. We may all agree that we must have full insight into the particular overall contextual frame that may influence observed data. We need reliable, quantitative models accounting for speech sound durations and ways of testing these models.

Nooteboom is somewhat sceptical about statistical studies on corpuses of connected speech, which may obscure real regularities, and he advocates for well controlled laboratory experiments for testing specific ideas. Indeed - but this latter statement could be turned around to claim that there is a real need for insightful interpretations within a linguistic and pragmatic frame of data from large corpuses of connected speech, and that tendencies observed in "lab speech" might not be equally valid for normal text reading. An optimal combination is needed. There is also a need to relate words in connected speech quantitatively to single word utterances. I have a feeling that short "lab speech" sentences occupy an intermediate position which needs to be better understood and modelled in relation to the two extremes.

I shall have reason to expand on Nooteboom's main issue about stress groups versus words. A large part of Nooteboom's paper is devoted to the defence of the word as a basic unit and to express scepticism about the stress group. The controversy outlined by Nooteboom goes beyond the intentions of Fant and Kruckenberg, [4], but it provides him with an incitement to review recent work, some from his own department. This is interesting material per se but merely adds to the established notion of the word as a basic unit, which we do not deny. The controversy appears somewhat superficial. Nooteboom leaves it to us and others to provide evidence in support of the stress group. This will be one of the objects of my review. Nooteboom's discussion of rhythmical properties is limited to within-word structures. I shall provide a broader basis for the discussion of speech rhythm in relation to stress groups and pauses and to temporal units larger than the stress group.

2. THE STRESS GROUP AS A UNIT OF TEMPORAL ORGANIZATION

The stress group or foot is a domain of speech which incorporates one main stress. The boundaries may be defined so that the stress group comprizes a number of complete syllables. This is the case of the metrical foot. However, the most common definition of the domain of a stress group is the interval between two successive stressed vowels. The latter convention is usually adopted for the study of stressed timed languages as Swedish and English in accordance with rhythmical consideration of stressed vowel onsets approximating the so called P-centers, the locations of perceived beats. These are found to be displaced ahead of the vowel onset if preceded by a cluster or an unvoiced consonant, [14, 16]. In the analysis of French prosody the stress group is generally considered to end with the last phoneme of an accented syllable. Martin [15] refers to such stress groups as "prosodic words", which become minimal units in an intonational analysis. We have followed a similar principle in the analysis of French prose reading, and we have found specific patterns of durational increase associated not only with prepause stresses but also with minor stresses inside a clause or a phrase [7, 8]. In order to attain a closer conformity with syntactic units Jassem et al [10] suggest a wider definition of the stress group not restricted to ending or starting with a stressed syllable. The stress group has also played a role in phonological systems, e.g. Selkirk [17].

The stress group is accordingly an accepted unit in phonetics and it remains to evaluate its merits and limitations. Here follows a brief summary:

(1) The basic function of the stress group is to serve as a frame for studying quasi-rhythmical aspects of speech as a sequence and alternation of stressed and unstressed syllables [12].

(2) In connected speech the duration of a stress group not spanning a pause or a region of phrase juncture lengthening increases with the number of phonemes or syllables contained in an approximately linear fashion

$$T_n = a + bn \qquad (1)$$

where b is the average increment per added unit, syllable or phoneme, and a is an offset value which represents the average stress induced lengthening. We may accordingly identify the average duration of an unstressed syllable as b, whilst a + b represents the average duration of a stressed syllable. Apparently the ratio (a + b)/ b is a measure of stressed/unstressed contrast that can be used as a correlate of an individual or general speaking style [4, 5].

(3) A more detailed analysis reveals a weak tendency of isochrony in stress timed languages, e.g. stressed syllable compensatory shortening when the number of following unstressed syllables is increased. This issue has been taken up in several papers to this congress. In our experience the effect is small, in our Swedish data base about 15 ms per added unstressed syllable. Campbell [2] reports somewhat larger values. It is my impression that it is more pronounced in isolated polysyllabic words or short lab sentences than in connected speech. Also it is associated more with the first and second added unstressed syllables than with additional syllables, see e.g. Strangert [18].

(4) A closer approximation to isochrony appears in read poetry and may be studied in terms of stress groups as a supplement to the formal syllable based metrical foot [11].

(5) The stress group is a convenient unit for discussing tempo, i.e. speech rate. The average duration of a phoneme within a stress group is T_n/n, where T_n is the duration and n is the number of phonemes contained. This observed value may be compared to a predicted value of b + a/n from free foot statistics, Eq.1. One aspect of speech rhythm is the alternation of tempo within and between phrases. Some of these variations are predictable from the text in terms of the density of stressed syllables of content words, which shows systematic variations. A text neutral phrase rhythm of decelerations and accelerations may thus be computed as a reference to which adds the speaker's subjective interpretations [8].

(6) An apparent tie exists between mean stress group durations and pause durations. We have observed a tendency of pauses plus associated prepause lengthening to approximate an average interstress interval [4]. This is typically the case of sentence internal pauses of the order of 300 - 500 ms. Pauses between sentences are longer. We have found multi-modal distributions of pause durations, with some additional correction for prepause lengthening, to approximate two or three or four quanta of the order of an average interstress interval. This is found not only for Swedish but also for English and French. A rhythmical coherence of pauses and tempo with obvious analogy to music performance is typical of a relaxed reading of good speakers, and it is also typical of the reading of metrically structured verse. We have evidence that the average interstress interval within a short time memory span of about 4 seconds preceding a pause or something like the last eight free feet

synchronizes an internal beat generating clock which sets a preferred pause duration. This is exemplified by Fig.1 for Swedish sentence internal pauses.

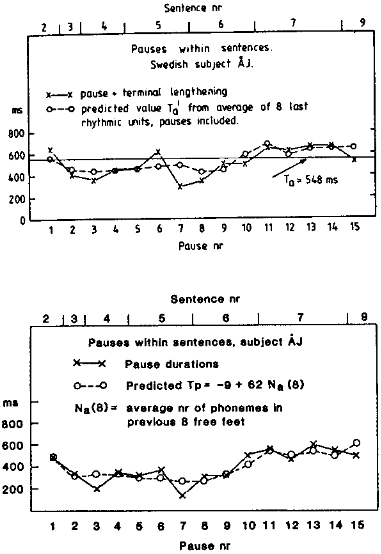


Fig.1. Sentence internal pauses predicted from the last eight free feet, above in terms of mean foot duration, below in terms of average number of phonemes per foot.

Here we have gone one step further and tested the hypothesis that the number of phonemes within a stress group would have the same predictive power as durations. There is a close correlation between number of phonemes per stress group and pause duration. A prediction of pause durations between complete sentences in French is shown in Fig.2.
It is apparent that the eight free feet local reference provides a better prediction than the average foot duration of the whole text. However, it must be stressed that these rhythmical traits are speaker dependent and become upset in conscious efforts to change the overall speech rate or speaking mode. Also, there remains to clarify the underlying perceptual and speech motor mechanisms.

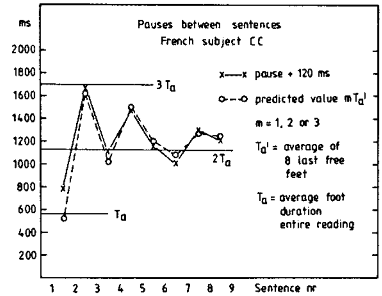


Fig.2. Prediction of pause lengths between sentences in French. The local eight free feet reference provides a better prediction than the long time average foot duration.

3. FINAL REMARKS

Nooteboom's main argument is that much knowledge of speech structure is tied to the word as an organizational unit, whilst this is not the case for stress groups. In my discussion I have stressed the role of the stress group as a unit for structuring rhythmical phenomena and for quantifying stressed/unstressed contrasts. The statement of Fant and Kruckenberg [4] that, in connected speech, the stress group overrides the word is valid in the sense that the stress group as a unit is orthogonal to the word, i.e. it occupies an independent tier of temporal structure above the word level imposing additional constraints. One example is the finding of Bruce [1] that the alternation of weaker and stronger unstressed syllables in Swedish constitutes a rhythmical pattern with a greater consistency within stress groups than within words.
In retrospect, our provocative statement has a wider significance than the role of the stress group. The underlying notion is that of the large differencies frequently encountered between words in isolation and words in context affecting overall duration as well as relative segment durations and patterns. As Nooteboom points out, these depend on a complex of conditioning factors and interactions within and outside a linguistic frame that need to be better understood. The differencies are at times

drastic. The average duration of prepositions is only 20% of the isolated citation form value. Data from "lab speech" experiments are not always representative of the reading of continuous texts. Thus, the recursive segment duration models of Lindblom [13], based on systematically enlarged word and sentence structures, have been quite influential, but they do not seem to have sufficient predictive power for connected text reading. On the other hand, Carlson et al [3] report a significance of stress group alignments in preserving synthesis quality.

At last, a few words about the limitations of stress group statistics. For studies of stressed/unstressed contrasts the regression constants a and b provide a gross measure only. For a more detailed analysis we have to go inside the stress group and perform separate studies of stressed and unstressed syllables and their segmental components [7, 8]. In science we need to use each unit at its best advantage. The stress group is not without interest. It is an established unit.

REFERENCES.

[1] Bruce, G. (1984), "Rhythmic alternation in Swedish", in *Nordic Prosody III*, 31-41, University of Umeå, Almqvist & Wiksell Int.

[2] Campbell, W.N. (1988), "Foot-level shortening in the spoken English corpus", in *Proceedings of FASE 88*, Edinburgh, 489-494.

[3] Carlson, R., Granström, B. and Klatt, D. (1979), "Some notes on the perception of temporal patterns", *Frontiers of Speech Communication Research* (B. Lindblom & S. Öhman, editors), 233-244, Academic Press.

[4] Fant, G. and Kruckenberg, A. (1989), "Preliminaries to the study of Swedish prose reading and reading style", *STL-QPSR 2/1989*, 1-83.

[5] Fant, G., Kruckenberg, A. and Nord, L. (1990), "Acoustic correlates of rhythmical structures in text reading", *Nordic Prosody V*, 70-86, University of Turku, Painosalama Oy.

[6] Fant, G., Kruckenberg, A. and Nord, L. (1989), "Rhythmical structures in text reading. A language contrasting study", *Eurospeech 89*, Vol. 1, 498-501.

[7] Fant, G., Kruckenberg, A. and Nord, L. (forthcoming), "Durational correlates of stress in Swedish, French and English", Proceedings of the Second Seminar on Speech Production, Leeds, May 1990. To be published in *Journal of Phonetics*.

[8] Fant, G. Kruckenberg, A. and Nord, L. (1991), "Temporal organization and rhythm in Swedish", *ICPhS* 1991.

[9] Fant, G., Kruckenberg, A. and Nord, L. (1991), "Language specific patterns of prosodic and segmental structures in Swedish, French and English", *ICPhS* 1991.

[10] Jassem, W., Hill, D.R. and Witten I.H. (1984), "Isochrony in English Speech: its Statistical Validity and Linguistic Relevance", (D. Gibbon & H Richter, eds.) *Intonation, Accent and Rhythm. Studies in Discourse Phonology*. Walter de Gruyter, Berlin, New York.

[11] Kruckenberg, A., Fant, G. and Nord, L. (1991), "Rhythmical structures in poetry reading", *ICPhS* 1991.

[12] Lehiste, I. (1977), "Isochrony reconsidered", *Journal of Phonetics*, 5, 253-263.

[13] Lindblom, B. (1975), "Some temporal regularities in spoken Swedish," (G. Fant & M. Tatham, eds.) *Auditory Analysis and Perception of Speech*, 387-396, Academic Press, London.

[14] Marcus, S.M. (1981), "Acoustic determinants of perceptual center (P-center) location", *Perception & Psychophysics* 30, 3, 247-256.

[15] Martin, Ph. (1982), "Phonetic realisations of prosodic contours in French", *Speech Communication* 1, 283-294.

[16] Rapp, K. (1971), "A study of syllable timing," *STL-QPSR* 4/1971, 14-19.

[17] Selkirk, E.O. (1984), *Phonology and Syntax. The Relation between Sound and Structure*, The MIT Press, Cambridge, MA.

[18] Strangert, E. (1985), *Swedish speech rhythm in a cross-language perspective*, Almqvist & Wiksell International, Stockholm.

TEMPORAL ORGANIZATION AND RHYTHM IN SWEDISH.

Gunnar Fant, Anita Kruckenberg and Lennart Nord

Department of Speech Communication and Music Acoustics, KTH,
Box 700 14, S-100 44 STOCKHOLM, SWEDEN.

Phone 46 8 790 7872, Fax 46 8 790 7854

ABSTRACT

This is a report on a pilot study of speech and pause timing in various modes and speeds of prose reading. We have also performed an analysis of the reading of word lists conforming with the text. The degree of durational reduction in connected speech compared to the isolated words varies with the particular word class and allows an hierarchical ordering of content and function words. Stressed syllables tend to expand more than unstressed syllables in a change from a normal to a distinct reading mode. From the overall statistics of the growth of stressed and unstressed syllables with number of phonemes one can predict a major part of the fluctuation of speech rate within sentences and between phrases. The prediction error represents the reader's deviation from a neutral unengaged reading.

1. INTRODUCTION

In the last few years we have been engaged in studies of prose reading and reading style. These studies have largely been concerned with Swedish. Our major reference for this work is that of Fant and Kruckenberg [3], see also Fant, Nord and Kruckenberg [2] from an early stage of the project with discussions of segmentation techniques, and Fant, Kruckenberg and Nord [4] summarizing how stress foot statistics relate to speech style and rhythmical traits in speech pausing. Recently, the project has been extended to incorporate a language contrasting study [5] and a separate contribution to this congress, [6]. Another extension of our work is to poetry reading, [7].

Studies of the acoustic realization of text reading potentially cover a wide range of problems and methodology related to segmental and suprasegmental structures and the influence of speaker type, text and speaking style. In the present report we shall concentrate on essential differences in durational patterns associated with variations in overall speech rate and distinctiveness. It is well known that a word spoken in the natural context of a sentence may be highly reduced compared to the same word spoken in isolation. We are in a position to provide some quantitative data on normal reduction rates ordered with respect to word class.

The experimental data to be reported here are largely limited to syllable and word durations in text reading. What are the main consequences of a change in speech rate and/or in distinctiveness? How much of the long time speech rate is governed by pauses? How do syllable durations contract and expand at increasing and decreasing speech rate? Do stressed and unstressed syllables behave differently?

A basic problem is how to define speech rate quantitatively. A count of words per minute is not very informative. We need to separate speech from pauses to define an effective speaking time and an average duration of phonemes within sentences or phrases. The local average speech rate varies from one phrase to the next and displays a pattern of quasiperiodical alternations that constitute a higher order rhythmical property of connected speech. We shall attempt to separate the two major factors of the local speech rate, namely that which can be predicted from the particular text and that which has been added by the reader to mark his interpretation and realization of a speaking style.

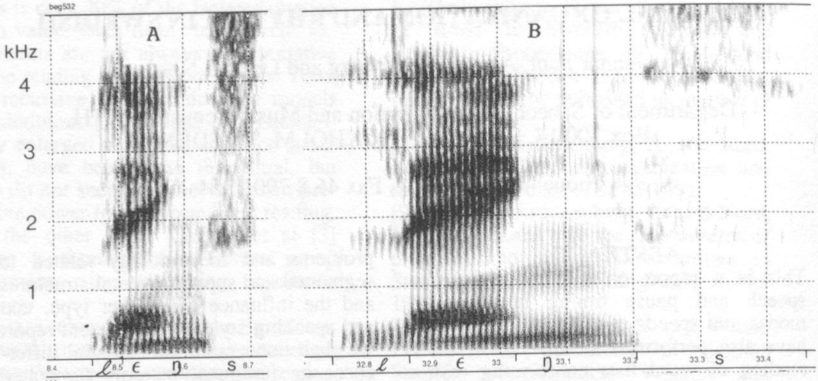


Fig.1. Spectrograms of the Swedish preposition "längs", uttered in a phrase, A, and in isolation (word list), B.

In this paper we shall leave out the lower level aspects of rhythm related to a perceptual average of interstress intervals which we have found to influence speech pauses. This aspect has been extensively treated in our earlier studies, [3], [4].

2. SPEECH - PAUSE TIMING

Our standard text of nine sentences from a novel was read by our reference subject ÅJ, a Swedish language expert employed by the Swedish Radio. He was also the main subject in our earlier studies. The essential data concerning speech and pause durations in a series of four readings representing normal, faster, slower and a distinctive mode of reading are summarized in Table I. A main conclusion, not unexpected, see e.g. Strangert [8], is that the variations in reading mode are associated with substantial variations in overall pause durations. In our data this is combined with rather moderate variations in effective speech time. Thus, in slow reading the total pause time is almost the double of that in fast reading, whilst the effective speaking time and thus mean phoneme durations differ by 11.5% only.

Total pause time within sentences vary relatively more with reading mode than pauses between sentences. This is largely a matter of the number of pauses which increases with decreasing speech rate. In the distinct mode the number of sentence internal pauses was about twice that of normal reading, whilst the average of these pause durations were not much different, of the order of 400 ms. The distinct reading mode displayed the lowest overall speech rate, but this is accomplished with less overall pause duration than in the slow reading and a pause/reading time ratio not much larger than that of normal reading, 30% versus 28%.

3. WORDS SPOKEN IN ISOLATION

It is well-known that words in the natural context of an utterance may vary appreciably in duration compared to words spoken in isolation. The difference may be dramatic such as in highly reduced function words. Thus, in the primary segmentation we often have to assign the /h/ phoneme a zero duration since although heard it may be manifested not by a separate segment but by a subtle modification of a source function only. Short unstressed vowel may occupy one pitch period only or lose voicing in an unvoiced context. On the other extreme a focally emphasized word usually gains a duration close to that when it is spoken in isolation. A typical example of reduction is shown in Fig.1 which pertains to the Swedish preposition "längs", which

Table I. Speech - pause timing in different reading modes

	Normal	Faster	Slower	Distinct
Total reading time (sec)	57.1	51.0	66.8	70.3
Words per minute	130	146	111	106
Pauses, total time (sec)	16.2	12.8	23.9	21.4
Between sentences (sec)	10.6	9.3	14.1	11.5
Within sentences (sec)	5.5	3.5	9.8	9.9
Number within sentences	13	10	18	24
Effective reading time (sec)	41.0	38.2	42.8	48.9
Total pause time as a fraction of total reading time in %	28	25	36	30
Mean phoneme duration in ms (n=547)	75	70	78	89

in context occupies only 28% of the duration when spoken in isolation. Here we also note features such as final lengthening of the /s/ in the isolated form and the phrase initial shortening of the /l/ in the context version.

Fig.2 shows average data of durational reduction as a function of word class. There is a hierarchy headed by adjectives and nouns, which retain about 75% of their isolated reference duration followed by verbs, numerals, adverbs, pronouns, prepositions, auxiliary verbs, and conjunctions down to the extreme of articles that retain on the average 21% only of their isolated mode duration. Content words retain more than 45% of the reference duration and function words less than 45%.

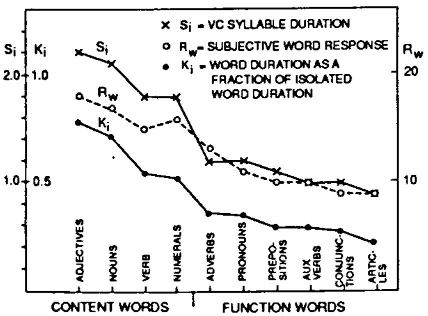


Fig.2. Average data of durational reduction as a function of word class, K_i, VC syllable duration index, S_i, and subjective word response, R_w.

As shown in Fig.2 word classes display the same hierarchical order when represented by a normalized measure of the sum of the duration of the vowel and the following consonant within the maximally stressed syllable. This so called syllable duration index S_i, [3], closely correlates with subjective prominence values derived from continuous perceptual scaling of syllables and words within the same text reading. Also included in Fig.2 are the perceptual estimates of relative word prominence, R_w, which apparently display the same hierarchical order. However, the total span of scale values expressed as ratios comparing adjectives with nouns is different, the R_w being compressed versus the S_i syllable duration index, whilst the degree of reduction versus isolated word form, K_i, displays a larger range. Thus in synthesis by rule of isolated words one should not simply adapt a common expansion factor operating on values typical of connected speech. If so, the function words will be heard as too short.

An observation from the reading of the list of isolated, more precisely separate, words is that of a remarkable isochrony achieved without the aid of a periodic prompter. Average word intervals measured with reference to vowel onsets of stressed syllables, came out close to 2 seconds with some drifts up and down and a standard deviation of 80 ms only within a group of five successive words. There were indications that a system of locating synchrony beats ahead of stressed vowels when preceded by a consonant cluster in accordance with a P-center approach, Browman and Goldstein [1], would have reduced the spread of word intervals.

Another observation from the reading of the word lists is a deviation from linear growth of overall word duration with number of phonemes in the word. As shown in Fig.3 a quadratic regression analysis revealed a mean trend of

$T_n = 240 + 120n - 4n^2$ (1)

where n is the number of phonemes of a word. Apart from the negative quadratic term the coefficients of this regression equation are about twice that found for foot durations of the connected speech text reading.

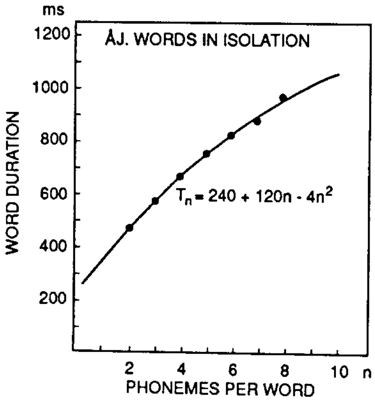


Fig.3. Duration of isolated words (from word lists) versus number of phonemes.

4. SYLLABLE DURATION

The primary aim of the study of syllable durations was to attain measures of basic units to be correlated with varying speech rate and reading style. For this purpose we marked syllables as stressed versus unstressed and whether they had a prepause location to separate out the effect of final lengthening. We shall here review some essential findings only. In a global view there are twice as many unstressed syllables as stressed syllables, whilst the average duration of the unstressed syllables is about one half of that of stressed syllables. This accounts for an approximate balance between stressed and unstressed parts of speech. In the normal text reading stressed syllables averaged 3 phonemes and a duration of 279 ms, whilst unstressed syllables averaged 2.3 phonemes and 127 ms. In comparison we found for the distinct reading mode a mean duration of 319 ms for stressed syllables and 140 ms for the unstressed syllables. Because of the limited text material these data have an uncertainty of the order of 5 ms. With this limitation in mind there is a rather weak significance of a 14% increase of stressed syllable duration in distinct versus normal reading, whilst the difference in unstressed syllables is 10% only. A closer study of the readings revealed that the distinct reading did not lead to a lower speech rate in all parts of the text. There was a similar lack of uniformity comparing normal, slower and faster reading mode. We therefore made a selective analysis contrasting only those words which differed in intended mode. As a result we found for the distinct mode a 22% increase of stressed syllable duration and 11% in unstressed syllable duration compared to normal reading. The corresponding values for slow versus normal reading was 10% and 3% respectively and -5% and -10% for fast versus normal reading. A possible interpretation is that unstressed syllables suffer more than stressed when speech rate is increased securing a stability of stressed syllables, whereas in the slow and distinct modes the stressed syllables are emphasized. This remains to be validated from a larger speech material. However, we may also interpret the results by looking for a ratio of the total duration of stressed syllables versus the total duration of unstressed syllables. Within the selected contrasting material we noted a stressed/unstressed ratio of 1.04 for the normal mode, 1.08 for the fast mode, 1.10 for the slow mode, and 1.14 for the distinct mode.

What systematic variations may we observe inside syllables? According to preliminary data an expansion of a stressed syllable from its normal mode to a more distinct mode generally affects consonants more than vowels, and phonemically long vowels are percentagewise less flexible than short vowels. A relatively greater load of consonants was also found in [3] comparing a distinct speaker with a less distinct speaker. Syllable durations vary systematically with the number of phonemes. Fig.4 provides a regression analysis for normal and distinct reading. There is a clear tendency of linear regression, especially for unstressed syllables which average

$d = -4 + 57.5n$ (2)

ms for normal and

$d = 4 + 61n$ (3)

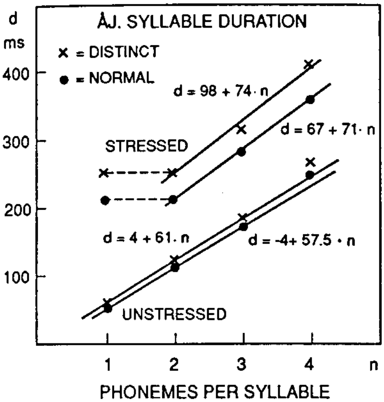


Fig.4. Duration of stressed and unstressed syllables versus number of phonemes in normal and distinct reading.

for the distinct mode. In stressed syllables the single vowels are phonemically and phonetically long and have to be treated separately. In the range of n = 2 - 4 phonemes we noted a regression

$d = 67 + 71n$ (4)

for the normal mode and

$98 + 74n$ (5)

for the distinct mode. Here we observe more clearly the relatively larger distinct versus normal difference in stressed syllables than in unstressed syllable duration.

5. LOCAL SPEECH RATE

When listening to the reading of a paragraph of a text there appears a pattern of alternating accelerations and decelerations of the perceived tempo. These variations occur within the domain of a sentence or a phrase. In order to catch the main variations we divided the complete text of 9 sentences into 26 parts of varying length by segmenting before all pauses and other apparent syntactic boundaries. The size of these units varied considerably, from 0.6 to 2.9 seconds with a mean value of 1.5 seconds, which corresponds to about 9 syllables, three of which stressed. For each of these phrases or complete sentences we calculated a measure of mean phoneme duration. A prediction was next carried out on the basis of the linear regression equations for stressed and unstressed syllables, Eq. 2 - 5. Stress was handled strictly binary, no attempt being made to introduce scalar modifications according to word class. One reason was that some function words were emphasized. We took care in estimating standard values of phrase terminal lengthening, 200 ms for a monosyllabic stressed word at a major phrase boundary, 85 ms for an unstressed syllable before a pause at clause and phrase boundaries inside a sentence, and 50 ms at the end of a complete sentence. Phrase initial shortening was not considered. The prediction was thus essentially based on the number of stressed syllables and unstressed syllables and the specific number of phonemes of each category.

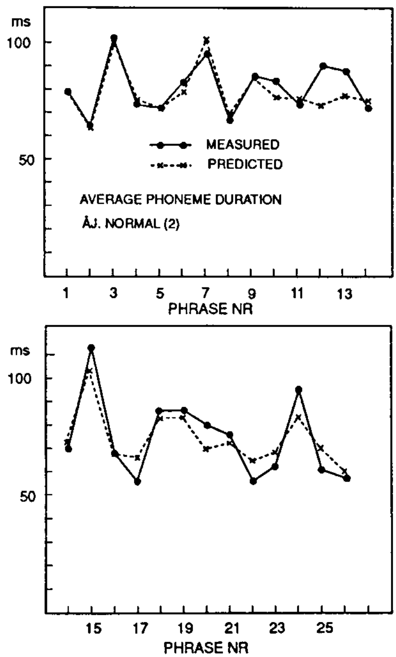


Fig.5. Measured and predicted average phoneme duration in 26 phrases of the main paragraph.

As seen in Fig.5 the prediction of mean phoneme duration within a phrase was successful with an average fit of 6% and occasional rather close matches. One apparent gain in dealing with relatively large units is that differences in phoneme inherent durations average out. A consequence of the overall good fit is that we may separate the two major factors underlying variations in local speech rate. One is a prediction of local speech rate from the text alone. The other major factor is the reader's modulation of the tempo enhancing some parts above the neutral prediction level and undershooting at other places. A grammatical and semantic analysis of the text can explain most of the main deviations. Phrases attaining focal attention by outlining a scene in the story alternate with explanatory and commentary phrases that attain less weight. The overall span of mean phoneme duration is large, ranging from 58 ms to 105 ms. A common pattern within a sentence is that the mean phoneme duration starts low, rises to a peak and decays. In other words, a deceleration followed by an acceleration of local speech rate.

It is remarkable that this tendency to a large extent also prevails in the predicted data, suggesting that essential parts of the local speech rate is determined by the text, e.g. by the relative density of stressed syllables and the occurrence of major clause boundaries. It is significant that the high speech rate of the final phrase is predicted from the fact that none of the eight words was a proper content word. The low local speech rate, indicated by the large peak at phrase 3, is due to the occurrence of two monosyllabic content words, an adjective and a noun, at the end of the phrase.

6. FINAL REMARKS

There remains much to be learned about the manifestation of various reading modes and speech rates, e.g. in the domain of individual phonemes and a contextual frame. There is apparently only a partial correlation between slow speech and distinct speech. We also need further experience from analysis of interstress intervals and their possible relation to the quantification of pauses in the various modes. Much of the analysis reported here is based on the syllable as a unit. A representation in terms of stressed and unstressed syllables has a more effective descriptive power than an analysis in terms of interstress intervals alone. However, there is a close interrelation. We have attempted a prediction of phrase durations as in Fig.5 on the basis of interstress parameters alone, i.e. the a and b parameters of a linear regression of foot durations, $T_n = a + bn$, where b is the increment per added unstressed phoneme in the foot and a the added stress component. The outcome is almost as good as in terms of the syllable based approach.

ACKNOWLEDGEMENTS

These studies have been supported by grants from The Bank of Sweden Tercentenary Foundation, The Swedish Council for Research in the Humanities and Social Sciences and The Swedish Board for Technical Development.

REFERENCES

[1] Browman, C.P. and Goldstein, L. (1988), "Some Notes on Syllable Structure in Articulatory Phonology", *Phonetica 45*, 140-155.

[2] Fant, G., Nord, L., and Kruckenberg, A. (1986), "Individual Variations in Text Reading. A Data-Bank Pilot Study", *STL-QPSR 4/1986*, 1-17.

[3] Fant, G. and Kruckenberg, A. (1989), "Preliminaries to the study of Swedish prose reading and reading style", *STL-QPSR 2/1989*, 1-83.

[4] Fant, G., Kruckenberg, A. and Nord, L. (1990), "Acoustic correlates of rhythmical structures in text reading", *Nordic Prosody V*, 70-86.

[5] Fant, G., Kruckenberg, A. and Nord, L. (forthcoming), "Durational correlates of stress in Swedish, French and English", Proceedings of the Second Seminar on Speech Production, Leeds, May 1990. To be published in *Journal of Phonetics*.

[6] Fant, G., Kruckenberg, A. and Nord, L. (1991), "Language specific patterns of prosodic and segmental structures in Swedish, French and English", *ICPhS* 1991.

[7] Kruckenberg, A., Fant, G. and Nord, L. (1991), "Rhythmical structures in poetry reading", *ICPhS* 1991.

[8] Strangert, E. (1990), "Pauses, syntax and prosody", *Nordic Prosody V*, 294-305.

ISOCHRONY, UNITS OF RHYTHMIC ORGANIZATION AND SPEECH RATE

K. J. Kohler

Institut für Phonetik und digitale Sprachverarbeitung
Kiel, Germany

ABSTRACT

The questions of isochrony, units of rhythmic organization and speech rate are discussed with regard to Sieb Nooteboom's keynote paper in this semi-plenary session.

1. GENERAL REMARKS

There is no doubt that, for an adequate assessment of any temporal effect in speech, we have to take a multiplicity of factors at different levels as well as their interactions into account, that we need to devise and test quantitative models to cope with the data, that statistical analyses of connected speech and of well controlled laboratory experiments complement each other, and that we need a better empirical foundation of the distinction between spontaneous and prepared speech. I also fully subscribe to the importance of phonetic explanation of the mechanisms responsible for speech timing beside the simple description of observable regularities. What I am going to take issue with concerns the topics of isochrony, of units of rhythmic organization and of the relationship between vowel reduction and increased speech rate.

2. ISOCHRONY IN SPEECH PRODUCTION

The strong isochrony hypothesis has been disproved. It has been shown for German [5] that with increasing articulatory complexity and number of syllables in rhythmical sequences of identical nonsense syllables compression to isochronous feet becomes less and less feasible because of the time constraints of articulatory movements; and even if compression is possible it results in the perception of increased spech rate in the case of achieved isochrony. But the lack of compression, i.e. the proportional expansion, also results in a change of overall tempo, this time a decrease. So, in order to stay within the same perceived rate of delivery the speaker has to compress, but this compression must not reach isochrony in this type of logatome syllable chains with unreduced vowels and consonant(s) (clusters).

The use of more natural syllable strings, which not only conformed to the phonotactics of syllables, but also to the rules of syllable chaining in German by selecting reduced vowels in unstressed positions of nonsense words, showed two complementary timing effects [5]:

(a) Disyllabic and monosyllabic feet of the same stressed syllable complexity (vowel quantity, consonant clusters) and within the same speech tempo have duration ranges that tend not to be statistically different, due to a shortening of the stressed syllable before /ə/, whereas polysyllabic feet, although also showing stressed-syllable compression, did not reach complete isochrony.

(b) The comparison of long vs.

short stressed syllables in 2- or 3-syllable feet ("Pahne" vs. "Pinne" or "Pahnige" vs. "Pinnige") yielded a complementary adjustment of the durations of the reduced unstressed vowels.

These data thus support a weak isochrony hypothesis at least for German, and other so-called stress-timed languages, e.g., English and Dutch, may be supposed to behave likewise. There is a tendency to compress as the number of syllables within the same frames increases, but this compression quickly approaches a ceiling when the number of syllables exceeds two. On the other hand, there is also the tendency to vary the durations of reduced unstressed vowels in opposition to the preceding stressed vowel, being a complementary aspect of a tendency to foot isochrony.

These data can, at first sight, also be referred to the word level because word and foot coincided in these experiments. But why should there be a tendency to make mono- and bisyllabic words of the same length by stressed-syllable compression and unstressed syllable compensation? There is nothing in the linguistic category of the word that could determine such a behaviour, whereas a superimposed rhythmic principle can easily explain it and a number of other phenomena:

(1) The ordering in German "mit Pfeil und Bogen" as against English "with bow and arrow" is not semantically, but rhythmically conditioned: the mono- rather than the disyllabic noun is put before the conjuction to get a more even sequence of foot durations than would be the case with the reversed order, and this grouping cuts across word boundaries.

(2) Articulatory reduction is at work irrespective of words and word boundaries [6]. Words may disappear altogether, and they may be treated as syllabic appendices to preceding words, even bridging phrase structure boundaries, e.g. in "Hast du einen Moment Zeit?" (*Have you got a moment to spare?*) [ˈhaspm momɛn ˈtsaɪt]. The deletion of [ə] in [dənən], derived from "du einen", follows the [ə] elision in "die geschnittenen Rosen" (*the cut roses*), although there is a phrasal boundary between "hast du" and "einen Moment". The reduction can go further to [ˈhas məmɛn ˈtsaɪt], where "du" and "einen" have disappeared from the phonetic surface, and the prestress syllable [mo] is also reduced. Finally, [ˈhas mmɛn ˈtsaɪt] can result, showing a further reduction of the unstressed part of the content word "Moment". All these processes are in keeping with the rhythmic principle to make feet as equal in duration as possible. Function words are obvious candidates to assist in this compression because they are unstressed in the unmarked case and signal redundantly coded syntactic functions rather than lexical meaning, but unstressed syllables of content words undergo the same reductions.

(3) In verse, the rhythmic principle is regularised as in / Humpty / Dumpty / sat on a / wall / / Humpty / Dumpty / had a great / fall / / All the king's / horses and / all the king's / men / / Couldn't put / Humpty to/gether a/gain / This is only possible because there is an underlying rhythmic principle in speech, which triggers the tendency towards isochrony independent of the chaining of words.

Nooteboom refers to the Swedish data from read speech in Fant and Kruckenberg [2] in support of his dismissal of isochrony as a factor in speech production. But continuous texts, i.e. accidental corpus rather than systematic experimental materials can neither prove nor disprove such a rhythmic principle because in connected speech a great number of timing factors operate, and they may easily override tendencies towards isochrony, as was demonstrated in Kohler [4]. The reference to Fant and Kruckenberg actually runs counter to Nooteboom's statement that "statistical studies on corpuses of connected speech obscure real regularities: there remains a need for testing specific ideas with well controlled materials in laboratory experiments". It was precisely this methodological prerequisite that determined the experimental design followed in the Kiel studies of speech timing, which devised language materials and data collection procedures in a hypothesis-driven fashion to systematically test and possibly reject the isochrony assumption [5]. But the reference is also at odds with Nooteboom's statement that "the systematic effects on speech sound durations of anyone particular factor can only reliably be assessed when we take the effects of many other factors, on different levels of speech processing, into account". In continuous, ad hoc texts the many different factors and their interactions cannot be reliably separated.

3. WORDS OR STRESS GROUPS AS UNITS OF TEMPORAL ORGANISATION?

Nooteboom provides a very clear and categorical answer to this question: "Words are important units for temporal organization of speech, stress groups are not." And he adduces four reasons:

(a) Speech pauses always occur at word boundaries, never at stress group boundaries that do not accidentally coincide with word boundaries, and boundary phonemes of emphasized or informative words, not boundary phonemes of emphasized or informative stress groups tend to show increased duration and reduced coarticulation.

It is very important to define what is meant by pauses: pauses for syntactic and semantic structuring are usually placed at word boundaries, hesitation pauses can be anywhere in the syllable chain, and pauses for emphasis can also be inside words before the stressed syllable, as in "I warn you: don't for ... get." (compare this with "I warn you: don't for-fucking-get."). Even if it could be argued that in this case the pause or an inserted swear word occurs after a stripped-off prefix and therefore at a linguistic not a rhythmic boundary, examples such as "po ... tato and to ... mato" refute this as well. And in verse a slowing down can produce pauses at any stress group boundary irrespective of word boundaries. Thus "Humpty Dumpty" can be read with pauses at all the foot boundaries including "to/gether" and "a/gain". Instead of having pauses at foot boundaries the boundary phonemes in all the examples quoted may be lengthened and detached from their environment for special emphasis. It is the stressed syllable that gets the extra prominence either in order to highlight the word it is in for semantic reasons, or the foot it is in, for rhythmic reasons.

(b) Beside phrase-final lengthening, there is word-final lengthening, which could not be explained in terms of isochronous intervals.

This finding does not justify the exclusion of the stress group as a timing unit. Of course, it is possible to have word-induced duration control, as was shown for German "eine gezeigt" ([*has*] *shown one*) vs. "einige zeigen" ([*will*] *show some*) in Kohler [3, 4]. This is due to the content structuring of speech, but its occurrence does not cancel out rhythmic structuring; on the contrary, the latter can obliterate word-related timing as the same investigations demonstrate.

(c) Nooteboom refers to unpublished data by van Santen that show considerable duration effects of the number of syllables and the stressed-syllable position in words, but not in stress groups.

The data in Kohler [3, 4] point to effects in both units.

(d) A principle of economy compels us not to introduce more units than are necessary to account for our data, and there are no publications where it has been convincingly shown that data cannot be explained without recourse to stress groups.

First of all, economy is certainly a useful criterion in phonetic data <u>description</u>, but when it comes to <u>explaining</u> the observed phenomena we are bound by what can give us the deepest insight into the widest possible empirical domain, and economy is secondary to this consideration because why should speech timing, or any other phonetic event in human production and perception processes, be entirely governed by economy. Secondly, the occurrence of greater compression in German "eine gezeigt" ([*has*] *shown one*), as against "eine zeigen" ([*will*] *show one*), in some recorded data [3], points to a timing factor that cannot be equated with the word as the only relevant rhythmical unit.

Furthermore, van Dommelen has shown [1] that in a falling, as against a level, F0 contour, combined with vowels from a duration continuum spanning the quantity opposition /a/ vs. /a:/ in German, the perceptual quantity switch occurs at greater vowel durations, provided the syllables are embedded in a rhythmic sequence, irrespective of word boundaries, e.g. "Er hat As [ˈas] / Aas [ˈa:s], Assen [ˈasən] / aßen [ˈa:sən], Masse [ˈmasə] / Maße [ˈma:sə] verstanden." (*He understood ace/carrion, aces (dat.) / (they) ate, mass/ measurements.*) This result is replicated when the stimuli of the disyllabic word pairs are presented in isolation, but it is, in this position, reversed to shorter vowel durations for the monosyllabic pair, in keeping with Lehiste's findings [7]. What is important here is simply the presence or absence of a following unstressed syllable without reference to word division. A possible explanation for these opposed effects can be sought in a perceptual syllable lengthening in falling F0, which changes the rhythmic patterning of syllable sequences and thus the speech rate; the quantity assessment of the physical vowel duration then occurs against a slower tempo frame and therefore appears shorter. If there is no rhythmic frame surrounding the test syllable, especially no following unstressed syllables, there is no independent tempo assessment, and the perceived lengthening affects the vowel directly.

The answer to the question of rhythmical units in speech should not be an "either or", but a "both and". Words are certainly important units for the temporal organisation of speech, but stress groups are as well, and the two interact. In verse the rhythmic principle dominates, in continuous connected, spontaneous speech the word (content) aspect gets more prevalent, but the rhythmic principle never disappears. Just as words have to be put into a segmental frame, so they also have to be fitted into a rhythmic one, and both segments and timing are affected by the content structuring of utterances.

4. VOWEL REDUCTION AND INCREASED SPEECH RATE

Nooteboom rules out that vowel shortening due to a higher speech tempo leads to vowel reduction, by referring to data in van Son and Pols [9]. But these data were obtained from a "single highly experienced professional speaker", and I therefore think that such a categorical exclusion of vowel reduction in increased speech rate of spontaneous speech is unjustified. What is essential here is that, given the need of speakers to be understood under different speech production conditions, phonetic variation, including vowel spectra in different tempo frames, can be located along a hyper-hypo scale to guarantee sufficient discriminability with as little effort as is necessary in the particular communicative situation [8]. So, speakers can execute precise movements to reach targets irrespective of speech rate if they put in the necessary effort to achieve increased discriminability for listeners, but they may also slur if they think the effort is not worthwhile, and it is the latter attitude that eventually results in language change.

5. REFERENCES

[1] VAN DOMMELEN, W. A. (1991), "F0 and the perception of duration", *Proc. XIIth Intern. Congr. of Phonetic Sciences*

[2] FANT, G. and KRUCKENBERG, A. (1989), "Preliminaries to the study of Swedish prose reading style", *STL*, *QPR*,2, 1-83.

[3] KOHLER, K. J. (1983), "Prosodic boundary signals in German", *Phonetica*, 40, 89-134.

[4] KOHLER, K. J. (1984), "Temporal control at the utterance level in German"; in VAN DEN BROECKE and COHEN (eds.) "*Proc. of the Xth Inter. Congr. of Phonetic Sciences*"; pp. 197-200, Dordrecht: Foris Publications.

[5] KOHLER, K. J. (1986), "Invariance and variability in speech timing: from utterance to segment in German", in PERKELL and KLATT (eds.) "*Invariance and Variability in Speech Processes*"; pp. 268-289, Hillsdale, New Jersey, Lawrence Erlbaum.

[6] KOHLER, K. J. (1990), "Segmental reduction in connected speech in German: phonological facts and phonetic explanations"; in HARDCASTLE and MARCHAL (eds.) "*Speech Production and Speech Modelling*", pp. 69-92, Dordrecht/Boston/London: Kluwer Academic Publishers.

[7] LEHISTE, I. (1976), "Influence of fundamental frequency on the perception of duration", *Journal of Phonetics*, 4, 113-117.

[8] LINDBLOM, B. (1990), "Explaining phonetic variation: a sketch of the H & H theory", in HARDCASTLE and MARCHAL (eds.) "*Speech Production and Speech Modelling*", pp. 403-439, Dordrecht/Boston/London: Kluwer Academic Publishers.

[9] VAN SON, R. J. J. H. and POLS, L. C. W. (1989), "Comparing formant movements in fast and normal rate speech", in TUBACH and MARIANI (eds.) "*Eurospeech 89*", *vol. 2*, pp. 665-668 Edinburgh: CEP Consultants.

PROSODY IN SITUATIONS OF COMMUNICATION: SALIENCE AND SEGMENTATION

Anne Cutler

MRC Applied Psychology Unit, Cambridge, UK.

ABSTRACT

Speakers and listeners have a shared goal: to communicate. The processes of speech perception and of speech production interact in many ways under the constraints of this communicative goal; such interaction is as characteristic of prosodic processing as of the processing of other aspects of linguistic structure. Two of the major uses of prosodic information in situations of communication are to encode salience and segmentation, and these themes unite the contributions to the symposium introduced by the present review.

1. INTRODUCTION

Communication is what speech is for. Everything about speech is somehow involved in the relationship between speaker and listener. Is there anything special to say about the role of prosodic structure in this relationship?

One rather negative claim that has shown up in a number of forms is that prosody is in some sense not central to the message being communicated. Among the reasons cited are that prosody encodes affect, which, while it may be communicated, is not part of linguistic structure; or that the dimensions of prosody are duration, intensity and fundamental frequency, and since every speech sound must have some duration, intensity and fundamental frequency, prosody simply falls out of the fact that speech is realised acoustically. The fact that most orthographies do not encode prosody is sometimes seen as supporting evidence for the claim that prosody is inessential.

These days it is presumably unnecessary to argue against this point of view. However, the contributions to the present symposium certainly provide counter-evidence to it. In this introductory review paper, I shall present evidence from studies of speech processing showing that the processing of prosody is subject to the same interacting constraints of the perception and production systems as affect the processing of other aspects of linguistic structure.

2. PROSODY

There is, of course, no one-to-one mapping between form and function in prosody, although for administrative convenience many researchers often act as if there were. Strong correlations certainly exist, for instance between certain kinds of pitch movement and the presence or absence of syntactic closure, but if we know one thing about prosodic function, it is that its relationship to prosodic form is highly complex and to a considerable degree context-dependent.

This symposium is not a theoretical treatment of prosody from either single perspective, however; it is a discussion about prosody in situations of communication. The complexity of the relationship between form and function implies matching complexity in the prosodic processing which speakers and listeners perform in the course of communicating. In the following section I review some of the considerable recent literature on the interaction of perceptual and production processes, with emphasis on the perception and production of prosody.

The complexity of prosodic structure, and the necessity for hierarchical structural descriptions, is a recurring theme also in the five other contributions to the present symposium. In this introductory paper, I have chosen to follow two further themes which run through the symposium: the way prosodic structure can express relative *salience*, and the way it can communicate information about *segmentation*, at various levels of linguistic structure.

3. SITUATIONS OF COMMUNICATION

Let us define a situation of communication for our present purposes as a speaker speaking and a listener listening. (This is not to deny that there are many other kinds of communication, and some of them - sign language, for instance - certainly involve prosody.) The speaker's production processes and the listener's perceptual processes are obviously not independent, if only in the trivial sense in that one operates on the output of the other. However, there are some interesting further aspects of non-independence. Speech production processes can be actively constrained by characteristics of the perceptual process; and such effects can certainly be observed in the processing of prosody.

3.1 Perceptual constraints on production

As I have argued elsewhere [10], speakers' choices in production are often quite obviously constrained by the needs of listeners. This happens even at what one might consider quite low levels. For instance, why are the utterances of a speaker with a pipe clenched between the teeth not incomprehensible? If the processes of production were to run their normal course, the output might be considerably distorted; instead, adjustments occur (see e.g. [26]), with the effect that the processes of perception are enabled to run their normal course. Similarly, consider the Lombard reflex [27]: when ambient noise increases, speakers involuntarily speak more loudly. Interestingly, speakers in this situation adjust the individual formant frequences of their speech to compensate for the spectral characteristics of the noise [31]. The result, once again, is that the output sounds as close to the speaker's normal output as is possible.

At a slightly higher level, we see the same constraints operating on phonological processes of elision and assimilation. The process of palatalisation, whereby an alveolar stop and a following palatal glide become affricated, can apply across a word boundary - thus *did you* becomes [dɪdʒu] - and the effect is obviously to obscure the onset of the post-boundary word. But as Cooper and Paccia-Cooper [6] have shown, palatalisation across a word boundary is significantly less likely if the post-boundary word is unpredictable - for instance, low frequency, or contrastively stressed. The effect of this is that the words which the listener most needs to hear are less likely to be obscured. Likewise, speakers making up nonce words prefer to choose affixes which leave the base word intact over affixes which require stress shifts or vowel changes (so *dowagerish* is preferred to *dowagerial*; [8]); again, the effect is that listeners can make sense of the new word because it contains a known word unaltered within it.

It is unsurprising that effects of this kind are apparent also in the realm of prosody. The mis-stressing of words impairs word recognition most severely if the stress shift causes vowel quality changes [2, 14]; and when speakers make a slip of the tongue involving mis-stressing, they are most likely to correct it if a vowel was changed [9]. Furthermore, they are more likely to add contrastive stress to the correction if there is high contrast between the error and the intended word [25]. Thus both the frequency and urgency of error repair are directly correlated with the likelihood that the error will disrupt comprehension.

Likewise, the work of two contributors to this symposium has shown how well-attuned are the processes of accent placement to listeners' needs. Fowler and Housum [20] showed that deaccented productions of words in a story could function as better retrieval cues for listeners than the same words accented on first mention. We know that listeners hearing a story construct an overall representation of the story situation [3, 22]; Fowler and Housum speculated that deaccenting could function as a signal to listeners that the concept in question is *already in* the story representation. Thus on hearing a word which in the phonetic context is obviously deaccented, listeners automatically access the already-constructed representation; for this reason such words function particularly effectively as retrieval cues. Similarly, Terken and Nooteboom [32] found that true-false decisions could be made more rapidly if new sentence subjects were accented but previously mentioned subjects were deaccented.

3.2 The Role of Speaker Awareness

Speakers' choices when they are deliberately trying to make themselves clear are also well attuned to listeners' needs. When marking word boundaries, for instance with a pause, speakers pay most attention to marking exactly the boundaries which listeners most often overlook, i.e. boundaries before weak syllables [13]. When trying to make syntax explicit, they add syntactic markers such as relative pronouns and complementisers [33], the presence of which, as perceptual research (e.g. [21]) has shown, makes syntactic processing significantly easier.

Prosody can be consciously used by speakers who are trying to be clear. Thus speakers who realise a listener is having difficulty understanding tend to speak more slowly, louder, and with raised pitch [5]. One communication situation in which this very noticeably happens is when an adult is talking to a child. A recent study by Fernald [19] has shown how effectively prosody can be used in this way. Fernald recorded the same mothers talking either to their infant child or to their husband, in specific types of interaction: expressing approval, attracting attention, giving solace etc. She then filtered all the recorded utterances and asked listeners to identify the type of interaction involved in each. The listeners' choices corresponded with the original context to a significantly greater degree for the infant-directed utterances than for the adult-directed utterances. Since the filtering process had left nothing in the speech signal intact except for the prosody, it would seem that, as Fernald concluded, speech to infants is more heavily loaded than speech to adults on prosodic signals of interactive intent.

In most speech situations, however, speakers are not making deliberate efforts to speak clearly. And as

Lehiste showed in a classic study [23], the availability of prosodic cues which will be of use to listeners may depend crucially on speaker awareness of potential problems for the listener. Prosody can in many cases very effectively signal which of two alternative syntactic parses is intended, for instance for syntactic ambiguities such as *The German teachers attended a meeting*, or *She hit the man with the stick* (see, e.g., [30]). In Lehiste's experiment, speakers read out a number of sentences, some of which were syntactically ambiguous; Lehiste then ascertained whether or not the speakers had been aware of the ambiguity, and which interpretation they had intended in their reading. The speakers then produced the sentences twice more, consciously intending each of the two different interpretations. All the versions were then played to listeners, who, Lehiste found, could much more accurately judge which interpretation had been intended in the versions produced with awareness of the ambiguity. Where the speaker had been unaware of the ambiguity, in fact, the listener judgements were often at chance.

3.3 The Speaker-Listener Contract

We can use the term *speaker-listener contract* to signify the proposal that participants in spoken communication have a shared goal: maximising the probability of successful message transmission. As the above review suggests, prosody is as much involved as any other aspect of linguistic structure in speakers' efforts to do their part in achieving this goal. The evidence reviewed included contrastive stress on error corrections; deaccenting of previously mentioned referents; and explicit cues to speech segmentation at the word and the phrase level. Thus both salience and segmentation figure in prosodic contributions to realisation of the speaker-listener contract.

4. SALIENCE

In a language which has sentence accent, listeners accord a high priority to the task of detecting where accent falls in a speaker's utterance. Prosodic cues are exploited to enable listeners to direct attention to the location of sentence accent [7]. If part of the normally available prosodic information is absent, listeners will exploit what remains [15]; but it seems that no one prosodic dimension is paramount in signalling accent location, because conflict between different sources of prosodic information (e.g. rhythm and pitch) leaves listeners unable to predict where accent will occur [11]. The importance of seeking accent location is explained as a search for focussed, or semantically central, aspects of the speaker's message [16].

The processing advantage enjoyed by accented words does not of course imply that if every word in an utterance were to be accented, the listener could process the entire utterance at a faster rate. Salience is necessarily a relative concept. As the work of Fowler and Terken, cited above, has conclusively shown, appropriate deaccenting is just as informative, and just as important, as accent.

In this symposium the contributions of Fowler, Ladd and Terken all make a further contribution to our understanding of the phonology and processing of sentence accent. As Ladd argues, relative salience expresses a syntagmatic relationship (between nodes in a metrical tree, in the metrical notation which Ladd uses), which co-exists with paradigmatic category distinctions between levels of accent (or, in Ladd's terms, levels of sentence stress). Ladd's intention in making this proposal is to reconcile apparently conflicting views of stress: on the one hand, the consensus of contemporary phonologists that stress is an abstract relational construct, and on the other, the paradigmatic approach whereby stress is a property of syllables, which has proven persistently useful to non-phonologists (such as syntacticians and of course psycholinguists).

The role of relational structure in the expression of salience is also central to Terken's contribution, which focusses on the way in which the processes of speech production translate such relational structure into relative acoustic salience (in the fundamental frequency contour, in this instance), and the way in which the processes of speech perception interpret fundamental frequency variation as information about relative salience.

Fowler and Levy extend our understanding of how relative salience in a context finds expression in linguistic output by drawing parallels between lengthening and shortening effects in both prosodic and lexical forms. Unpredictable topics are referred to by longer expressions, and/or the words expressing them are realised with greater duration. The effect is to provide listeners with more speech evidence for less predictable concepts. This is powerful evidence for the operation of the speaker-listener contract at multiple levels of linguistic structure.

5. SEGMENTATION

Segmentation is one of the listener's major tasks; boundaries must be identified between units at several linguistic levels. Firstly, the continuity of the speech signal results in very few reliable cues to word boundaries being realised; listeners therefore have to exploit whatever sources of information they can to work out how speech signal divide up into individual words. Secondly, listeners must group words into phrases, that is, they must detect syntactic boundaries. Thirdly, they must identify larger units of semantic structure, sometimes referred to as topic structure [4], or paragraph structure [24]. And fourthly, they must perceive structure at the interactional level, i.e. speaker turns.

Prosody contributes to the listener's performance of all these segmentation tasks. At the lexical segmentation level, listeners can exploit their linguistic experience to develop heuristic segmentation procedures based on where word boundaries are most likely to occur in their native language; in English, I have argued, such procedures are based on the predominance of strong initial syllables in the vocabulary [12]. At the syntactic level, as was discussed above, prosodic cues to boundaries are readily exploited by listeners [23, 29, 30].

In comparison with the quite large amount of research on lexical juncture, and yet larger body of work on syntactic boundaries, segmentation of discourse into topic or paragraph units has received relatively little attention. (Three studies in the early 1980s should be mentioned: Brown, Currie and Kenworthy [4] reported that speakers tended to raise the pitch of their speech when introducing a new topic; Menn and Boyce [28] reported the same finding in parents' conversations with children. Lehiste [24] analysed the average duration of phonetic segments and words in non-final, phrase-final and paragraph-final position; she found both phrase-final and (somewhat greater) paragraph-final lengthening.) It is therefore timely that the contribution to this symposium by Bruce describes an ongoing project which has as one of its principal aims the investigation of prosodic cues to segmentation at this level of linguistic structure.

Segmentation of conversation into participant turns is, finally, addressed in this symposium by Couper-Kuhlen. The literature on prosodic cues to turn-taking has been bedevilled by confusion between the speaker and listener perspectives; Duncan [18], for instance, isolates several prosodic characteristics of speakers' turn-final uttterances and terms them "cues" without, however, any evidence that listeners actually use them as such (see Cutler and Pearson [17] for a critique). Couper-Kuhlen reports evidence that co-operative rhythmic synchronisation of speech occurs in smooth turn-taking; in this study the listeners' reception of speakers' signals is attested by the synchronisation of the initial rhythmic intervals of the new turn (produced by the listener-turned-speaker) with the final rhythmic intervals of the old turn produced by the previous speaker. Like the other contributors, Couper-Kuhlen also highlights the importance of hierarchical structure in prosody, such structure being fundamental to the turn-taking metric which she proposes.

6. CONCLUSION

It is no surprise to find that salience and segmentation form unifying themes for contributions to a symposium on prosody. According to Bolinger [1], these (or, in his words, obtrusions for prominence and the expression of closure) are the two major language-universal uses of prosody. In situations of communication, much of speakers' and listeners' prosodic processing is devoted to these goals.

One thing to note about the importance of prosodic segmentation cues is that it mirrors the importance of segmentation in orthographic representations - lexical segmentation is explicitly coded in nearly all orthographies, and syntactic segmentation in most; higher-level segmentation is likewise signalled by textual devices. As this review has tried to show, and as the symposium will further stress, both salience and segmentation are central to successful communication, and prosody is thus central to linguistic structure.

7. REFERENCES

[1] Bolinger, D.L. (1978), "Intonation across languages", in J.H. Greenberg (Ed.) *Universals of Human Language*, Stanford: Stanford University Press.

[2] Bond, Z.S. & Small, L.H. (1983), "Voicing, vowel and stress mispronunciations in continuous speech", *Perception & Psychophysics, 34,* 470-474.

[3] Bransford, J. & Franks, J. (1971), "The abstraction of linguistic ideas", *Cognitive Psychology, 3,* 331-350.

[4] Brown, G., Currie, K.L. & Kenworthy, J. (1980), *Questions of Intonation*, London: Croom Helm.

[5] Clark, J.E., Lubker, J. & Hunnicutt, S. (1988), "Some preliminary evidence for phonetic adjustment strategies in communication difficulty", in R. Steele & T. Threadgold (Eds.) *Language Topics: Essays in Honour of Michael Halliday*. Amsterdam: J. Benjamins.

[6] Cooper, W.E. & Paccia-Cooper, J. (1980), *Syntax and Speech,* Cambridge, MA: Harvard University Press.

[7] Cutler, A. (1976), "Phoneme-monitoring reaction time as a function of preceding intonation contour", *Perception & Psychophysics, 20,* 55-60.

[8] Cutler, A. (1980), "Productivity in word formation", *Papers from the Sixteenth Regional Meeting, Chicago Linguistic Society,* 45-51.

[9] Cutler, A. (1983), "Speakers' conceptions of the functions of prosody", in A. Cutler & D.R. Ladd (Eds.) *Prosody: Models and Measurements,* Heidelberg: Springer.

[10] Cutler, A. (1987), "Speaking for listening", in A. Allport, D.G. MacKay, W. Prinz & E. Scheerer (Eds.) *Language Perception and Production: Relationships between Listening, Speaking, Reading and Writing,* London: Academic Press.

[11] Cutler, A. (1987), "Components of prosodic effects in speech recognition",

Proceedings of the Eleventh International Congress of Phonetic Sciences, Tallinn, Estonia; Vol. 1, 84-87.
[12] Cutler, A. (1990), "Exploiting prosodic probabilities in speech segmentation", in G. Altmann (Ed.) *Cognitive Models of Speech Processing,* Cambridge, MA: MIT Press.
[13] Cutler, A. & Butterfield, S. (1990), "Durational cues to word boundaries in clear speech", *Speech Communication, 9,* 485-495.
[14] Cutler, A. & Clifton, C.E. (1984), "The use of prosodic information in word recognition", in H. Bouma & D.G. Bouwhuis (Eds.) *Attention and Performance X: Control of Language Processes,* Hillsdale, NJ: Erlbaum.
[15] Cutler, A. & Darwin, C.J. (1981), "Phoneme-monitoring reaction time and preceding prosody: Effects of stop closure duration and of fundamental frequency", *Perception & Psychophysics, 29,* 217-224.
[16] Cutler, A., and Fodor, J.A. (1979), "Semantic focus and sentence comprehension", *Cognition, 7,* 49-59.
[17] Cutler, A. & Pearson, M. (1985), "On the analysis of prosodic turn-taking cues", in C. Johns-Lewis (Ed.) *Intonation in Discourse,* London: Croom Helm.
[18] Duncan, S. (1973), "Toward a grammar for dyadic conversation", *Semiotica, 9,* 29-47.
[19] Fernald, A. (1989), "Intonation and communicative intent in mothers' speech to infants: Is the melody the message?", *Child Development, 60,* 1497-1510.
[20] Fowler, C.A. & Housum, J. (1987), "Talkers' signalling of 'new' and 'old' words in speech, and listeners' perception and use of the distinction", *Journal of Memory & Language, 26,* 489-504.
[21] Hakes, D.T. (1972), "Effects of reducing complement constructions on sentence comprehension", *Journal of Verbal Learning & Verbal Behavior, 11,* 278-286.
[22] Haviland, S.E. & Clark, H.H. (1974), "What's new? Acquiring new information as a process in comprehension", *Journal of Verbal Learning & Verbal Behavior, 13,* 512-521.
[23] Lehiste, I. (1973), "Phonetic disambiguation of syntactic ambiguity", *Glossa, 7,* 107-122.
[24] Lehiste, I. (1980), "Phonetic characteristics of discourse", *Transactions of the Committee on Speech, Acoustical Society of Japan, 4,* 25-38.
[25] Levelt, W. & Cutler, A. (1983), "Prosodic marking in speech repair", *Journal of Semantics, 2,* 205-217.
[26] Lindblom, B., Lubker, J. & Gay, T. (1979), "Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation", *Journal of Phonetics, 7,* 147-162.
[27] Lombard, E. (1911), "Le signe de l'élévation de la voix", *Annales des Maladies de l'Oreille, du Larynx, du Nez et du Pharynx, 37,* 101-119.
[28] Menn, L. & Boyce, S. (1982), "Fundamental frequency and discourse structure", *Language & Speech, 25,* 341-383.
[29] Scott, D.R. (1982), "Duration as a cue to the perception of a phrase boundary", *Journal of the Acoustical Society of America, 71,* 996-1007.
[30] Streeter, L.A. (1978), "Acoustic determinants of phrase boundary perception", *Journal of the Acoustical Society of America, 64,* 1582-1592.
[31] Summers, W.V., Pisoni, D.B., Bernacki, R., Pedlow, R. & Stokes, M. (1988), "Effects of noise on speech production: Acoustic and perceptual analyses", *Journal of the Acoustical Society of America, 84,* 917-928.
[32] Terken, J. & Nooteboom, S.G. (1987), "Opposite effects of accentuation and deaccentuation on verification latencies for given and new information", *Language & Cognitive Processes, 2,* 145-163.
[33] Valian, V.V. & Wales, R.J. (1976), "What's what: talkers help listeners hear and understand by clarifying syntactic relations", *Cognition, 4,* 115-176.

THE EXPLOITATION OF PITCH IN DIALOGUE

G. Bruce

Department of Linguistics and Phonetics, Lund, Sweden

ABSTRACT

Our study concerns the prosody of spontaneous dialogue centered around the examination of pitch and with exemplification from Swedish. Here we deal with the methodology of this research and also present some results in summary. We have undertaken four types of analysis: analysis of dialogue structure, auditory (prosodic) analysis, acoustic-phonetic analysis, and analysis-by-synthesis. Typically, the same pitch patterns that we have met in read, laboratory speech tend to occur also in spontaneous dialogue. Variation in overall pitch range and its relation to categories of dialogue structure is discussed and found to be a potentially important means for use in the sectioning and development of a dialogue.

1. INTRODUCTION

The present paper reports on and summarizes our current research on the prosody of spontaneous dialogue conducted at the Department of Linguistics and Phonetics at Lund. Our study of dialogue prosody is related to a research project called CONTRASTIVE INTERACTIVE PROSODY ('KIPROS'), which started in 1988 and is supported by the Bank of Sweden Tercentenary Foundation. The object of study is dialogue prosody in a contrastive perspective in French, Greek and Swedish. The ultimate goal of the project is to develop a model for French, Greek and Swedish dialogue prosody. For recent reports from the project work see [8], [9], [10].

Two important general questions that we have been addressing are the following 1) Do we find the same, well-known prosodic patterns in spontaneous dialogue as we have met earlier in read, laboratory speech? 2) How are the prosodic patterns observed related to dialogue structure and interactive categories?

The first question relates to our "old" research tradition in prosody and the general model of prosody we have been developing in Lund ([2], [4], [5], [13], [14]). Our research on prosody in a spontaneous speech framework will give us an indication of how well we have been able to simulate natural prosody in a laboratory speech environment. The second question is related to the "new" research setting for our study of prosody: spontaneous speech and dialogue. What are the factors that govern the specific choice of prosodic patterns for the speakers involved?

In the present report we will deal with the methodology that we have been developing in our study of dialogue prosody and also present in summary some results from our work. The exemplification here will be taken exclusively from Swedish.

2. METHODOLOGY

Our research strategy in the project work has been to study a fairly restricted sample of speech material in relative depth and from different angles. We have been conducting four different kinds of analysis: 1) analysis of the dialogue structure itself without specific reference to prosodic information, 2) auditory analysis in the form of a prosody-oriented transcription, 3) acoustic-phonetic analysis centered around the examination of pitch, and 4) analysis-by-synthesis by the use of text-to-speech.

2.1. Analysis of Dialogue Structure

We have been considering three different aspects of dialogue structure which we have found reason to keep apart in our analysis.

Textual aspects pertain to the development of a dialogue as a text, which may involve one or more speakers. Specifically we are thinking of the division of a dialogue into different 'speech paragraphs', each of which has a certain coherence from the point of view of topic structure. From this point of view the speakers' turns may be characterized as introducing, continuing on or terminating a certain topic.

By *interactive aspects* we refer specifically to the analysis of a dialogue as to how it is carried on in terms of the initiatives (actions) and responses (reactions) taken and given by the speakers involved. This kind of analysis is comparable to a more traditional one into speech act categories such as questions and answers.

Turn taking aspects refer to the specific regulation of the speakers' turns in a dialogue, such as taking, receiving, keeping, and giving away the turn.

2.2. Auditory Analysis

The auditory analysis in terms of a prosodic transcription is kept distinct from the analysis of dialogue structure. Therefore, our prosodic transcription does not contain categories such as question intonation, continuation tone etc. It is only at a later stage, when we are relating the auditory prosodic analysis - as well as the acoustic-phonetic analysis - to the analysis of the structure of the dialogue itself, that we may establish such potential categories.

Basically it is an orthographic transcription of what has been recorded. To this segmental transcription are added prosodic features selected from our model of prosody. While it does not contain potentially very interesting features such as change in speech tempo, loudness and voice quality, our system does encode five prosodic features: accentual prominence, phrasing, pitch range, boundary tones and pausing. Our notation is with one exception fairly broad, and the symbolization is as far as possible in accordance with the new, current IPA system [16].

Prominence. The analysis of prominence levels was made in terms of three binary features: 1) The lowest level of prominence (apart from unstressed), mere stress with no accent, coded [ˌx], 2) A higher level of prominence, accented, coded [ˈx], 3) The highest level of prominence at the phrase or utterance level, focally accented, coded [ˈˈx] .

Phrasing. In the analysis of prosodic phrasing we assume two types of boundaries: a minor phrase boundary for an accentual phrase [|] and a major phrase boundary, corresponding to a division into regular prosodic phrases [||].

Pitch range. Our notation of pitch range represents a fairly narrow phonetic transcription, as this has been in the focus of our attention. Overall pitch range for a major prosodic phrase has been analyzed syntagmatically in relation to the neighbouring phrases and may assume five different values: [→] = same [↗] = slightly raised, [↑] = markedly raised, [↘] = slightly lowered, [↓] = markedly lowered.

Boundary tones. Within a prosodic phrase and for a given pitch range, initial and final boundary tones are judged to be either raised (marked value = [ꜛ]) or non-raised (unmarked). This means that the range of, for example, a final pitch rise, notated as a high boundary tone, can vary considerably but still be transcribed as the same category.

Pausing. In our transcription system we have assumed that where a real pause is perceived, two degrees of pause length are noted: short [(.)], and long [(..)].

Exemplification of our prosodically oriented transcription has been given in earlier reports (cf. [8], [9], [10]).

2.3. Acoustic-Phonetic Analysis

We consider the auditory analysis in terms of a prosody oriented transcription to be a useful basis for the acoustic-phonetic analysis of dialogue prosody: the qualitative and quantitative study of prosodic patterns from acoustic recordings of F0 and speech waveform. Our analysis has been centered around pitch. The standard procedure for us has been to have the recorded material digitalized on the VAX 11/730 at our laboratory and analyzed using the API program of the ILS package, where pitch extraction is done with a modified cepstral technique. A first part of this analysis consists in isolating relevant pitch patterns for accentuation, phrasing, boundary signalling and pitch range, where an intermediary phonological (or abstract phonetic) representation in terms of H(igh) and L(ow) turning points has proved helpful (see e.g. [8]).

2.4. Analysis-by-Synthesis

An important and powerful method in our modelling of dialogue prosody and particularly the exploitation of pitch is analysis-by-synthesis. The research tool which we have been using is the multilingual text-to-speech system developed by Carlson and Granström [12]. The prosody rules of the Swedish text-to-speech system have recently been modified by Bruce & Granström [6], [7]. The idea is to use rule synthesis as a control instrument for checking the adequacy of our model of dialogue prosody and as a direct way of testing alternative analyses. There are still, however, several limitations for its exploitation in the specific study of dialogue prosody and in simulating spontaneous speech in interaction, so that at the present stage several typical ingredients of spoken dialogue could not be implemented in the syntheses. In spite of these limitations we have found that rule synthesis can be a valuable instrument in dialogue prosody research.

The speech synthesis used here allows one to choose from a small set of speaking voices. Two different voices have been selected for participating in our simulated dialogue, the so-called regular male voice and the deep male voice. In our use of rule synthesis, the starting point is a phonetic transcription of prosodic features, basically the same features as described above under auditory analysis.

3. RESULTS

3.1. Laboratory Speech vs. Spontaneous Dialogue

When studying the prosody of spontaneous dialogue against the background of having studied it in a laboratory speech environment, we have encountered relatively few surprises. Although we do not mean to underestimate the difference between read and spontaneous speech, it is our general impression that the difference in prosodic patterning, particularly pitch patterns, between a specially designed, read test material and a spontaneous dialogue is less than we had expected. A typical example of the relative similarity between laboratory speech and spontaneous speech is the following.

The location of a focal accent in (Standard) Swedish represents a pivot (cf. [13]) of a prosodic phrase or utterance. The pivotal character of the focal accent in Swedish can be illustrated by its role in determining the presence or absence of a downstepping pitch contour in read speech material (cf. [3]). In a pre-focal position, up to the focal accent of a phrase (or a whole utterance) there is typically no downstepping, but instead successive non-focal accents occur on more or less the same pitch level. However, after a focal accent, the downstepping of successive non-focal accents is a characteristic pitch pattern. This downstepping seems to be the expression of equal prominence of successive post-focal accents within the phrase.

It is interesting to note that in our spontaneous dialogue speech there are several, typical examples of downstepping and non-downstepping pitch patterns, which seem to be triggered by the placement of focal accent in very much the same way as describe above. For a perspicuous example of this see [8].

3.2. Dialogue Structure and Pitch Range

A fundamental question in the study of dialogue prosody is of course how the prosodic patterns observed are related to the structure of the dialogue itself in terms of textual, interactive and turn regulating aspects. One case in point here is the variation and changes in overall pitch range, which has been in the focus of our interest. Differing degrees of attention generally seem to correlate with variation in range. A more specific hypothesis has been to ascribe variation in pitch range a possible connection with boundaries in the dialogue structure, for example to speech paragraphs or to the introduction of a new conversation topic (cf.[1], [15]).

In the particular Swedish dialogue that we have studied in some detail - a radio listeners' conversation over the telephone with the program leader of a popular radio program "Ring så spelar vi" - the following regularities appear. The combined introduction of a new topic and interactive initiative is reflected by an increase in pitch range in 75 % of the cases. This can be contrasted with the combined continuation of a topic and an interactive response (which is a category four times as common in the actual dialogue), for which we find an almost equal distribution of increase, decrease and no change in pitch range.

3.3. Dialogue Prosody and Speech Synthesis

In our search for regularities of variation in overall pitch range we have also used speech synthesis. Different versions of a dialogue section have been implemented in the rule synthesis. Two versions of our synthesis attempts are interesting for the present discussion. The first one is a neutral version of the dialogue section, where only default utterance prosody is used with no attempt to simulate interaction. Thus the same pitch range is used for the consecutive prosodic groups of the dialogue section.

The second version presents an attempt - in addition to the neutral utterance prosody - to simulate one aspect of dialogue prosody, namely the variation of pitch range for interactive purposes.

A comparison of the two synthesized versions of the actual dialogue section - the neutral version and the pitch range version - clearly shows that variation in overall pitch range may be considered a potentially important means for use in the development of a dialogue and its division into speech paragraphs (see further [11]).

4. REFERENCES

[1] BROWN, G., CURRIE, K. and KENWORTHY, J. (1980), *Questions of intonation*, London: Croom Helm.

[2] BRUCE, G. (1977),*Swedish word accents in sentence perspective*, Lund: Gleerup.

[3] BRUCE, G. (1982), 'Developing the Swedish intonation model', *Working Papers 22*, 51-116, Lund: Dept. of Linguistics.

[4] BRUCE, G. (1985), 'Structure and functions of prosody', *Proc. French Swedish Seminar on Speech* , eds. B. Guerin and R. Carré, 549-559, Grenoble.

[5] BRUCE, G. and GÅRDING, E. (1978), 'A prosodic typology for Swedish dialects', *Nordic Prosody*, eds. E. Gårding et al., 219-228, Lund: Dept. of Linguistics.

[6] BRUCE, G. and GRANSTRÖM, B. (1989), 'Modelling Swedish intonation in a text-to-speech system', *STL-QPSR 1*, 17-21, Stockholm:KTH, Speech Transmission Laboratory.

[7] BRUCE, G. and GRANSTRÖM, B. (1990), 'Modelling Swedish prosody in text-to-speech: phrasing', *Nordic Prosody V*, eds. K. Wiik and I. Raimo, 26-35, University of Turku: Phonetics.

[8] BRUCE, G. and TOUATI, P. (1990a), 'On the analysis of prosody in spontaneous dialogue', *Working Papers 36*, 37-55, Lund: Dept. of Linguistics.

[9] BRUCE, G. and TOUATI, P. (1990b), 'Analysis and synthesis of dialogue prosody', *Proc. ICSLP 90 Vol. 1*, 489-492, Kobe, Japan.

[10] BRUCE, G. and TOUATI, P. (1991), 'On the analysis of prosody in spontaneous speech with exemplification from Swedish and French', To appear in *Proc. ESCA Workshop on Phonetics and Phonology of Speaking Styles*, Barcelona.

[11] BRUCE, G., WILLSTEDT, U. & TOUATI, P. (1990), 'On Swedish interactive prosody: analysis and synthesis', *Nordic Prosody V*, eds. K. Wiik and I. Raimo, 36-48, University of Turku: Phonetics.

[12] CARLSON, R and GRANSTRÖM, B. (1986), 'Linguistic processing in the KTH multilingual text-to-speech system', *In Proc. ICASSP 86, Vol.4*, 2403-2406, Tokyo.

[13] GÅRDING, E. (1981), 'Contrastive prosody: a model and its application', *Studia Linguistica 35*, 146-166.

[14] GÅRDING, E. (1982), 'Swedish prosody', *Phonetica 39*, 288-301.

[15] HIRSCHBERG, J. and PIERRE-HUMBERT, J. (1986), 'Intonational structuring of discourse', *Proc. 24th Meeting of the Association of Computational Linguistics*, 136-144, New York.

[16] I.P.A. (1989), 'Report on the 1989 Kiel Convention', *Journal of the International Phonetic Association 19 (2)*, 67-80.

A RHYTHM-BASED METRIC FOR TURN-TAKING

E. Couper-Kuhlen

University of Constance, Germany
University of Zürich, Switzerland

ABSTRACT

A rhythm-based metric for turn-taking is introduced here and its implications for the identification of marked and unmarked transitions in everyday verbal interaction are spelled out. Empirical evidence to support two specific predictions of a rhythm-based metric is presented, showing (i) that for transition times <0.2 sec. overlap, latching, and micro-pausing can be predicted in relation to the tempo of prior talk, and (ii) that for transition times >0.2 sec. a rhythmic analysis makes a better prediction of noticeability for pauses in the 0.2-0.6 sec. range. It will be argued that a metric based on speech rhythm is superior to one based on absolute time because it allows for a more reliable identification of communicatively significant timing.

1. INTRODUCTION

Early instrumental studies of spoken English utterances showed isochrony to be present only under ideal conditions [2]. However, the more recent discovery of P-centers [6] has reopened the debate [4]. In order to produce perceptual isochrony between a set of monosyllabic words, it has been shown that it is necessary to advance or retard the acoustic onsets of syllables according to their phonetic make-up. In one case the amount of offset shown to be necessary was 80 ms. or 16% of the duration of the interval involved [6]. This suggests that interstress intervals in connected speech could vary at the worst by as much as 16% with respect to the duration of a prior interval and still be considered isochronous. Presumably, however, the ratio for permissible variation within the bounds of perceptual isochrony increases when the intervals under consideration are polysyllabic and/or contain phrase boundaries. In an auditory/acoustic investigation of spontaneous English speech [1], I have found cases of perceptual isochrony in which the percent difference to a preceding interstress interval is as much as 30%. Perceptual isochrony is, however, rarely if ever found with % differences greater than this.

In everyday English conversation, perceptual isochrony is not constant; there are often syncopated beats and noticeable shifts in tempo. Nor is it continuously present. However, it does tend to become particularly pronounced and clear at the ends of turns in the conversations I have examined. This suggests that it may play a role, perhaps even a facilitative role, in the temporal coordination of turn-taking.

2. TRANSITION TIMING: THE UNMARKED CASE

My working hypothesis is that the norm for transition timing in everyday English conversation is temporal coordination between the first beat of the new turn and the last two beats of the prior turn such that an isochronous sequence results. Beats are typically created by syllables with relative prominence at some one level of a prosodic hierarchy constituted by syllables, feet, phonological phrases and intonation phrases. Isochronous patterns arise when every syllable, every unreduced syllable, every (pitch-) accented syllable and/or every intonational onset or nucleus in a stretch of speech is timed at perceptually regular intervals of time. In contrast to earlier studies on isochrony then, the phenomenon in this approach is not restricted to an uninterrupted succession of stressed syllables only.

Once a rhythmic pulse is established in speech, it creates the expectation that it will continue. Therefore, temporary interruptions, provided they are not too long, can be tolerated without causing the rhythmic gestalt to break down. When a pre-established rhythmic pulse coincides with silence rather than with some prosodically prominent syllable, this is referred to as a silent beat.

Given the above hypothesis, prototypical unmarked timing for turn transition might be represented schematically as in Figure 1, where A and B are different speakers. As this diagram suggests, B must position the first prosodic prominence of the new turn such that it follows the last two prominences of A's turn at an approximately equal interval of time. Any intervening non-prominent syllables (whether post-tonic in A's turn or anacrustic in B's turn) will need to be incorporated into the transition interval in such a way that overall isochrony is not disturbed. If there are numerous non-prominent syllables intervening, this may entail latching or even minimal overlap, depending on the tempo of prior speech. If there are few or no non-prominent syllables to fill the interval, a small 'pause' may be called for, again depending on prior tempo.

3. TRANSITION TIMING: A MARKED CASE

According to the present hypothesis, marked timing in turn transition occurs when the first prosodic prominence of a second turn comes either earlier or later than the next rhythmic pulse following a TRP in the prior turn. Only 'late' timing will be considered here.

A prototypical 'late' second turn can have two variants, represented schematically in Figures 2a/b. In both variants the first rhythmic pulse following the last two beats of A's turn is unfilled, i.e. the next turn is late. But in Figure 2a, when B's turn does begin, its first prosodic prominence coincides with a continuation of the pulse, which retrospectively converts the intervening silence into a silent beat and re-establishes the rhythm. In Figure 2b, by contrast, the first prosodic prominence of the new turn does not coincide with the pre-established pulse and it can be assumed that the gestalt-like rhythm rapidly breaks down. The status of silence following a TRP in the marked transition is thus quite different from its potential status in an unmarked transition. In the marked case it contains or coincides with a rhythmic pulse, whereas in the unmarked case any silence which occurs is incorporated into the rhythmic interval between two pulses.

4. EVIDENCE FOR A RHYTHM-BASED METRIC

The rhythm-based metric hypothesized here makes a number of predictions with respect to transition timing which can be investigated empirically. The following will deal with two of these, using two randomly

Figure 1. Prototypical unmarked timing for turn transition

```
                                  TRP
A:    6    (..)    6    (..)    6    (..)
                   ^             ^          ^

B:                                    (..)  6
```

[6 represents a prosodically prominent syllable, (..) optional non-prominent syllables, and ^ the rhythmic pulse established by a perceptually regular succession of 6s.]

chosen fragments from natural, spontaneous English conversation (one telephone, one face-to-face) as a data base. In both cases prosodic prominences and isochronous rhythms were determined auditorily by two trained specialists, working first independently and later as a team. The intervals between perceptually regular prosodic prominences were then measured acoustically, from vowel onset to vowel onset, using wideband sonagrams as produced by a Kay Elemetrics Digital Sonagraph Model 5500. Only those patterns in which intervals deviated by less than 30% from the duration of a prior interval were retained. In addition to rhythmic intervals, all transition times were measured to the nearest one-hundredth of a second.

4.1. Prediction of Overlap, Latching or Micro-pause

From the stipulation that prosodic prominences must be timed regularly in the unmarked case, it follows that the rhythmic interval containing a turn transition may be empty, or filled with a varying number of post-tonic or anacrustic syllables. A rhythm-based metric would predict that for the unmarked case overlap, latching or micro-pause between two turns is a function of how many (and possibly what kinds) of syllables the transition interval contains, in interaction with the rate of the rhythmic pulse (tempo). Where there are few or no non-prominent syllables to be incorporated, a micro-pause is likely, provided the tempo is not fast. Where there are several or many non-prominent syllables, latching or in the extreme case overlap is likely, provided the tempo is not slow. If tempo is held more or less constant, there should be less 'space' between turns the more syllables there are to be incorporated, and vice versa.

When all transition times of 0.2 sec. or less (including cases of minimal overlap between non-prominent syllables) in the conversational fragments examined are now grouped according to the tempo of the rhythmic pulse in surrounding talk, the pattern which emerges supports the prediction that transition times will be shorter where more syllables must be incorporated. In transition intervals of 0.71-0.72 sec., for instance, a 0.2 sec. pause occurs with no incorporated syllables, a micro-pause of 0.1 sec. with 2 incorporated syllables and latching with 4 incorporated syllables. In transition intervals of 0.78 sec. there is a micro-pause of 0.1 sec. with 2

Figure 2a. Marked 'late' timing for turn transition (rhythm preserved)

					TRP				
A:	6	(..)	6	(..)	6	(..)			
			^		^		^		^
B:								(..)	6

Figure 2b. Marked 'late' timing for turn transition (rhythm disturbed)

					TRP				
A:	6	(..)	6	(..)	6	(..)			
			^		^		^		^
B:							(..)	6	

incorporated syllables but overlap with 6 incorporated syllables.

4.2. Identification of Noticeable or Significant Pausing

From a comparison of unmarked vs. marked timing, it also follows that silence occurring between two turns may either fall within a rhythmic interval, or instead coincide with a rhythmic pulse. In the latter case the silence can be expected to be more salient or noticeable, since interlocutors presumably monitor the rhythmic pulse in order to use it as an orientation in timing their entries to the floor. On the other hand, longer pauses are arguably more likely to contain a rhythmic pulse and for this reason may be more noticeable than shorter ones anyway [3].

When all transition times of 0.2 sec. or more in the conversational fragments examined are ranked according to absolute duration, the following pattern emerges. In general, longer pauses are more likely to coincide with rhythmic pulses; pauses longer than 0.7 sec. were never incorporated in the data examined. However, pauses shorter than this were not uniformly incorporated. There were 4 cases of incorporation to 7 of non-incorporation for pauses in the 0.2-0.6 sec. range. Thus, pauses of intermediate duration may or may not contain a rhythmic pulse, in function of the tempo and rhythm of surrounding talk. This suggests that conversationalists may be aware of and attribute signficance to some transitional silences in this range but not to others. An informal test of pause noticeability in the fragments under consideration offers initial support for the prediction that pauses which contain a rhythmic pulse will be more noticeable than those which are incorporated.

5. CONCLUSION

Although more evidence must still be brought to bear on these issues, transition timing in the conversational fragments examined here appears to support a rhythm-based metric for turn-taking in English conversation. If fully confirmed, the hypothesis would offer support not only for a prosodic hierarchy such as that advocated e.g. by Nespor/Vogel [7], but also for a non-concatenative, hierarchical model of speech processing [5]. Just as it can be argued that speech segments are not processed sequentially, so interactional timing may be 'processed' not in terms of cumulative duration but with respect to a rhythmic pulse.

6. REFERENCES

[1] COUPER-KUHLEN, E. (forthcoming), "*English speech rhythm: Form and function in everyday verbal interaction*".

[2] CLASSE, A. (1939) "*The rhythm of English prose*", Oxford: Blackwell.

[3] DUEZ, D. (1985) "Perception of silent pauses in continuous speech", *Language & Speech 28*, 377-389.

[4] FOWLER, C.A. (1979) "'Perceptual centers' in speech production and perception", *Perception and Psychophysics 25*, 375-386.

[5] MARTIN, J.G. (1972) "Rhythmic (hierarchical) versus serial structure in speech and other behavior", *Psychological Review 79*, 487-509.

[6] MORTON, J./MARCUS, S./FRANKISH, C. (1976) "Perceptual centers (P-centers)", *Psychological Review 83*, 405-408.

[7] NESPOR, M./VOGEL, I. (1986) "*Prosodic Phonology*", Dordrecht: Foris.

SOME WAYS IN WHICH FORMS ARISE FROM FUNCTIONS IN LINGUISTIC COMMUNICATIONS

Carol A. Fowler and Elena T. Levy

Haskins Laboratories, New Haven, Connecticut 06511

ABSTRACT

We are examining some ways in which talkers signal discourse structure to listeners. Earlier research had suggested that words are shortened in acoustic duration the more redundant they are, and other findings suggested that the lexical length of referring expressions varies as a function of their role in spoken discourse. Comparing across the lines of research, we have speculated that the two levels of shortening may occur in response to variation in some of the same discourse variables. The research on which will report offers supportive evidence for the variable, order of mention in an episode unit.

1. INTRODUCTION

Some language forms arise from language use. More or less as the character of a riverbed reflects the dynamical forces that have formed it or as a fossil tooth reflects the dietary habits of a former chewer, some common phonological and lexical forms of languages may reflect the constraints on talkers and listeners that have given rise to them.

Some constraints are articulatory and perceptual. In particular, the literature provides evidence of striking parallels between certain phonological systematicities of a few languages and phonetic regularities that are universal or nearly so. The parallels have been taken to suggest that the phonological forms arose as elevations from, and conventionalizations of, articulatory dispositions of the vocal tract [12] triggered, perhaps, by systematic misperceptions of members of a language community [13]. Some parallels, among others, are the following. In nearly all languages that have been examined, final voiced obstruents are partially devoiced, while in some languages, a phonological voicing distinction among obstruents is neutralized word finally. In many languages, vowels are shortened in measured acoustic duration as consonants are added to the syllable rhyme, while in some languages, phonologically long vowels can only occur in open syllables or followed by at most one short consonant. Historically, loss of a consonant in the rhyme of a syllable has occasionally triggered phonological lengthening of a preceding vowel. In many languages, intonation contours exhibit downdrift or declination, which tracks the falling subglottal pressure of the lungs [6], while some languages have intonational downstep rules, and some tone languages have downstepping lexical tones.

Articulatory dispositions and mishearings do not, of course, exhaust the communicative constraints that may shape language forms. In our presentation, we will examine effects of two additional hypothetical constraints: speaker efficiency and comprehensibility of the linguistic message. As for the articulatory and perceptual constraints, these superordinate constraints may give rise to parallel regularities at distinct linguistic levels--in this case, lexical/syntactic and prosodic/phonetic. In contrast to the phonological and phonetic correspondences described above, in the case of these additional parallels, we do not identify a directional arrow--that is, an indication that features at one linguistic level derive from those at another. Rather, we speculate that the same communicative pressures may exert themselves concurrently at several levels of linguistic structure and may leave parallel traces behind. The features on which we report are a durational shortening (or lengthening) of words that are less (more) expected by the listener and a lexical shortening of referring terms under approximately the same conditions.

As for the phonetic effects, Bolinger [2,3] suggests that words that are unexpected in their contexts (e.g. "mowed" in "he mowed home") are lengthened in duration as compared to their duration in contexts where they are expected ("he mowed the grass"). Perhaps compatibly, Lieberman [11] has found that spoken words excised from contexts in which they are predictable are less identifiable than the same words excised from contexts in which they are unpredictable. Even out of context, spoken words that are chronically likely to be produced (that is, high frequency words of a language) are shorter in duration than unlikely words (e.g. [15]); this holds even for more and less frequent nonhomographic homophones [14].

Similar effects are found in spontaneous speech [5] and, to a lesser extent, in read discourse [4]. Words produced for the first time are durationally longer than the same words repeated (as long as they have the same referent on both occasions; see [1]). On the listeners' side, second occurrences of words are generally more predictable from their contexts than are first occurrences, and there is some evidence [5] (but see [1]) that the durational reduction itself has communicative significance to listeners.

These effects may indicate at least that speakers reduce words when they know that the listener can get by with a less adequate acoustic signal, because the context predicts the word. In addition, however, if the findings on listeners' perceptions are real, they may show that listeners use durational reduction as information that a word is "old" and hence refers back to material earlier in the discourse. In turn, information that a word is old may facilitate retrieval of relevant earlier material.

Turning to the findings of lexical shortening, when terms for new referents are coined, their names often are long, and their meanings are sometimes decipherable from their component morphemes ("automobile", "videocassette recorder"). When real-world referents of these new terms are talked about frequently and become commonplace, their names often shrink and become less transparent ("car"; "VCR"; cf [16]). Thus, there is a wearing away of terms with use that is reflected also in the finding that high frequency words of a language are shorter than low frequency words [16] (see also [8]).

On a shorter time scale, in spoken discourse, a similar phenomenon can be observed. Givon [7] suggests a principle whereby less predictable or less accessible topics in a discourse tend to be coded using more linguistic material than is used to code more predictable and accessible topics. In particular, in an analysis of referring terms, he suggests that referring terms vary in length depending on predictability and accessibility along the following continuum (abbreviated slightly here) from least to most accessible: modified full NPs, full NPs, stressed pronouns, unstressed pronouns, zero anaphor.

Compatibly, in an analysis of the spontaneous narrations of four speakers (who recounted a film that they had seen to naive listeners), Levy [9,10] found a strong relation between the length of referring terms (references to either of two male characters in the film) and two measures of accessibility of the referents to the listener. In particular, referring expressions were longer when the immediate context of the targeted expression was "noncoreferential" than when it was coreferential. (A coreferential context is one in which the last male reference to occur in a parallel position to the target reference is coreferential with it.) In addition, longer expressions occurred in "sparse" rather than "dense" contexts (where a dense context referred to an immediately preceding paragraph in the discourse in which the targeted character was more frequently mentioned than were other characters). Interestingly, Levy identified another variable that was associated with the length of a referring expression that is particularly analogous to findings of Fowler and Housum [5]. She found that longer referring expressions were used to refer to a character's first, as compared to subsequent, mentions in an episode unit of the discourse.

2. OUR ONGOING RESEARCH

The research on which we will report examines the relation, if any, between the two levels of length variation that we have described. In particular, we are examining the acoustic durations of full NP expressions referring to the two main characters in the film narrations collected by Levy [9]. We know from that earlier study that referring expressions in these narrations exhibit lexical length variation in response to the three discourse variables: coreferentiality and density of prior mention and order of mention in an episode unit. In the narratives of four of the eight speakers that we have examined to date, we find consistent effects of order of mention in a episode within the narratives such that first mentions of full NPs in an episode are durationally longer than subsequent mentions, even when the character has been mentioned previously in the narrative. For one main character, across the four talkers, first-mentioned full NPs are 42 ms longer than subsequent mentions on average ($F(1,149) = 3.88$, $p = .05$). For the other main character, first mentions are longer by 107 ms on average than subsequent mentions ($F(1,121) = 9.56$, $p = .0025$). We have also looked for effects of the discourse variables, coreferentiality of prior mention and density of prior mention, but, in the two speakers in whom we have examined the data, these variables do not affect acoustic duration of referring expressions in a consistent way.

We speculate that the systematic variation that the literature reveals and that we have found in the phonetic durations of expressions and in their lexical length, may originate in a sort of tradeoff between a talker's goal of verbal efficiency and the requirement that listeners be able to recover the intended communicative message. As a communication goes forward, some topics arise temporarily as central and hence as accessible to and predictable by the listener, while other topics are less central, accessible and predictable. Compatible with a goal of efficiency, speakers will shorten, in either or both of two ways, terms relating to the accessible, predictable topics. The pattern of shortenings and lengthenings themselves may be informative to listeners, however, who then can use evidence of durational reduction or that a referring term is inexplicit as information that a referent is "old" and can determine from the fact that a referring term is inexplicit that the referent is viewed by the speaker as focal to a topic.

3. REFERENCES

[1] BARD, E., & BREW, C. (1990), *Psycholinguistic studies on incremental recognition of speech: An introduction to the messy and sticky*. University of Edinburgh.

[2] BOLINGER, D. (1963), "Length, vowel, juncture", *Linguistics, 1*, 5-29.

[3] BOLINGER, D. (1981), *Two kinds of vowels, two kinds of rhythm*, Bloomington, Indiana : Indiana University Linguistics Club.

[4] FOWLER, C. A. (1988), "Differential shortening of repeated content words produced in various communicative contexts", *Language and Speech, 31*, 307-319.

[5] FOWLER, C. A. & HOUSUM, J. (1987), "Talkers' signalling of 'new' and 'old' words in speech and listeners' perception and use of the distinction", *Journal of Memory and Language, 26*, 489-504.

[6] GELFER, C. (1987), *A simultaneous physiological and acoustic study of fundamental frequency declination,* PhD Dissertation, City University of New York.

[7] GIVON, T. (1985), "Iconicity, isomorphism and non-arbitrary coding in syntax", In J. HAIMON (Ed.), *Iconicity in syntax* (pp. 187-219), Amsterdam: John Benjamins.

[8] LANDAUER, T., & STREETER, L. (1973), "Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition", *Journal of Verbal Learning and Verbal Behavior, 12*, 119-131.

[9] LEVY, E. T. (1984), *Communicating thematic structure in narrative discourse: The use of referring terms and gestures*, PhD Dissertation, University of Chicago.

[10] LEVY, E. T., & MCNEILL D. (in press), "Speech, guesture and discourse", *Discourse Processes.*

[11] LIEBERMAN, P. (1963), "Some effects of semantic and grammatical context on the production and perception of speech", *Language and Speech, 6,* 172-187.

[12] MACNEILAGE, P., & LADEFOGED, P. (1976), "The production of speech and language",In E. C. CARTERETTE, & M. P. FRIEDMAN (Ed.), *Handbook of perception: Language and speech,* (pp. 75-120), New York: Academic Press.

[13] OHALA, J. (1981), "The listener as a source of sound change", In C. MASEK, R. HENDRICK, R. MILLER, & M. MILLER (Ed.), *Papers from the parasession on language and behavior,* (pp. 178-203). Chicago: Chicago Linguistics Society.

[14] WHALEN, D. Unpublished data.

[15] WRIGHT, C. (1979), "Duration differences between rare and common words and their implications for the interpretation of word frequency effects", *Memory and Cognition, 7*, 411-419.

[16] ZIPF, G. (1935/1965), *The psycho-biology of language,* Cambridge, MA: MIT Press.

INTEGRATING SYNTAGMATIC AND PARADIGMATIC ASPECTS OF STRESS

D. Robert Ladd

Department of Linguistics, University of Edinburgh

ABSTRACT

The purely syntagmatic analysis of stress assumed in metrical phonology is difficult to reconcile with the fact that words can in practice be classified as either stressed or unstressed. Various properties of stress (vowel reduction, nuclear stress, and the use of stress to signal focus and deaccenting) can be integrated into metrical phonology if we define three paradigmatic levels of stress in terms of the prosodic categories phrase and foot.[1]

1. A STRESS PARADOX

When linguists (and others) discuss the *function* of stress, they normally find it sufficient to indicate stress in any given sentence by capitalising or underlining stressed words. This notational convention implies that stress is a paradigmatic property that can apply to a word more or less independently of what happens to adjacent words. Moreover (though we may allow for the possibility of secondary stress), it strongly suggests that stress is a fairly categorical property: either a word is stressed, or it isn't.

When discussing the *phonological nature* of stress, however, linguists have been led to construct increasingly elaborate theories that emphasise its syntagmatic and non-categorical aspects. Metrical phonology (e.g. [3],[6]), in particular, emphasises that stress does not involve paradigmatic features but only syntagmatic relations, of which the theory's central notational device, the weak-strong branching node, is emblematic. Nor do metrical trees define any sort of categorical distinction between stressed and unstressed (or even a three-way distinction among primary, secondary, and unstressed), because in theory there is no limit to the depth of stress subordination they can express.

How can the evidence for the syntagmatic or relational view be reconciled with the practical usefulness of the categorical capital-letter stress notation?

2. PROSODIC CATEGORIES

Despite the success of metrical phonology in expressing the syntagmatic aspects of stress, everyone acknowledges that, in some way, at least some properties of stress are not relational or syntagmatic at all. The most conspicuous problem in the description of English stress is vowel reduction, and the existence of minimal pairs like *raider* and *radar*. Both of these are strong-weak, but there's a further difference of prominence between the reduced weak syllable of *raider* and the unreduced weak syllable of *radar*. We also encounter the converse problem, namely structures in which the relational representation demands a difference in relative prominence, but in which the stresses in question appear to be equal. For example, Culicover and Rochemont [1] suggest that the "multiple primary stress[es]" in a sentence like

(1) John told BILL about SUSAN, and SAM about GEORGE

have equal relative prominence. They propose (p. 127) that "in order to

---

[1]This paper is a "bonsai" version of a paper presented at the 7th East Coast Conference on Linguistics (ESCOL 7) at Ohio State University in September 1990, and is part of work in progress on prosodic structure. Thanks to Anne Cutler for the bonsai metaphor.

accommodate [such] instances ..., [metrical] theory must be modified so as to allow prosodic nodes to dominate two s[trong] sisters..."

The "equal-primaries" case has never been given very much attention, but the problem of vowel reduction has been debated extensively. Liberman and Prince [3] (henceforth LP) treated such cases as *raider-radar* in terms of a feature [+/-stress] that could be applied to terminal elements of the stress tree - syllables - more or less without regard to their place in structure. Thus:

(2)

s w	s w
raider	radar
+ -	+ +

But in an early response to this analysis, Selkirk [6] proposed to get rid of the [+/-stress] feature by adding **prosodic categories** to the abstract relational structure posited by LP. Thus in place of (2) we will have

(3)

ω	ω
Σ	Σ_s Σ_w
σ_s σ_w	σ σ
raider	radar

In *raider*, the word (ω) consists of a single stress foot (Σ), within which there is a strong-weak relation between the two syllables (σ). In *radar*, on the other hand, the word consists of two stress feet between which there is a strong-weak relation, and each stress foot contains only a single syllable.

This provides a solution to the problem of non-syntagmatic properties, because prosodic categories, unlike the purely relational nodes in the LP metrical trees, can have intrinsic - paradigmatic or nonrelational - phonetic properties defined independently of their place in structure. In Selkirk's words, "a syllable which *is a* stress foot will never be interpreted as a weak [unstressed] syllable ... [B]eing a stress foot always implies some degree of prominence". In what follows I suggest that we might use judiciously selected prosodic categories to give formal definitions of the (apparently) paradigmatic "levels of stress" that non-phonologists find so useful - and incidentally, resolve the equal-primaries problem as well.

3. DEACCENTING AND FOCUS

3.1. Syntagmatic...

First let's consider the apparently unrelated problem of **deaccenting** that I discussed in my thesis [2] - the use of reduced prominence to signal that an item is already in the discourse, given information, etc. Deaccenting is of interest because, superficially, it appears to support the paradigmatic, "capital-letter" view of stress, and yet, on the closer inspection I gave it in my thesis, it appeared to be analysable purely in syntagmatic terms.

An example of classic deaccenting is seen in (4):

(4) The only stuff written about this is in German, and I can't READ German, so I guess I'll work on something else.

Here *German* is deaccented because it's repeated in the discourse context; as I noted in my thesis, the stress is on *read* not for any "positive" reasons, such as focus or contrast, but specifically in order to deaccent *German*, which would otherwise be stressed. Ostensibly, the stress simply "shifts" from one word to another.

This stress-shift account is consistent with the non-relational view of stress (note the use of capital-letter notation in (4)!) But because the LP account of the phonology of stress seemed superior in other respects, I was concerned in my thesis to establish that deaccenting is phonologically syntagmatic or relational, and to get away from treating deaccenting data in terms of the presence or absence of stress on this word or that. Specifically, I showed that there are certain aspects of deaccenting that are puzzling under the commonsense account, but which can be readily explained if we treat deaccenting not as stress shift, but as a *reversal of relative strength in a metrical tree*.

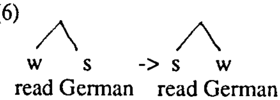
The main such problem I dealt with was the case of *rightward shift* of stress. In the classic case of deaccenting - as in (4) - stress shifts to the *left* compared to the normal location. In some cases, however, deaccenting shifts stress to the right:

(5a) ("normal")
A: Anything happen while I was out?
B: My PARENTS called.

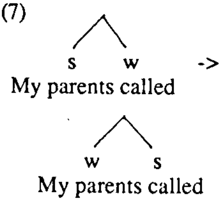
(5b) (*parents* deaccented)
A: Maybe we should call your parents and tell them.
B: My parents CALLED - they already know.

What I suggested in my thesis is that both leftward and rightward shift could be given a unified description in terms of reversed strong-weak nodes. So the trees (or relevant subtrees) in (4) and (5) will be modified as in (6) and (7) respectively:

(6)



(7)



This node-reversal analysis works for a wider range of cases than a straight leftward stress-shift rule.

3.2. ...or paradigmatic?

Nevertheless, there are cases that the analysis doesn't fit very comfortably. For example, there are sentences in which the semantic/pragmatic effects of deaccenting are acheived by the use of distinct pitch accent patterns on accented words. In my thesis I discussed the case of sentences like

(8a) The butcher charged me a thousand bucks!

With one type of pitch accent *butcher* may be interpreted as an epithet for "doctor", while with a different type of pitch accent *butcher* is interpreted literally. This is exactly the difference of interpretation produced by deaccenting or not deaccenting *butcher* in sentence-final position, as in

(8b) I'd like to strangle the butcher!

Unlike (8b), however, the difference in (8a) is not readily interpreted in terms of node reversal. In both readings there would seem to be a weak-strong relationship between *butcher* and *bucks*, and it is rather the different *paradigmatic* choice of pitch accent on *butcher* that conveys the intended interpretation.

Similar phenomena can be observed in the use of sentence stress to signal narrow focus or contrast, as seen in example (9). The context of this utterance was a discussion of somebody who used to be able to speak German well but had then spent a long time living in Sweden and now spoke good Swedish but had trouble with German. My contribution to the discussion was:

(9) That's what happened to MY FRENCH - it used to be good, but then I spent a year in Germany and ended up with good German, and now whenever I want to speak French I get German interference all over the place.

The relevant part of this discourse is the very beginning: *That's what happened to MY FRENCH*. There's clearly a double contrast or focus intended here: on the one hand, we're talking about *my* linguistic abilities rather than those of the person who lived in Sweden, and on the other hand, we're talking about knowledge of *French* getting lost rather than knowledge of German. If we didn't intend the extra focus or contrast on *my*, *my* would be unstressed; it would be somewhat shorter, possibly with a somewhat centralized vowel, and without any sort of pitch accent.

The problem for the reversed-nodes analysis is that the phonological modifications that signal the "deaccenting" of *butcher* or the "focusing" of *my* cannot be described in syntagmatic terms. Both effects are clearly prosodic, but do not involve reversed nodes. Both could, however, be described in terms of modifications of a "normal" or "expected" **level of stress**. In the case of the focus on *my*, *my* is still weak relative to its strong sister *French* (in its original context it clearly had "secondary stress") but we perceive it as focused because it is stronger or more prominent *than it would be in a non-focal context*. That is, it is stronger than some other *paradigmatic* possibility, namely complete lack of stress. Similarly, the pitch accent on *butcher* in the "epithet" interpretation of (8a) is (in Pierrehumbert's terms [4]) a *prenuclear* **H*** - and hence arguably *secondary* stress - while that in the "literal" reading is a nuclear **H*** plus **L** phrase accent - and hence primary. The deaccenting is thus also signalled paradigmatically, by making *butcher* weaker or less prominent than it would be in a non-deaccenting context.

It thus seems that it was a mistake to try to reduce deaccenting to a matter of relative strength - i.e. to reversal of a syntagmatic strength relation. Instead, to a considerable extent, the signalling of focus and deaccenting is based on a neutral (unmarked, default) *degree* of prominence for any given part of speech. Based on that neutral level of prominence, focus (newness, contrast, etc.) is signalled by an increase in the degree of prominence, or **promotion,** while deaccenting (givenness, coreferentiality, etc.) is signalled by a decrease in the degree of prominence, or **demotion**. It seems to be a reasonable generalisation that pronouns, prepositions, and the like are normally unstressed; if they have stress (even secondary stress), it is interpreted - paradigmatically - as conveying some sort of focus. Nouns, on the other hand, normally have primary stress; if their stress is reduced (even to secondary stress), it is interpreted as deaccenting.

4. LEVELS OF STRESS

In order to make descriptive statements of the sort just made, we have to be able to treat the notions of primary stress, secondary stress, and unstressed as degrees of prominence that are statable independently of any given utterance context - i.e. paradigmatically. How can we integrate these notions into the metrical description of stress that we want for other reasons? I propose to do this by defining them in terms of prosodic categories.

Let us posit two prosodic categories, **foot** (F) and **phrase** (P). (For expository purposes I assume here that phrase is the next higher prosodic category above foot, though I'm well aware that this runs counter to the most recent work.) Foot has properties of the sort that Selkirk talked about - unreduced vowel quality and full syllable duration - and is equivalent to Selkirk's Σ. Phrase has primarily intonational correlates - it's the domain of an intonation contour. As Selkirk suggested, the difference between stressed and unstressed is the difference between being a foot and not being a foot. Thus the difference between the two renditions of *my French* could be something like the following (assuming in (10a) the notion of structural extrametricality discussed in [5]):

(10a) P
/ |
| F
| |
σ σ
my FRENCH

(10b) P
/ \
F_w F_s
| |
σ σ
MY FRENCH

The difference between primary and secondary stress, meanwhile, is the difference between being the strong foot of a phrase and being a weak foot. Thus in (10b), the stress on *my* is secondary, while that on *French* is primary or nuclear.

This means we can define neutral prominence for a noun as

(11) P
|
F_s

and "reduced" or "deaccented" prominence as

(12) P
|
F_w

For a pronoun, etc., neutral prominence is

(13) F
|
σ_w

which by definition cannot be deaccented or made less prominent, while increased prominence for focus, contrast, etc. is

(14) P
|
F

The structure in (14), with only F instead of F_w or F_s, says that the very fact of being a foot is enough to signal increased prominence on a function word. It doesn't matter whether the foot is weak or strong in the phrase (i.e. it doesn't matter whether it has primary or secondary stress). This is unlike the situation with nouns in (11) and (12), where the difference between weak and strong in the phrase is exactly what signals deaccenting.

Notice that this analysis subsumes the proposal in my thesis to treat deaccenting as syntagmatic strength reversal. The reason for this is that the strength reversal will happen automatically *assuming certain well-formedness conditions on trees*. Take the case of *My parents CALLED*. The neutral version of this (as in 5a) would be

```
(15)      P
         / \
       Fs   Fw
My parents called
```

To deaccent the subject, we must give it secondary stress, i.e. put it in the configuration shown in (12). But we can't just do that in the tree as it stands, because that would yield the ill-formed structure

```
(16)      P
         / \
       Fw   Fw
My parents called
```

It's therefore necessary to promote the verb, yielding the correct tree (as in 5b)

```
(17)      P
         / \
       Fw   Fs
My parents called
```

In other words, the node reversal happens indirectly, as a consequence of reducing the prominence level on the subject and repairing the resulting violation of well-formedness conditions on metrical trees. The phonological essence of deaccenting is the paradigmatic demotion from primary to secondary stress, not node reversal itself.

5. CONCLUSION

The stress paradox with which we began can be resolved if we take primary stress, secondary stress, and unstressed to be paradigmatic categories, while treating fine differences of relative prominence to be a matter of syntagmatic structure.

Among other things, this makes it possible to reconcile the fundamental assumptions of metrical phonology with Culicover and Rochemont's view, discussed earlier, that the primary stresses in an utterance are equal. The way in which they are equal is *paradigmatic*: both are the primary stresses of their respective phrases. At the same time, there are good reasons to suggest that *syntagmatically* the two P nodes are in a weak-strong relationship, which justifies the traditional view that one fo the primaries is the nucleus of the whole sentence. The two claims need not be seen as incompatible.

6. REFERENCES

[1] Culicover, Peter W., and Michael Rochemont (1983). Stress and Focus in English. *Language* 59: 123-165.

[2] Ladd, D. Robert (1980). *The Structure of Intonational Meaning: Evidence from English.* Bloomington, Indiana University Press.

[3] Liberman, Mark; Prince, Alan (1977). On stress and linguistic rhythm. *Linguistic Inquiry* 8: 249-336.

[4] Pierrehumbert, Janet (1980). *The Phonology and Phonetics of English Intonation.* PhD Thesis, MIT.

[5] Pierrehumbert, Janet, and Mary Beckman (1988). *Japanese Tone Structure.* Cambridge, Mass., MIT Press.

[6] Selkirk, Elisabeth O. (1980). The Role of Prosodic Categories in English Word Stress. *Linguistic Inquiry* 11: 563-605.

PRODUCTION AND PERCEPTION OF PROSODIC PROMINENCE

Jacques Terken

Institute for Perception Research, Eindhoven, The Netherlands

ABSTRACT

The variation of prosodic prominence in human speech is ascribed both to pragmatic and lexical/metrical factors. In order to account for the influence of pragmatic factors, the traditional Given/New distinction must be replaced by a hierarchical ordering, reflecting the relative importance of expressions. In addition, prominence differences serve a demarcative function. On the perceptual side, the demarcative function seems more relevant than the pragmatic function: it seems unlikely that listeners can use prominence differences to recover fine gradations in relative improtance.

1 INTRODUCTION

Variation of prosodic prominence is an important feature of natural speech. By this we do not only mean the presence or absence of prominence such as established by the distribution of pitch accents, but also the relative differences in prominence for accented syllables. Synthetic speech which does not contain such relative differences in prominence sounds rather dull.

This observation immediately raises a number of questions, both with respect to the production and perception of prosodic prominence. This paper will briefly discuss some of these questions. Before doing so, we will first indicate how the present discussion relates to the traditional classifications of prosodic prominence.

Prosodic prominence is usually defined in terms of variation in duration, F0 and amplitude. In the frequency domain, prominent speech units coincide with appropriately timed F0 changes (or local F0 maxima or minima). In the temporal domain, prominent speech units are lengthened in comparison with non-prominent units: unit duration exceeds the duration that would be expected on the basis of speech rate, phonological class and phonemic context if the unit were non-prominent. The perceptual tolerance for temporal variation is quite large (Nishinuma & Duez, 1989), and F0 variation appears to be the most reliable acoustic correlate for prosodic prominence. For that reason, prominence is usually discussed in terms of the distribution of pitch accents: speech units can be either prominent by virtue of the presence of a pitch accent or non-prominent if there is no pitch accent. This makes prominence a binary feature. However, a finer differentiation can be made for prominent speech units. Liberman & Pierrehumbert (1984) present evidence that speakers can very reliably comply with the instruction to make a word more or less prominent. The effect is that the F0 maximum increases if the speaker is asked to pronounce the word with a greater "degree of overall emphasis or excitement".

Our purpose is to incorporate such quantitative differences between accented speech units into the treatment of prosodic prominence.

2 THE PRODUCTION OF PROSODIC PROMINENCE

Traditionally, prosodic prominence has been related to information structure (e.g. Halliday, 1967): Given information is expressed by unaccented, i.e. non-prominent expressions, and New information by accented, i.e prominent expressions. In this treatment, no satisfactory account was given as to the location of accents (i.e. the location of prosodic prominence) within the expressions conveying New information. Later treatments, building on this framework, have related prosodic prominence to the focus structure of the discourse. Expressions can be [$+focus$] or [$-focus$]. The assignment of [$\pm focus$] is driven by pragmatic factors such as the Given/New status. Within a [$+focus$] expression, metrical rules determine the position of the accent. These metrical rules are sensitive to syntactic properties of the sentence, such as functor-argument relations. Stylistic considerations determine whether additional words or syllables will be accented in [$+focus$] expressions. In this way, the search for factors determining the assignment of focus can be separated from the search for rules determining the assignment of accents within focal expressions.

Now, the question is to which level relative differences in prominence must be ascribed. There are two broad classes of models:

1. according to one class of models, the speaker may decide that not all [$+focus$] expressions are to be focussed upon to the same extent, for reasons which have to do with the pragmatic context and/or the thematic structuring of the sentence. The consequence is that we replace the binary feature [$\pm focus$] by a n-ary valued feature [$\alpha\, focus$]. The mapping of [$\alpha\, focus$] onto F0 values is done by means of a grid. We will call this the VARIABLE FOCUS view;
2. according to a different class of models, the speaker assigns [$\pm focus$] to expressions on the basis of pragmatic considerations, and prominence differences originate from the mapping of [$+focus$] onto F0 values. Particular models may differ in the way this mapping is conceived of. Some models attribute prominence differences to the outcome of the metrical rules which determine the location of accents in [$+focus$] expressions. Essentially, this boils down to replacing the binary feature [$\pm\, accent$] by a n-ary feature [$\alpha\, accent$]. Other models attribute prominence differences to lexical factors: each part of speech has associated with it a fixed prominence (cf. Allen, Hunnicutt & Klatt, 1987). Still other models attribute prominence differences to the operation of prosodic rules such as downstep. We will call this the MAPPING view, since it ascribes prominence differences to factors which come into play when [$+focus$] is mapped onto F0 values.

Before we can discuss the different options, the status of α must be considered. From a linguistic point of view, it is implied that α is a nominal variable, i.e. the different values that α can take are qualitatively different and distinctive. We want to be more lenient and to avoid these implications. Instead, we consider α as an ordinal variable: the values of alpha are defined as the set {more, less, equal}. Also, the requirement of distinctiveness must be replaced by a requirement that different values of α are associated with different felicity conditions.

The variable focus view accounts for prominence differences between focal expressions, since it assigns to each focal expression a particular value for α. Support for this view can be found in different sources. In the introductory section, we have already referred to Lieberman & Pierrehumbert (1984). Kruijt (1985) shows that F0 maxima are lower in focal expressions referring to Given referents than in focal expressions referring to New referents (Given is used in the sense of "mentioned in the immediately preceding context", cf. Chafe, 1974; Brown, 1983; Terken, 1985). Wells (1986) describes an experiment in which listeners were presented with utterances isolated from their context. He found that rankings of the relative importance of information conveyed by expressions in the utterances were systematically related to the prosodic characteristics of the expressions. Thus, it appears that prosodic features can be used by speakers to convey gradations in focus. Needham (1990) shows that focal expressions referring to non-typical parts of a previously mentioned whole are associated with higher F0 maxima than the same focal expressions referring to typical parts (we assume that this is a priming effect).

These findings can be accounted for in a natural way, if we take inspiration from proposals within the area of computational linguistics, which are intended primarily for anaphora resolution (e.g. Asher & Wada, 1988; Hajicova, Kubon & Kubon, 1990). They elaborate on the intuition that a hierarchical ordering can be established on the items which have been made accessible by the discourse. Anaphora resolution is guided by the hierarchical ordering of the set of candidate antecedents for a given anaphoric expression. For instance, Hajicova *et al.* describe an algorithm by which a salience index can be computed for the set of accessible items (i.e. the stock of shared knowledge).

In order to relate these proposals to relative differences in prosodic prominence, we must extend the idea of a hierarchical ordering to the information which is to be transmitted by the speaker:

1. for items in the stock of shared knowledge, prosodic prominence is directly related to their hierarchical ordering: the most accessible items are least prominent;

2. for items which are to be transmitted, a hierarchical ordering is established by the relative weight which is assigned to them by the speaker; this relative weight is affected by thematic roles (more central roles carry more weight than less central ones) and priming effects (information which can be more easily activated from the information in the stock of shared knowledge carries less weight than information which can be less easily activated); prosodic prominence is highest for items which are highest in the hierarchy.

If this account of prominence differences between focal expressions in terms of pragmatic factors is appropriate, prominence differences within focal expressions require a different explanation; such within-expression differences cannot be accounted for in terms of different values of $[\alpha\, focus]$, since we have assumed that $[\alpha\, focus]$ is assigned to each focal expression as a whole. It can be argued, however, that such within-expression differences do not emerge from pragmatic influences but from lexical and/or phonological factors. This means that the mapping view would be appropriate for relative differences in prominence within focal expressions.

Terken (1991b) describes findings related to the issue of prominence variation in read aloud text. The materials consisted of referring expressions embedded in texts read aloud by a professional speaker. Although there were clear differences in prominence within these expressions, no general pattern emerged. In order to determine the perceptual tolerance for different prominence patterns, listeners were presented with manipulated versions of expressions containing three accented words, embedded in their sentential context. The expressions constituted the maximal projections of Noun Phrases or Preprositional Phrases.

In general, listeners had equal preference for two different prominence patterns. In one pattern, which would be predicted on the basis of rhythmic alternation (Monaghan, 1988), maximal prominence was associated with the "edges" of the expression. In the other pattern, maximal prominence was associated with the left edge of the ex-

pression, and prominence decreased as a function of serial position. This pattern more resembles a downstep pattern. Both patterns were strongly preferred over a pattern in which maximal prominence was on the middle of three accented words. In both cases, prominence differences can be said to have a demarcative function: words with major prominence are at both edges or at the left edge of a syntactic constituent.

In addition, preferences for a particular prominence pattern were modified by lexical factors: if the word at the left edge was a semantically weak adverb such as "rather", there was a preference for reduced prominence at the left egde, and major prominence was shifted to the middle word.

These findings are supported by unpublished data from read aloud isolated utterances. In these data, a typical decrease in prominence appears to be associated with serial order within referring expressions containing several accented words: each following accent is less prominent than the preceding one.

As a consequence, we have two factors which contribute to differences in prosodic prominence: pragmatic factors govern the relative prominence of $[+focus]$ expressions, and lexical and phonological factors affect the relative prominence of accented words within $[+focus]$ expressions. The phonological factors concern the demarcation of syntactic constituents.

3 THE PERCEPTION OF PROSODIC PROMINENCE

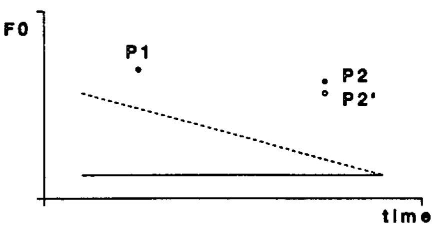
If listeners should be able to recover the value of $[\alpha\, focus]$ from relative differences in prominence, we must assume that they can establish prominence differences between non-adjacent pitch accents. Due to the transient character of the speech signal, it seems unlikely that this can be done on the basis of F0 values directly. Therefore, it must be assumed that relative prominence is recoded in terms of a grid-like structure, where the grid defines a set if "iso-prominence" curves. However, as Lieberman demonstrated already in 1965, even expert listeners cannot do so reliably on the basis of acoustic information.

This implies that, although there may be reliable relations between $[\alpha\, focus]$ and relative differences in prominence on the part of the speaker, it is unlikely that a one-to-one mapping of $[\alpha\, focus]$ can be recovered by the listener. Instead, we assume that listeners can give rather accurate judgments of relative differences in prominence between pairs of prominent syllables. For two successive accents A and B, A can be more or less prominent than B, or they can be equally prominent. If A and B are equally prominent, or B is less prominent than A, and there is a third accent C which is less prominent than B, from this it follows that C is also less prominent than A. However, if B is more prominent than A, the relation between A and C cannot be established.

If this is valid, it would imply that relative differences in prominence have primarily a demarcative function by telling the listener when a new constituent starts, and that the signalling of focus strength is of secondary importance. Only if successive constituents each contain just one accented word, can the speaker signal focus strength by means of relative differences in prominence.

Further experiments are needed to determine whether this picture is valid. Before conducting these tests, we need more insight into the perception of prominence differences. The primary acoustic correlates of prosodic prominence are well-known, but it is not fully clear how they contribute to the perception of prominence differences. In particular, there is no fixed procedure to determine which one of two accented words will be perceived as more prominent. Since pitch information will play an important part in such a procedure, an experiment was done addressing the question of how F0 variation contributes to prominence (Terken, 1991a).

In the experiment, utterances containing two accented syllables were presented to listeners. The two accented syllables were each associated with a frequency maximum, P1 and P2, respectively. The temporal distance between the two accented syllables was kept constant. The listeners were asked to adjust P2 so that the second accented syllable was judged to have the same prominence as the first accented syllable, for different values of P1. There were two different conditions. In one condition there was no baseline declination and different values of P1 were associated with variations in the distance between the topline and the baseline. In the other condition, the slope of the baseline was varied, so that the distance between topline and baseline within the utterance remained constant, but the scaling of the contour within the overall frequency range varied. A schematic representation of the results is shown in Figure 1. Here, for a given P1, the corresponding P2 which gives the same prominence as P1 is shown for the conditions with baseline declination (P2′, the dashed line indicates the baseline, the open circle indicates the position of P2′ giving the same prominence as P1) and without baseline declination (solid line, filled circle). P2 is adjusted to lower values in utterances with baseline declination than in utterances without baseline declination.



As can be seen from the schematic representation, a steeper slope of the baseline was associated with an upward shift of the initial part of the contour within the overall frequency range. Now, we assume that the speaker, when he is higher up in the overall range, has less room to bring about prominence variations by means of F0 variation due to a ceiling effect. In general, the range of F0 values employed and the position of the latter part of the contour within the overall range vary little for a given speaker. From these considerations, it may be concluded that the prominence associated with a given F0 maximum is affected by the actual frequency range employed by the speaker: if he can employ a small frequency range in the beginning of the utterance due to an upward shift of the baseline, it appears that the listener also expects a small frequency range near the end of the utterance. Further experiments are required to find out how these results generalize to utterances with three accents and to varying time intervals between the accented syllables. On the basis of these additional experiments, perceptual tolerances can be established for F0 variation, and questions can be anwered with respect to the communcative function of relative differences in prosodic prominence.

4 CONCLUSION

On the production side, prominence differences appear to be associated both with pragmatic factors affecting the information status of focal expressions, and with lexical and phonological factors affecting the relative prominence of accented words within focal expressions. In order to account for the pragmatic factors, the notion "information structure" must be defined in terms of a hierarchical ordering instead of the binary "Given/New" distinction.

On the side of the listener, it seems unlikely that this hierarchical ordering can be recovered from the speech signal on the basis of differences in prominence. Here, prominence differences appear to have primarily a demarcative function. Only in relatively simple sentences, may prominence differences help the listener to recover intended differences in relative importance.

References

Allen, J., Hunnicutt, M.S. & Klatt, D. (1987). *From text to speech: the MITalk system*, Cambridge U.P.

Asher, N. & Wada, H. (1988). A computational account of syntactic, semantic and discourse principles for anaphora resolution. *Journ. of Semantics*, **6**, 309-344.

Brown, G. (1983). Prosodic structure and the Given/New distinction. *Prosody: Models and Measurements*, eds. A. Cutler & R.D. Ladd, Springer Verlag, Berlin, 67-77.

Chafe, W.L. (1974). Language and consciousness. *Language* **50**, 111-133.

Hajicova, E., Kubon, P. & Kubon, V. (1990). Hierarchy of salience and discourse analysis and production. *Coling 90, vol. III*, ed. H. Karlgren, 144-148.

Halliday, M.A.K. (1967). Notes on contrastivity and theme II. *Journal of Linguistics* **3**, 199-244.

Kruyt, J.G. (1985). *Accents from speakers to listeners. An experimental study of the production and perception of accent patterns in Dutch*, Unpubl. Doct. Diss, Leyden.

Liberman, M. and Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. *Language sound and Structure*, eds. M. Aronoff and R. Oehrle, MIT Press, Cambridge, 157 - 233.

Lieberman, P. (1965). On the acoustic basis of the perception of intonation by linguists, *Word* **21**, 40-54.

Monaghan, A.I.C. (1988). Generating intonation in the absence of essential information. *Proceeding of Speech '88, Seventh FASE Symposium* eds. W.A. Ainsworth and J.N. Holmes, 1249-1256.

Needham, W.P. (1990). Semantic structure, information structure and intonation in discourse production. *Journ. of Memory and Language* **29**, 455-468.

Nishinuma, Y. & Duez, D. (1989). Perceptual optimization of syllable duration in short French sentences. *Proceedings of European Conference on Speech Communication and Technology, Eurospeech 89*, eds. J.P. Tubach & J.J. Mariani, Vol. II, 694-697.

Terken, J. (1985). *Use and Function of Accentuation. Some Experiments.* Unpublished doctoral dissertation, University of Leyden.

Terken, J. (1991a). Fundamental frequency and perceived prominence of accented syllables. *JASA*, in press.

Terken, J. (1991b). Synthesizing natural sounding intonation for Dutch: rules and perceptual evaluation. *Computer Speech and Language*, submitted.

Wells, W.G.H. (1986). An experimental approach to the interpretation of focus in spoken English. *Intonation in Discourse*, ed. C. Johns-Lewis, Croom Helm, London, 53-75.

SOME GENERAL REMARKS ON DESIGNING LINGUISTIC MODELS OF INTONATION

Robert Bannert

Institute of Linguistics, Stockholm, Sweden

ABSTRACT
The main purpose of this contribution is to collect and put to discussion some general thoughts and ideas for the modelling of intonation. Two different aspects of this approach may be discerned: An interior aspect concerning the structuring and the elements of intonation models proper and an exterior aspect which focusses on intonation in a wider linguistic and non-linguistic context. Some notes of the perspectives on future research will round up my reflections.

1. INTRODUCTION
Surveying the development of linguistic intonation research of the last decades, one is somewhat surprised by the number and scope of all the research efforts. This is true of such areas as phonetics, general linguistics, psychology, and psychoacoustics.

Within the field of phonetics, a great number of descriptions of the Fo-movements in sentences of various languages have seen the light, and usually they are claimed to be models of intonation. In most cases, intonation equals the movements of fundamental frequency (Fo) or the succession of tones (pitch) throughout a linguistic utterance. As a rule, the linguistic specification of the input is not highly elaborated. Considering the universal dimension of human language, however, the largely differing and seemingly contradictory phonetic structure of the models is quite surprising. This is particularly true of languages that are very closely related to each other. The question arises of course if this diversity in modelling really reflects linguistic variation. We know very well, though, that variation of Fo represents a rather simple system: for linguistic purposes, Fo either rises or falls within different ranges in certain segments and syllables. There may be different reasons for this variety of description which I do not want to discuss here. For a general perspective, I would like just to put forth some remarks. First, I will consider some interior aspects of intonation models and will commence with the area of phonetics proper. Second, I will turn to some exterior, linguistic and non-linguistic aspects. They should altogether be considered important because we ought not lose sight of the total range of our endeavours, namely studying language as the most significant and effective means of man's communication for social interaction in a given situation.

2. ASPECTS OF INTONATION MODELS
Starting from the numerous and diverse descriptions of intonation in the literature, some comments on the models will touch their structure and their parts.

2.1. Levels of description
No explication is required in order to understand why so many of the descriptions of intonation pertain to production, e.g. are speaker oriented. Given the Fo analysed rather easily from the acoustic signal, it appears relatively uncomplicated to put into rules the different movements of Fo as a consequence of linguistic parameters systematically varied and thus, step by step, generating an Fo-curve. Another motive for this approach is to be found in the synthesis-by-rule of a given language. First, synthetic speech had to

be made more natural and second, the method of rule synthesis itself in an interactive approach makes it possible to assess the formulation of the rules directly. This method of analysis-by-synthesis has also turned out very profitable for the development of intonation models.

Less numerous are listener oriented intonation models. It is quite understandable that models of production aim at generating the total variation of Fo - possibly undistorted by the influences of microprosody - throughout an utterance. Opposed to this, models of perception explicitly aim at the essentiel parts of the Fo-curve that are necessary and sufficient for the listener to identify some linguistic features. It might not be completely correct to maintain that listener oriented models of intonation are more important, revealing or simply more interesting than speaker oriented ones. However, we know very well that the Fo found in the signal and produced by the speaker shows a large amount of redundancy. Furthermore, it remains to be seen whether Fo alone, in fact, carries the responsability for the perception of melody of language or in what ways Fo interacts with other acoustic parameters in speech perception.

2.2. Macrointonation
Apart from some early attempts to describe intonation in continuous speech, most of the intonation models of the preceding years are based on the sentence representing the unit of analysis and description. This may, however, not only be considered a reflex of the corresponding unit of analysis of general linguistics of the last decades. From a methodological point of view, it appears quite advisable, especially when obeying the demand for a rigorous control of the variables chosen, to start investigating rather simple and manageable units. From this follows almost of necessity that test sentences, under the given conditions, cannot be spoken freely as in natural verbal communication but are read aloud from sheets of paper in the laboratory. Everybody who has investigated intonation (and prosody in general) has found themselves in a situation where some informants had to be discarded because they were simply not able to produce the test sentences under the various conditions demanded.

Due to the limitation to one sentence (often syntactically and semantically simple), it appeared relatively easy to visualize the range of variation of Fo throughout an utterance using geometric means such as straight lines: base line and top line with certain declinations and focal lines, thus the tonal grid. Among the intonation models of hierarchical representation, however, one version is to be found which manages without any geometrical devices such as straight lines or higher functions. At the very beginning of the process of generating intonation step by step, the range of variation of Fo throughout the utterance is defined by Fo-levels that are related to significant points of the utterance. The dvnamic nature of Fo-movements or pitch variation, at least in German, is expressed by a formula including the contribution of various phonetic and linguistic features.

The attempts to write rules and to present visually the intonation of sentences syntactically more complex and small texts lead to the development of the geometric framework of lines: direction and range of the grid were allowed to vary in accordance with varying syntactic, semantic, and pragmatic conditions. Future research with spontaneous speech will show us perhaps that the phenomenon of uniform declination may be a typical feature of sentences and prose read aloud. It may perhaps also teach us that models based on automatic and stereotype declension of Fo will have to be basically restructured.

As general linguists have turned to discourse, the general interest of many phoneticians, including prosodists, has turned to spontaneous, natural speech. Therefore, in a future not too far remote, we will, quite sure, learn more about the gross structuring of naturally spoken language. On this new route, the question concerning the tonal units larger than the stress group, e.g.

macrointonation (and macroprosody), will represent a concern of central significance.

2.3. Parts of the model: tonal elements

Intonation models contain linguistically and prosodically motivated basic elements. Reviewing the models, many questions arise with reference to the various parts of the models: What is the theoretical status of (word-) accent and focus accent? Wherein lies the difference? Are there languages without (word-)accent and focus accent? Does the sentence have a tonal component of its own as expressed in the hierarchically designed models? What is the status of sentence final and non-sentence final phrase boundary tones? What do the macroprosodic units actually look like in their tonal dimension? Are emphasis and contrast to be considered independent unities? What are the projections of feelings and attitudes onto Fo? How can we define these tonal elements, what function do they have, how can we specify them in the linguistic structure of the input? An approach which draws a sharp line of demarcation between universal and language specific units should also, and just for these questions, give significant and new insights. For just this respect, equal and different features in the models ought to be discerned clearly.

2.4. Aims and methods of description

It may easily be seen that intonation models are qualitatively and quantitatively different in their explicitness, they may be exclusively pure intonation models or they may be integrated in a comprehensive model of prosody. Generative production models of intonation are best understood as instructions to a machine to derive, step by step, the matching Fo-curve for a given sentence with clearly defined parameters. Seen in this perspective, this method is excellently suited, as mentioned earlier, to test elements and steps of the intonation algorithm by the direct feedback of the produced speech signal. Intonation models, however, as all models of (parts of) man's verbal communication, also have a broader aim. At best, the intonation model will reflect the manifold processes in man, at the same time speaker and listener, that are active during linguistic communication on the psychological, neurological, physiological etc. levels. Such a model would contribute greatly to the understanding of human communication; it might also be able to explain, in a natural way, why intonation looks and functions exactly as it does, in universal and language specific respect, as it is observed and described in all its variation and, at the same time, its systematicity.

There is some evidence from different sources today that it may indeed be justified to posit intonation (and prosody) at a primary level of language processing models. Rather than assigning a proper Fo-curve to a fully developed segmental sentence structure late in a linear model, prosodic and tonal structures, at a highly abstract and basic level, should be established in a parallal processing model. As a consequence of this, individual segments should be elaborated at a more peripheral level. Thus in a framework of basic prosodic macrostructures, phenomena like tempo, accent distribution (deaccentuation), focussing, emphasis etc., segmental reductions of time and spectrum, and omissions would be accounted for in a quite natural way.

3. ASPECTS OF LINGUISTIC INTONATION

Even if, for different reasons, Fo of linguistic utterances is modelled in isolation, very soon a stage will be reached where more than Fo alone is called for. By now it appears as a matter of fact that prosody, and especially intonation, is strongly intermingled with other parts and components of communication. Intonation, in fact, seems to exert a key function in communication. This is especially true in direct relation to semantics and pragmatics with respect to focussing.

3.1. Interdependence of intonation and other parts of language

Today we have realized that there are

strong relationships between intonation, considered as an independent structure of utterances, and syntax, morphology (word structure) and semantics, and furthermore, pragmatics. As a good challenge, applying the approach of bridging gaps between disciplines, the different linguistic threads could be tied together. Within the field of phonetics, a first and preliminary attempt has been made to look at the interplay of the temporal and tonal components of a prosody model consisting of just these two dimensions. As an outlook, it has been proposed to widen the perspective of the phonetic realm of prosody by also treating, besides the dimensions of time and Fo, voice quality and intensity in some kind of parallel processing of bottom-up and top-down information.

This aim appears almost overwhelming when all aspects of intonation and all its linguistic relationships are considered in spontaneous speech and in discourse. However, as intonation does function in this verbal context, the description and the explanation of the linguistic totality, nevertheless, has to be realized as our utmost and long-standing aim.

3.2. Interdependencies of intonation and the outer world

One aspect of intonation that up to now has not obviously been investigated within the framework of an integrated model is its relationship to the world outside the linguistic code. From this outer world we can observe the expression of various non-linguistic features such as e.g. joy, vexation, anger; refusal, persuasion; intimacy etc. Furthermore, when these paralinguistic features are concerned, pragmatics, psychology, and social relations between interlocutors come into play.

When intonation research has progressed that far and has successfully produced knowledge of this whole area, we, as phoneticians, will then have approached our real aim: the modelling of linguistic communication, including intonation, as the expression of man's social being and comprehension.

4. SOME PERSPECTIVES OF FUTURE RESEARCH

Fortunately it can be stated that intonation and prosody are receiving more and more significance. Last not least, this is shown by the large number of contributions concerning prosody to this congress. Given this platform, will it at present be able to say something about the future of intonation (and prosodic) research? To my opinion, some lines of development seem quite obvious and some points of discussion may be suggested.

Designers of intonation models of different languages will soon discover the necessity and appropriateness to try to reach a more uniform terminology and manner of construction of their models. In order to achieve this, a phase of comparative consideration within a typological perspective might bring fruitful results. In spite of the existence of various schools of describing intonation with their seemingly different theories and systems, it should be possible, from a universal point of view, to treat, among others, accent and tone languages within a common framework of intonation.

Further important steps in the development of the actual intonation research, I suppose, are the concentration on the perceptual aspect of intonation and, of course, the investigation proper of intonation in natural discourse. The question concerning macrointonational units in spontaneous speech will be of prominent significance because these units constitute the basic elements of a good model.

As intonation is closely interrelated with various linguistic and non-linguistic areas and these areas, by their form of organisation, belong to different disciplines, it will become of outstanding significance that phoneticians will learn to cooperate with researches of intonation of other areas. Well-meant attempts lately have also shown that such an integrated approach will require mutual good will, patience, and also skill and aptitude from all people involved.

INTONATION PARAMETERS IN PRODUCTION AND PERCEPTION

Eva Gårding

Lund University, Sweden

ABSTRACT

My contribution is a discussion of perceptual aspects of the acoustic parameters of a Swedish intonation model which has been applied to several prosodic systems [5,7,20]. I emphasize the importance of perceptual analysis and close with examples showing how a phonological analysis may depend on the descriptive level used by the analyst.1)

1. A PRODUCTION-ORIENTED INTONATION MODEL

The model for Swedish intonation that I have worked with is oriented towards production. One goal is to reproduce the intonation of an utterance in the form of an Fo curve for a temporally structured string of words with labels representing the lexically distinctive accents (Accent 1 and Accent 2), phrase and sentence level accents, and labels indicating phrasing, sentence type and dialect. The generative scheme and its rules are based on analyses of materials chosen to explore the effect of the main communicative functions of intonation on the Fo curve.2) Superposition is a guiding principle in the analysis which is reflected in the generative scheme. This means that a local movement pertaining to an accent or tone is seen as added to a slow movement (global or semiglobal) representing sentence or phrase intonation [3,7].

The model makes use of the following parameters, which are all visual correlates of acoustic events. Fig.1.

1. *Turning points* are the local maxima or minima of Fo which are related to an accented syllable in a systematic way. The timing of these points, which is largely 'fixed' to specified syllables in different intonations, is crucial for the distinction between the accents and typical for a particular dialect.
2. The *tonal grid* is the auxiliary parallel lines enclosing curve pieces with a uniform general direction (see figure). Drawn in this way the upper line of the grid passes through local maxima of the curve (marked by H(igh) in the figure) and the lower line through local minima (marked by L(ow). In practice not all the local maxima or minima reach the grid lines, which is a tonal reflex of the varying accentuation levels carried by the syllables enclosed by the same grid. In synthesis the grid is generated before the accent target points.
3. *Pivots* are those places in the curve where the grid changes direction or range. They usually occur in connection with focus and demark constituents in the global contour.

Interpolation is used in the generative program to fill in the blanks of the Fo pattern between the target turning points.

This model, which captures the gross features of Swedish intonation, makes it possible to describe the accent contrast as a matter of *timing* the Fo target points in relation to to the accented syllable. The dialectal variation in intonation becomes a result of different timing rules for word accent as well as sentence accent (nuclear stress). Modality and weighting are expressed by the tonal grid.

With this description the common feature of the relation between the accents in various dialects is that the accentual movement of A2 comes later than that of A1. Interpreted at a physiological level, the underlying accentual gesture is similar but timed differently in relation to the accented syllable. This interpretation has support in EMG data obtained from pitch controlling muscles [10].

The feature 'delayed peak' for accentual timing differences in English has been suggested by Ladd [17] and plays a role in the intonation analysis presented for German by Kohler [16].

2. PERCEPTION, A MISSING LINK

The explanatory value of the peak analysis is limited to the production side of speech. For Swedish it implies that speakers use their laryngeal muscles in a similar way to raise and lower pitch in order to produce a given tonal pattern and that the Swedish accents can be produced by timing a similar underlying laryngeal gesture in different ways in relation to the accent carrying syllable. The speaker's intention, on the other hand, is to reproduce tonal and rhythmical patterns and these patterns should also be described at a level that relates to perception. Such a description is important for the understanding of perception mechanisms in general, for the understanding of how prosodic patterns are acquired, learned and transferred to a foreign dialect or language, and for a comprehensive theory of intonation.

The following discussion of the possible perceptual relevance of the acoustic parameters described above are inspired by ongoing research, planned to explore perceptual correlates of intonation and accentuation.

1. *Turning points.* A given pitch curve can be efficiently reconstructed either from information about the position in time and frequency of the turning points or from suitable information about the falls and rises that make up the curve. Some of the turning points have very stable locations relative to the C/V boundary of the accented syllable which means that also the following falls and rises are relatively stable. Physiologically speaking a turning point indicates an onset-offset change in a nerve signal and like other transient signals this ought to have a strong attention-sharpening effect. It may correspond to what a neurologist would call a feature-trigger [e.g. Granit 13].

If we aim at a perceptual description of intonation, the falls and rises (or contrasts between two frequency levels) are more important than the turning points. Only a phonetician who has worked long enough with acoustic records can perceive the position of a turning point but almost anybody can hear a rise or fall if it is given the proper duration.

An experiment in which Fo peaks of different shapes were shifted in small steps in time over constant carriers has shown that a prosodic phrase pattern can be recognized when the position of the peak has produced patterns in which pitch movements over neighbouring vowels are prototypical. Our experiment suggests that an Fo peak is only indirectly important. It is the adjoining ramps over the vowels which have perceptual reality [8, 9].

2. *The tonal grid*, which may be level or exhibit various degrees of fall and rise, encloses a phrase pattern made up of syllables with varying degrees of accentuation. The fact that the Fo movements relative to the grid are retained in different intonations is seen as a consequence of the superposition principle, as exemplfied in Fig.2. The communicative value of the direction of the grid may be explained in the following way: the auditory system makes a running average of the Fo values of the phrase and these values are retained in memory and matched with stored prototypes.

Since a listener is capable of identifying an accentuation pattern in different intonations, it is bound to have some invariant features. Our recent work with contrastive Swedish accentuation patterns confirms that their relative patterns of duration, intensity and Fo are retained in different intonations. In particular, pitch relations between neighbouring vowels are perceptually important. In a given intonation an accentuation pattern is characterized by its typical pitch relations over neighbouring vowels which sets it off from other patterns. (Fig.3.)

From a perceptual point of view, then, intonation, represented by the grid, carries relative invariant features of an accentuation pattern which recur in a global rising, falling or level Fo pattern with different communicative values.

3. The perceptual importance of the *pivots* stems from a discontinuity of direction or range of the global Fo pattern. Just like the turning points the pivots are not directly perceptual units but may serve as attention getters. The new direction or range is the important signal. In Swedish the position of the main pivotal point of an Fo curve in relation to the text is dialect dependent. In South Swedish declaratives the main pivot occurs on the accented syllable of the focussed word and in Central Swedish dialects it has a position after the accented word, described as 'floating' by Bruce [2].

The *interpolation* follows a mechanical rule which has the purpose of connecting two target points with each other. It has been interesting to note in ILS experiments that this interpolation can neglect small-scale pitch obtrusions caused by the accents without any noticeable effect on perception [12]. In this way, a description of the curve in terms of perceived local Highs and Lows may differ from a description relying rigorously on production data. Our phonological interpretation of the ILS experiments is that the accent domain may encompass not only the following unaccented syllables up to the next accent (which is a current definition) but also the deaccented ones.

3. REPRESENTATION

Highs and Lows may be used to show for a given Fo curve how local maxima and minima relate to the text and to the surrounding highs and lows. For Swedish, this kind of analysis applied to the Stockholm dialect may yield High for A2 and Low for A1, marking a high or low turning point near the beginning of the vocalic segment as the distinctive feature [2]. A phonetic rule would generate the following pitch movement. The representation makes no claim to perceptual validity.

In an intonation analysis of Hausa (traditionally analysed as having two tones, High and Low), Lindau showed that a 'high' tone is manifested as a rise to a high point at the end of the syllable and a 'low' tone as a corresponding fall to low [18]. This finding suggested that it is the latter part of the syllable that has perceptual salience rather than the earlier part. Further support comes from perceptual experiments with Chinese [11] and with Swedish nonsense syllables [15].

4. CONCLUDING REMARK AND QUESTION

Our discussion leads to the following conclusion:

Parameters and their representations differ depending on the orientation of the model and the descriptive level chosen by the analyst.

It also leads to an interesting question: What side of speech should the phonological units represent, production, perception, both or neither?

1) The necessity of distinguishing between levels of description in speech analysis, has been strongly argued for by Repp [19].

2) In classical tradition we have considered the lexical-distinctive function,the grouping function with demarcative and connective features and the weighting, modal and expressive functions.

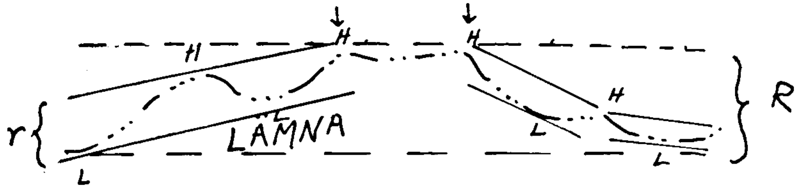


Fig.1. **Parameters in the production-oriented model.** Fo curve of *Man vill LÅMNA* (A2) *nåra långa* (A2) *nunnor* (A2) in declarative intonation (One wants to deliver some long nunnies). A2 is manifested as HLH turning points in focus, after focus as HL. Parallel lines represent the grid. 'R' denotes global range, 'r' is the vertical distance between the grid lines. The curve is from Bruce [1].

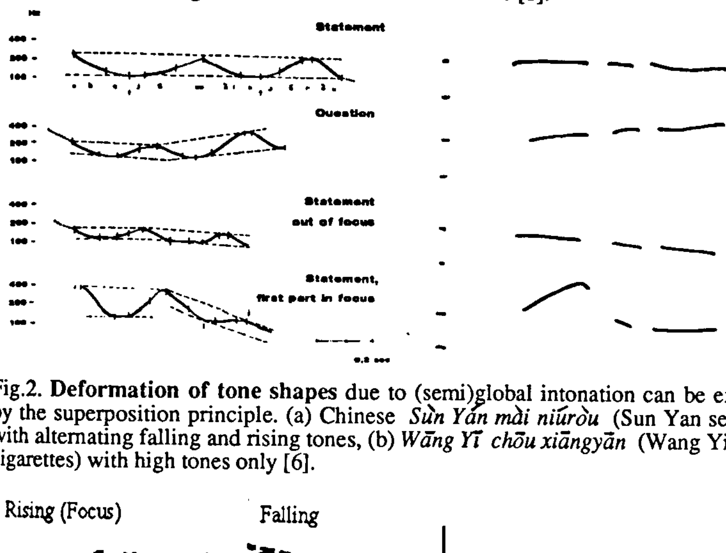


Fig.2. **Deformation of tone shapes due to (semi)global intonation can be explained by the superposition principle.** (a) Chinese *Sǔn Yán mài niúròu* (Sun Yan sells beef) with alternating falling and rising tones, (b) *Wāng Yī chōu xiāngyān* (Wang Yi smokes cigarettes) with high tones only [6].

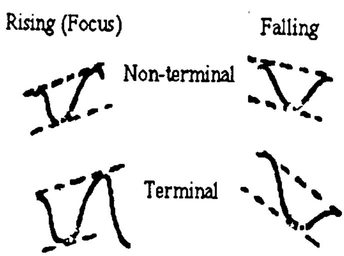


Fig.2(c).**Deformation of accent shapes.** Swedish *bara* A2 ('only' and placename) in rising and falling intonation. [4].

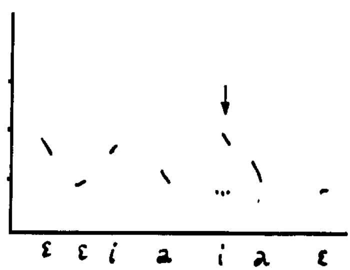


Fig.3. **Perceptual importance of intersyllabic relations.** Pitch curve in maximally unvoiced context. *Ester* (A1) *fick FATT*(A1) *i katten*(A1) (Ester got hold of the cat). Lowering the post-accented syllable (arrow) shifts categorization from A1 to an A2 compound (fatti(g)katten(A2) [11].

A NOTE ON INTONATION MODELS

Intonation models differ in scope and orientation. Moreover, to list some of the qualifying labels that have been used, they may be qualitative or quantitative, descriptive or generative, abstract or concrete. Analyses may depart from different theoretical assumptions (sometimes biassed by the analyst's native prosody!) and rely on different kinds of observations and measurements, described and interpreted at different levels in speech production or speech perception.

There is general agreement on the close correlation between intonation and Fo but there is disagreement on the best decomposition of a pitch curve, horizontally as well as vertically.

Evaluation of models seems to be simpler, at least if we adhere to the following criteria: A good model should be descriptivly adequate, general, parsimonious and have explanatory power.

5. REFERENCES

[1] BRUCE, G. (1977), Swedish word accents in sentence perspective. *Travaux de l'Institut de Linguistique de Lund XII*. Lund:Gleerup

[2] BRUCE, G. (1986), How floating is focal accent? *Nordic Prosody IV*, eds K.Gregesen & H. Basböll, 41-49. Odense University press.

[3] BRUCE, G. & GÅRDING E., (1978). A prosodic typology for Swedish dialects. *Nordic Prosody*, eds. E. Gårding, G. Bruce & R. Bannert, 219-228. Department of Linguistics. Lund University.

[4] GÅRDING, E. (1967), Prosodiska drag i spontant och uppläst tal, *Svenskt talspråk*, ed. G. Holm, 40-85. Stockholm: Almqvist & Wicksell.

[5] GÅRDING, E. (1983), A generative model of intonation, *Prosody: Models and measurements*, eds. A. Cutler & R. Ladd, 11-21. Springer.

[6] GÅRDING E. (1987), Speech act and tonal pattern in Standard Chinese: Constancy and variation. *Phonetica* 44:1, 13-29.

[7] GÅRDING, E. (1989), Intonation in Swedish, *Working Papers* 35, 63-88. Contribution to *Intonation Systems*, eds. D. Hirst & A. Di Cristo.

[8] GÅRDING, E. & ERIKSSON L. (1989), Perceptual cues to some Swedish prosodic phrase patterns - a peak shift experiment. Fonetik 89. *STL-QPSR* 1/1989, 13-16.

[9] GÅRDING, E. & ERIKSSON L. (1991), On the perception of prosodic phrase patterns. To be published.

[10] GÅRDING, E., FUJIMURA O., HIROSE H. & SIMADA Z. Laryngeal control of Swedish word accents. *Working Papers* 10, 53-82. Dept of Linguistics, Lund University.

[11] GÅRDING, E., KRATOCHVIL, P., SVANTESSON, J.O. & ZHANG J.(1986),Tone 4 and Tone 3 discrimination in modern standard Chinese. *Language and Speech* 29:3, 281-293.

[12] GÅRDING, E. & STENBERG, M. West Swedish and East Norwegian intonation. *Nordic Prosody V*, eds. K. Wiik & I. Raimu, 111-131, Turku.

[13] GRANIT, R. (1977) The purposive brain.The MIT Press.

[14] HIRST, D. & DI CRISTO, A. (1991), Intonation systems. To be published.

[15] HOUSE, D. (1990), Tonal perception in speech. *Travaux de l'Institut de Linguistique de Lund* 24. Lund University Press.

[16] KOHLER, K. (1987), Categorical pitch perception. *Proceedings of the 11th Congress of Phonetic Sciences*, 5:331-333, Tallinn: Academy of Sciences of the Estonian S.S.R.

[17] LADD, R. (1983), Phonological features of phonological peaks. *Language* 59, 721-59.

[18] LINDAU, M. (1986) Testing a model of intonation in a tone language. *J.A.S.A.* 80: 757-764

[19] REPP, B. (1981) On levels of description in speech research. *J.A.S.A.* 5: 1462-64.

[20] TOUATI, P. (1987), Structures prosodiques du suédois et du francais. *Travaux de l'Institut de Linguistique de Lund* 21. Lund University Press.

(For further references see [7])

INTONATION MODELS: TOWARDS A THIRD GENERATION.

Daniel Hirst

Institut de Phonétique d'Aix,
URA CNRS 261 Parole et langage
Aix en Provence, France

ABSTRACT

Intonation models up to now can be classified into two generations : single language descriptions and multi-language models. The next step will be the development of an integrated theory of intonation defining a number of independent levels of representation together with a specification of the relationship between these different levels. A sketch of a framework for such a theory is given as well as a number of questions we need to ask about the nature of the different levels of representation.

1. INTRODUCTION

All linguistic description is faced with the challenge of sorting out those facts which can be put down to some universal language faculty from those which are assumed to be specific to a given language. How do we decide, in other words, which facts are to be incorporated into a general model of language and which are to be analysed as language specific parameters of the model ?

The problem is particularly acute in the study of intonation which appears to be one of the most universal characteristics of human language and, paradoxically, at the same time one of the most specific characteristics of a given language or dialect.[17] The universal character of intonation is well established. A striking illustration of its language specific nature can be seen from the recent finding [24 and references there] that as early as 4 days after birth, babies appear to be capable of distinguishing recordings of their native language from recordings of other languages. The fact that similar results are obtained with low-pass filtered recordings but not with recordings played backwards suggests that such discrimination is based on prosodic information which can only be acquired during the pre-natal period.

I suggest in my contribution to this symposium that we can distinguish two generations so far in the history of intonation models, and that our knowledge of intonation has now reached the point where we can begin to envisage a model belonging to a third generation. In section 3, I outline what appear to me some of the desirable characteristics of such a model.

2. INTONATION MODELS.

2.1 Generation one: single language descriptions.

The first generation of intonation models consists of phonetic and/or acoustic descriptions of the intonation of particular languages. Probably the vast majority of research which has been carried out on intonation falls under this heading. Since such descriptions concentrate on the intonation of a single language, no principled distinction can be made between model and parameters. These descriptions, however, constitute the indispensable groundwork on which more general models can be built.

2.2 Generation two: multi-language models.

In the last fifteen years or so, a number of specific models have been proposed for the description of the intonation of several languages. Examples of models of this type are those which have been developped in Lund [10, 11] and Eindhoven [12, 13] as well as work in the generative phonology paradigm [28, 27, 14, 15, 21, 29].
These models differ from those of what I have called the first generation in that they explicitly tackle the problem of separating out universal characteristics of intonation systems, which are directly incorporated into the model itself, from language specific characteristics which constitute the parameters of the model. In addition, second generation models have a number of common characteristics. Firstly the models all incorporate a number of constructs (**tonal grid, declination line, focus marker, boundary tone** etc.) specifically designed for the description of intonation: Secondly, the various models are generally oriented towards a specific descriptive level - phonological, physiological, perceptual etc. Finally, the descriptions proposed within the framework of the various models are in general underdetermined by the data. Given a model and a set of data, it often seems possible to account for the data in more than one way by choosing a different set of parameter values.

2.3 Generation three: ?

What would a third generation model look like ? The major distinction between such a model and those of the second generation would be the existence of general principles constraining the model and its applications. The primitive constructs of the model should thus be determined as far as possible by more general linguistic principles and the choice of parameters for a description should be fixed on the basis of a limited number of empirical questions which can be asked about the intonation system of a language. A complete third-generation model of intonation would, moreover, not be restricted to a particular descriptive level but should provide a number of different levels of representation, including at least phonological, phonetic and acoustic levels.

An explicit characterisation of these different levels of representation will be crucial to the development of a coherent body of theory. The following remarks raise a certain number of questions concerning these representations together with some very tentative answers.

3. LEVELS OF INTONATIONAL REPRESENTATIONS.

Few linguists today would question the fact that intonation forms part of a speaker's overall cognitive representation of an utterance. Such a high-level abstract representation can be assumed to contribute both to the overall meaning of an utterance and to the way in which the utterance is pronounced. At the other extreme, the acoustic parameters of fundamental frequency and intensity, together with the slightly more abstract parameter of segmental duration, constitute the physical dimensions in which intonation can be expressed. Between these two extremes - phonology and acoustics - lies the whole field of phonetics. Ohala has recently argued [25] that "There is no interface between between phonology and phonetics". Di Cristo and I have suggested [17] that this is because phonetics is itself an interface between the cognitive (what can be thought) and the physical (what can be said).

3.1 Phonetic representations

Given the hybrid nature of phonetics as the link between the cognitive and the physical, it would seem to follow logically that a phonetic representation can only be defined on the basis of a prior theory of phonological representations. In fact, however, there is a constant interaction between phonology and phonetics. The more we learn about phonetic processes, the more we can incorporate into our phonetic model and the more we may wish to question traditional assumptions about phonological representations.

A number of techniques have been used for the generation of fundamental frequency for speech synthesis. Many of these techniques such as contour concatenation or neural networks leave unanswered the question of the nature of phonological representations. There are, however, at present at least two plausible candidates for a phonetic model of fundamental frequency curves.

The first of these, sometimes called the **Target and Transition Model,** assumes that an Fo curve is represented as a sequence of target points and that a (generally monotonic) interpolation function accounts for the portions of the curve between these target points [9, 10, 14, 15].

The second type which I shall call the **Pitch-Pulse Model** represents an Fo contour not as a sequence of pitch targets but rather as a sequence of instructions to raise or lower the pitch, the local pitch-movement being generally superposed on a more global movement. [26, 6]

It is obvious that at some level of abstraction, the two models are formally equivalent - any curve that can be generated as a sequence of targets can also be generated as a sequence of movements and vice versa.

On the level of speech production it has been argued [6] that a pitch-pulse model more adequately represents the actual physical mechanism underlying the generation of pitch contours in natural speech. Even this, however, does not guarantee that pitch contours are mentally represented as pitch changes. It is possible for example that a motor control mechanism such as a Forward Model [23, 1] is developped during the babbling stage, providing the speaker with a direct mental mapping between pitch targets and the impulses needed to generate the targets.

Evidence from acoustic modelling suggests that a model incorporating pitch targets more adequately accounts for the observed data than one generating pitch changes: in an experiment using controlled sentences [22] , relative peak levels were observed to be more highly correlated than were the corresponding rises or falls.

It has been claimed [4, 10, 17] that intonation systems generally make use of two distinct types of pitch levels : **relative** levels, determined with reference to the preceding pitch level, and **absolute** levels determined with reference to a wider context, perhaps even to an absolute speaker dependent value. While relative levels are easily coded in terms of pitch movements, absolute levels are less easily expressed in this way [27]

On the perceptual level, it has been claimed recently [20] that it is only during areas of spectral stability, as in a sustained vowel, that pitch patterns are interpreted as movement configurations rather than as pitch levels, but that elsewhere interpretation in terms of pitch levels is predominant.

It seems that much of the evidence points in favour of a mental representation of a pitch contour as a sequence of pitch targets even if this is not necessarily the form which serves as input to the pitch-producing mechanism in natural speech.

3.2 Surface phonological representations.

The surface phonological representation of a pitch curve can be assumed to consist of a sequence of phonetically interpretable symbols which can in turn be derived from a more abstract phonological representation. This would in many ways be the equivalent for intonation of the IPA transcription system for segmental phonology. An example of a first approximation of such a system is **INTSINT** [17], an INternational Transcription System for INTonation, which makes use of two types of symbols corresponding to the distinction between Absolute and Relative pitch levels mentioned above. Absolute pitch levels include :

↑ ↓
Top Bottom

as well as **Mid** which is assumed only to occur at the beginning of an Intonation Unit and is consequently unmarked. Relative pitch levels, defined with respect to the preceding pitch target include :

↗ ↘ ↱ ↳
Higher Lower Upstep Downstep

An application of this system to the F0 pattern of a continuous text in French [19] suggests the interesting result that all these targets can be defined with respect to three absolute speaker independent pitch levels corresponding to the speaker's mean Fo (Mid) and two levels (Top; Bottom) fixed at a half-octave interval respectively above and below the mean. It remains to be seen, however, whether similar results will be obtained for other speakers and other languages. It is also not clear how such a model can incorporate other factors such as variable pitch range [22].

3.3 Underlying phonological representations.

I suggested above that a surface phonological representation of intonation should have two characteristics. First, it should be phonetically interpretable, and secondly we should be able to derive it from a more abstract underlying phonological representation, the nature of which is obviously far from being clear. I have argued [15] that we should be guided by a general principle to the effect that all phonological primitives which are needed for the description of connected speech are needed for at least some languages to describe lexical contrasts. Thus for example stop aspiration is lexically distinctive in Hindi but not in English or French. English, however, unlike French, possesses a phonological rule adding aspiration to voiceless stops at the beginning of stressed syllables. The result is that in connected speech, aspiration can become distinctive in English (but not in French) as in the minimal pair : [ðɪspʰɔt] 'this port' ≠ [ðɪspɔt] 'this sport'.

If we compare the case of aspiration with that of word-stress we find a similar range of effects since word-stress is lexically distinctive in some languages (like Russian and English) but not in others (like French and Hungarian). The latter, however, contain rules assigning stress to the final (respectively initial) syllables of words so that stress can become distinctive in connected speech (cf minimal pairs in French : ['drapo'sɪbl] 'Drap possible' (possible sheet) ≠ [dra'po'sɪbl] 'Drapeau-cible' (target-flag). A similar argument can be made concerning tonal categories which, while they only need to be lexically specified for tone languages such as Chinese or Yoruba, remain available for the phonology of other languages in order to derive surface representations.

I mentioned above that as we learn more about phonetic processes we may well be led to modify our conception of underlying phonological representations. In the case of intonation, an explicit phonetic model suggests that tonal representations may be considerably sparser than in standard autosegmental accounts [7, 8]. I have suggested [14, 15, 16] that tonal segments are in fact linked not directly to vowels or syllables but rather to higher order phonological constituents as in Figure 1.

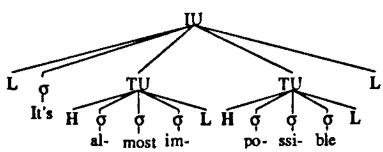


Figure 1 : Phonological representation for the sentence 'It's almost impossible'. **IU** = Intonation Unit, **TU** = Tonal Unit, σ = syllable, **H** = high tone, **L** = low tone.

A phonological structure of this type, removes the need for ad-hoc diacritic symbols indicating 'boundary tones' as proposed by Pierrehumbert [28]. Similar representations have since been proposed for a number of different languages including French [5], Swedish [2] and Japanese [27] as well as for an African tone language Kinyarwanda [3]. It seems probable that explicit acoustic modelling of other tone langauges could lead to similar results.

A representation of this type makes it possible to describe the variability found across the intonation patterns of different languages in terms of a small set of formal parameters. Thus if we compare the prosodic structure of English with that of French, we find that in English (as in Germanic languages in general), Tonal Units each contain one stressed syllable *followed* by an unlimited number of unstressed syllables, whereas in French (as probably in most Romance languages) Tonal Units contain one stressed syllable *preceded* by an unlimited number of stressed syllables. More formally Tonal Units can be described as *Left-Headed* in English and *Right-Headed* in French. The two languages also seem to differ in the sequence of tonal segments associated with Tonal Units in underlying representations. In English, each Tonal Unit contains the sequence [H L] as in Figure 1, whereas in French the underlying sequence seems to be [L H] [5]. Evidence from other languages [17] seems to suggest that the two parameters are independent.

The underlying representation in Figure 1 corresponds directly to the surface intonation pattern described as typical for many American English dialects which can be transcribed :

It's almost impossible

[⤒ ↘ ↗ ↓]

For most British English dialects, such a pattern is not at all typical. Instead a downstepping pattern is generally observed as in :

It's almost impossible

[⤒ ↳ ↓]

Both representations can be derived from the same underlying form illustrated in Figure 1 if we assume a further parameter which for British English allows just one single tonal segment to remain associated with each Tonal Unit except the last [15, 16].

The three parameters I have mentioned make it possible to generate an interesting variety of intonation patterns from otherwise similar underlying forms.

While it is obvious that the model I have sketched is far from constituting the Third Generation Model I evoked at the beginning of this discussion, I would argue that it possesses at least some of the characteristics.

4. CONCLUSIONS

I have presented a number of questions, together with some very tentative proposals concerning an integrated theory of intonation comprising several distinct levels of representation. My discussion of phonetic and acoustic representations has been exclusively concerned with fundamental frequency patterns although it is obvious that a complete description will need to account for variations of intensity and duration as well.

The aim of this presentation has been to sketch an overall framework such that at each step we can formulate our choice of representation in terms of formal parameters each of which can be determined by empirical investigation.

Although we are obviously still far from achieving such an integrated model of intonation, all of the models which I have somewhat disparagingly dubbed 'second generation' possess at least some of the desirable characteristics. Just as the accumulation of first-generation descriptions made it possible to develop such models in the late '70s, so the existence of such multi-language models will be the basis on which it will be possible to build a coherent general theory of intonation. The task for the '90s could thus be a concentration of collaborative (and/or competitive) work leading to the development of such a general theory.

REFERENCES

[1] BAILLY, G; BACH, M. LABOISSIERE, R.& OLESON, M. (1990) 'Generation of articulatory trajectories using sequential networks' *Proc. ESCA Workshop on Speech Synthesis* (Autrans), 67-70

[2] BRUCE, G. (1988) "How floating is focal accent ?" *Nordic Prosody* 4, 41-49

[3] CHAMBON, T. (1991) "Phonological interpretation of Fo variations in a Bantu language : Kinyarwanda".*Proc. Phon.Sci. XII* (vol. 2), 218-221

[4] CRYSTAL, D. (1975) 'Relative and absolute in intonation analysis' in *The English Tone of Voice* (Arnold, London), 74-83

[5] DI CRISTO, A. (in press) "Intonation in French". in Hirst & Di Cristo (eds) (in press).

[6] FUJISAKI, H. (1981), "Dynamic characteristics of voice fundamental frequency", *Proc. 8th FASE* (Venice)

[7] GOLDSMITH, J. (1976), *Autosegmental Phonology*, (Ph.D.thesis; MIT.)

[8] GOLDSMITH, J. (1990) *Autosegmental and Metrical Phonology.* (Blackwell, Oxford)

[9] GÅRDING, E. (1977) "The importance of turning points for the pitch patterns of Swedish accents.", in Hyman ed. (1977) *Studies in Stress and Accent.*, 27-35

[10] GÅRDING, E. (1983), "A generative model of intonation", in Cutler & Ladd (1983) *Prosody, Models and Measurements*, (Springer), 11-25

[11] GÅRDING, E. (in press), "Intonation in Swedish", in Hirst & Di Cristo (eds) (in press)

[12] 'T HART, J. & COHEN, A. (1973), "Intonation by rule : a perceptual quest.", *Journal of Phonetics* 1, 309-327

[13] 'T HART, J. COLLIER, R & COHEN, A.. (1990) *A Perceptual Study of Intonation: an Experimental-Phonetic Approach to Speech Melody*, (Cambridge University Press, Cambridge)

[14] HIRST, D.J. (1983) "Structures and categories in prosodic representations." in Cutler & Ladd (1983) *Prosody: Models and Measurements* (Springer, Berlin) 93-109

[15] HIRST, D.J. (1987) *La description linguistique des systèmes prosodiques : une approche cognitive.* (Thèse de Doctorat d'Etat, Université de Provence.)

[16] HIRST, D.J. (in press) "Intonation in British English" in Hirst & Di Cristo (eds.) (in press).

[17] HIRST, D.J. & DI CRISTO, A. (in press) "A survey of intonation systems." in Hirst & Di Cristo (eds). (in press)

[18] HIRST, D.J. & DI CRISTO, A. eds. (in press) *Intonation Systems : a Survey of Twenty Languages.* (Cambridge University Press; Cambridge)

[19] HIRST, D.J. NICOLAS, P. & ESPESSER, R. (1991) "Coding the F0 of a continuous text in French: an experimental approach." in *Proc. Phon.Sci. XII* (vol. 5), 234-237

[20] HOUSE, D. (1990) *Tonal Perception in Speech* (Lund University Press, Lund)

[21] LADD, D.R. (1986) "Intonational phrasing: the case for recursive prosodic structure." *Phonology Yearbook* 3, 311-340

[22] LIBERMAN, M & PIERREHUMBERT, J (1984) "Intonational invariance under changes in pitch range and length." in Aronoff & Oehrle (1984) *Language Sound Structure* 157-253

[23] LINDBLOM, B. LUBKER, J. & GAY, T. (1979) "Formant frequencies of some fixed-mandible vowels and a model of speech-motor programming by predictive simulation." *Journal of Phonetics* 7, 141-161.

[24] MEHLER, J. & DUPOUX, E (1990) *Naître Humain* (Seuil, Paris)

[25] OHALA, J.J. (199O) "There is no interface between phonology and phonetics : a personal view." *Journal of Phonetics*, 18, 153-171

[26] ÖHMAN, SVEN & LINDQVIST, J. (1966), "Analysis by synthesis of prosodic pitch contours.", *Proc.18th Int. Congress. of Psychology* (Moscow)

[27] PIERREHUMBERT, J.. & BECKMAN, M.. (1988), *Japanese Tone Structure.*, (MIT Press; Cambridge, Mass.)

[28] PIERREHUMBERT, J. (1980) *The Phonology and Phonetics of English Intonation.* PhD thesis; MIT.

[29] SELKIRK, E.(1984) *Phonology and Syntax : the Relation between Sound and Structure.* (MIT Press; Cambridge, Mass.)

THE ONSET OF CROSSLINGUISTIC DIFFERENCES IN SPEECH DEVELOPMENT

David Ingram

The University of British Columbia, Vancouver, Canada

ABSTRACT
Recent research indicates that the effects of the ambient language appear earlier than what was once believed. Here, two theories are discussed to account for this. One, the interaction theory, assigns this effect to the interaction between the child's perceptual and articulatory systems. The other, the NeoJakobson theory, makes the bolder claim that such effects reflect linguistic organization on the part of the child. Evidence for the latter is presented.

1. INTRODUCTION

The last twenty years have represented a tremendous increase in our knowledge about the abilities of infants to vocalize and perceive speech. There has been, however, a crossover concerning our knowledge in these two areas on infant development. Twenty years ago, I would have said that our knowledge of infant speech production was noticeably ahead that of infant speech perception. Based on the work of Irwin and others, we had a reasonable picture of the stages through which infants progress from cooing to babbling to the first words. Today, it looks like the greatest gains in our knowledge have taken place in our understanding of infant speech perception, inspired by the great interest generated through the methodological developments made in this area.

This is not to say that we have made no strides in the area of infant speech production. There has been the refinement of our knowledge of the stages of infant speech production, as seen in the important research of Oller and others (e.g. Oller [10]). Another development has been the initiation of important research on crosslinguistic influences on infant development. Today, however, I think it is fair to say that the latter work is only in its 'infancy', if you would excuse the pun.

In this paper, I will attempt to explain why I make this claim, and lay out what I see as the crucial issues which will influence research in infant speech production in the years ahead. I also hope that these remarks will

provide a useful backdrop for the other papers that have been prepared for this symposium on "Speech acquisition and development".

2. THE MATURATIONIST VIEW

Until quite recently, the dominant view of infant speech production has been what might be called a 'maturationist' view of development. This point of view, expressed by Lenneberg [7] and supported by data in Locke [8], sees the infant's speech as more or less controlled by a biologically determined sequence of development. Since this development is controlled by factors within the infant, there will not be any noticeable sign of the influence of the ambient language for a relatively long time. The following quote from Locke [8] (p. 84) captures this point of view quick succinctly:

"I will suggest that no genuine accommodations to the adult system will be evident until the child reaches the *systemic stage* of phonological acquisition, which probably occurs at some time after the first 50 words are in use".

There are at least two features of this proposal that need to be elaborated. One concerns the extent to which individual variation takes place. As argued in Locke [9], this does not mean that all children will vocalize in exactly the same way. Biological models of development still allow for variation. The critical point is that the variation that exists will be constant across linguistic environments.

A second aspect concerns the time at which crosslinguistic effects will appear. In the above quote, the fifty word stage is cited, but it appears that this is just an educated guess. There is nothing magical about this stage, and indeed, if we are dealing with a biological milestone, one would expect that age is more critical than stage. For example, suppose we were to compare two infants who are developing normally from all indications, except for language. We would anticipate that the infant who starts to speak at age three might show more adjustments to the ambient language than the child who starts at age one.

The issue of when crosslinguistic effects first appear is important for different reasons for different people. For the speech scientist, it is important in coming to understand the development of the speech apparatus. Further, because of the findings on the remarkable perceptual ability of infants, there is the question of how the perceptual system interacts with the articulatory system. If the latter is rather fixed by biological constraints, then its development is basically uninfluenced by the perceptual development taking place. I will refer to this issue as the 'perception-production issue'.

For the linguist, the question of when ambient effects begin is important for a different reason than just the perception-production issue. The linguist is more concerned with determining when the child has access to, and begins to construct, a linguistic system. For those who tend toward a maturationist point of view, much of the infant's early language, even up to the first 50 words as cited by Locke, is seen as prelinguistic. Such a position requires some discontinuity in development at some point when the infant shifts from a biologically based language to a more abstract and linguistically based one. I will refer to the question of when children begin to use linguistic principles as the 'linguistic issue'.

3. TWO ALTERNATIVE VIEWS

Most recently, at least two alternative positions have appeared to the one expressed by the maturationists. One point of view, referred to as the 'interactional hypothesis' in de Boysson-Bardies, Halle, Sagart & Durand [1], claims that the effects of the ambient language occur earlier than previously thought. This hypothesis has the following properties:

1. the early perceptual abilities of the infant enable it to show some effects of the ambient language at least during the later stages of the babbling period;
2. these effects are likely to be subtle at first, and may require more sophisticated analyses than previously done;
3. children's first 50 words will show crosslinguistic differences in their phonetic inventories.

The data for this position are primarily found in de Boysson-Bardies, Halle, Sagart & Durand [1] and in de Boysson-Bardies & Vihman [2]. In the former study, differences are found in the formant structure of vowels in the babbling of infants in French, Cantonese, English, and Arabic linguistic environments. The latter study expands indications of such differences through the study of consonantal patterns in French, English, Japanese, and Swedish infants.

The interactional hypothesis primarily focuses on the perception-production issue. As discussed in de Boysson-Bardies & Vihman [2], this interaction between perception and articulation takes place while the child is still by and large 'prelinguistic'. For example, they state "...we never assumed that selection on ambient language implied phonetic segmentation" (p. 17). Rather, they believe the following (p. 17):

"A segmentally unanalyzed acoustic representation may provide targets for a motor plan sufficient to initiate an epigenetic selection of articulatory gestures".

This point of view is one which I have referred to elsewhere as the Stanford theory (c.f. discussion in Ingram [4]). It sees the development of the first words as primarily devoid of any linguistic

organization into either phonemes or distinctive features.

An alternative to this belief is the opinion expressed years ago by Jakobson [6] that children show linguistic organization around the time of their first words. There are at least two aspects of Jakobson's theory which have been shown by more recent research to be incorrect. One was his conception of an abrupt shift in the infant's phonetic abilities once word acquisition appears. A number of recent studies have shown that this shift involves continuity rather than discontinuity (e.g. Vihman, Macken, Miller, Simmons & Miller [12]). The second error was that children of all linguistic environments show the same initial phonological system. As found in de Boysson-Bardies & Vihman [2], there are phonetic differences in infants in different linguistic environments from the onset of word acquisition.

Neither of these errors, however, directly negates Jakobson's primary claim that children use linguistic organization at the onset of word acquisition. My own research in this area (c.f. Ingram [5], for a summary) has maintained this aspect of Jakobson's original theory. I have referred to this position as a NeoJakobsonian point of view, since it retains the flavor of his ideas on this period, but rejects his claims about the transition from babbling to the first words. This point of view adds the following strong claim to the three already mentioned above:

4. crosslinguistic effects in children's phonetic inventories indicate that infants show linguistic organization of their words at the onset of acquisition, not at some later time in development.

These various points of view result in four positions about when effects of the ambient language appear and how what they indicate about the infant's linguistic abilities. These are summarized in Table 1.

Table 1. Four theories on infant phonological acquisition.
1. *Maturationist theory.* Infant babbling and the first words are determined biologically without linguistic processing or effects of the ambient language (e.g. Locke [9]).
2. *Interaction theory.* Infant babbling and the first words show effects of the linguistic environment, but the organization of this effect is prelinguistic.
3. *Jakobsonian theory.* The first words show linguistic organization but no effects of the ambient language.
4. *NeoJakobsonian theory.* The first words show linguistic organization and effects of the ambient language.

4. DIRECTION OF RESEARCH

The results of recent research of the sort being conducted by de Boysson-Bardies and her colleagues suggest that neither Maturationist theory nor Jakobsonian theory can be maintained. The unresolved issue is no longer when crosslinguistic effects appear but rather when they show linguistic organization. The resolution of this question requires detailed linguistic analyses of the first words of children across several linguistic environments.

One way in which this can be done is through the examination of the phonetic inventories of children acquiring different vocabularies. In several places, I have argued that children during the acquisition of the first fifty words or so acquire what I refer to as the basic phonological inventory. These are a basic set of consonants, vowels, and syllable types are used to determine the basic phonological features of the language.

Table 2 shows some preliminary results of such research from children in five linguistic environments. The English data are from Ingram [3], and the Quiche data from Pye, Ingram & List [11]). The other data are from unpublished data. The Dutch data are from Mieke Beers at the University of Amsterdam, and the Italian data are from Umberta Bortolini of the University of Padua. The French data are from my unpublished analyses of diary data.

Table 2. Basic consonantal inventories from children acquiring English, Quiche, Dutch, and Italian. (Capital letters are used to indicate alveopalatal sounds, e.g. S indicates the alveopalatal fricative).

English			*Quiche*				
m	n		m	n			
b	d	g	p	t	tS	k	?
p	t	k				x	
f	s	h	w	l			

Dutch			*Italian*			
m	n		m	n		
p	t	k	p	t	tS	k
	s	x	b	d	dZ	g
w		h	f	s	S	
				l		

French	
m	
p	t
b	d
f	s
	l

Since the data are very preliminary, and diverse in their analyses and collection procedures, it is of course necessary to be quite cautious in their interpretation. They suggest to me, however, that the extent of differences is extreme, certainly more than would be expected by the Interaction hypothesis discussed elsewhere.

There are of course similarities as might be expected. The data indicate that nasals are early, and that voiceless fricatives are preferred over voiced ones. There are also differences, however, which cannot be explained if a maturational view is maintained. This is perhaps best seen when looking at the fricative systems of the above languages. Some of the languages show early use of alveopalatal

fricatives and affricates, as does Quiche and Italian. English, on the other hand, shows little early use of these despite their presence in the language. While several of the languages show the early use of [f] and [s], these are later in Quiche and Dutch, where the velar [x] tends to be the first fricative.

Other differences can also be seen with the liquids. English has an [l] but it is not an early sound in the language. This is not the case with several of the other languages, however, where [l] is a basic sound. This is particularly striking in the Quiche data, where Pye, Ingram & List [11] report that it is one of the two most used sounds. They also provide a further analysis which reveals that it is also one of the most frequent sounds in the vocabulary addressed to young children.

Such differences even appear with the stop consonants where similarities usually abound. Four of the languages show early three-way place distinctions with early velar sounds. The French data, however, suggest that these velar stops may come in relatively later when compared with the fricatives.

The critical question becomes explaining the source of these differences. Since the children are selecting in several instances from very similar sounds, perceptual differences cannot account for the difference. In Pye, Ingram & List [11] it is proposed that the differences result from the children's linguistic organization of the more frequent sounds that they hear. This frequency, however, is type frequency rather than token frequency. That is, it is not important that a sound just be frequent, but that it also occur in a range of words, thereby providing information to the child about the sound's linguistic function. This difference between type and token frequency accounts for the fact that the voiced dental fricative in English is acquired very late. It has token frequency, but is restricted in appearance to a small range of function words.

A further argument for this interpretation comes from the patterning of the sounds used. If the interaction hypothesis were correct, there would be no reason to expect linguistically patterned systems at this point in acquisition. An examination of the consonants in Table 2, however, suggests otherwise. The sounds by and large fall into pattern sets where minimal contrasts can be proposed. Of course, much more detailed linguistic analyses of individual children will be needed to substantiate this claim, but the data in Table 2 are at least suggestive that such an interpretation is on the right track.

5. SUMMARY

Research on infant speech production is at an important stage in resolving the question of when children begin to show the effects of the ambient language. Recent research has suggested that infants may show ambient effects earlier than previously thought, once more sophisticated analyses are conducted on infant speech samples. Even more controversial is the interpretation of this finding. One point of view is to give this little linguistic significance, but to interpret it as indicative of a close connection between the infant's perceptual and articulatory systems. I have presented a more radical interpretation, arguing that it suggests early linguistic processing. Even more subtle and detailed analyses will be needed before this more latter issue will be resolved.

6. REFERENCES

[1] DE BOYSSON-BARDIES, B. HALLE, P., SAGART, L. & DURAND, C. (1989). "A crosslinguistic investigation of vowel formants in babbling", *Journal of Child Language*, 16, 1-17.

[2] DE BOYSSONS-BARDIES, B. & VIHMAN, M. (in press). "Material evidence of infant selection from target language: a cross-linguistic phonetic study". In C. Ferguson, L. Menn & C. Stoel-Gammon (eds.), *Proceedings of the Stanford conference on phonological development*. Parkton, Md.: York Press.

[3] INGRAM, D. (1981). *Procedures for the phonological analysis of children's language*. Baltimore, Md.: University Park Press.

[4] INGRAM, D. (1989). *First language acquisition: method, description, and explanation*. Cambridge: Cambridge University Press.

[5] INGRAM, D. (in press). "Early phonological acquisition: a crosslinguistic perspective". In C. Ferguson, L. Menn & C. Stoel-Gammon (eds.), *Proceedings of the Stanford conference on phonological development*. Parkton, Md.: York Press.

[6] JAKOBSON, R. (1968). *Child language, aphasia, and phonological universals*. The Hague: Mouton. (Translation by R. Keiler of original German version published in 1941).

[7] LENNEBERG, E. (1967). *Biological foundations of language*. New York: Wiley.

[8] LOCKE, J. L. (1983). *Phonological acquisition and change*. New York: Academic Press.

[9] LOCKE, J. L. (1988). "Variation in human biology and child phonology: a reponse to Goad & Ingram", *Journal of Child Language*, 15, 663-8.

[10] OLLER, D. K. (1980). "The emergence of the sounds of speech in infancy". In G. Yeni-Komshian, J. Kavanaugh & C. Ferguson (eds), *Child phonology: vol 1.production*. New York: Academic Press, 93-112.

[11] PYE, C., INGRAM, D. & LIST, H. (1987). "A comparison of initial consonant acquisition in English and Quiche". In K. E. Nelson & A. van Kleeck (eds.), *Children's language, Vol. 6*. Hillsdale, N. J.: Erlbaum.

[12] VIHMAN, M., MACKEN, M., MILLER, R., SIMMONS, H. & MILLER, J. (1985). "From babbling to speech: a reassessment of the continuity issue". *Language* , 61, 395-443.

EARLY SELECTION OF PHONETIC REPERTOIRE: CROSS-LINGUISTIC DIFFERENCES.

Bénédicte de Boysson-Bardies

CNRS - EHESS 54 boulevard Raspail,
75006 Paris, France

ABSTRACT

Cross-linguistic studies of infant productions show that among the wide range of possible sounds that prelinguistic infants can produce a choice is made in relation to the phonetic caracteristics of the inputs of the environment. Two studies demonstrated language-specific production patterns in vowels and consonants in the course of the first year.

We view early specific linguistic effects on infant vocal productions as the result of **selection** by infants from a range of possible sound-productions, and not as evidence of new sound **acquisition**.

1. VOWELS

The first study we present, was based on a statistical analysis of the relative distribution of sounds within the vowel space in the adult target language and in the vocalization patterns of infants, as defined by formant frequencies. Infant's vowels were extracted from 20 minutes of recording of twenty 10-month-old French, English, Algerian and Cantonese infants. The comparative analyses were based on computation of statistical distances between infants' vowel sets. They show that infants differ more between linguistic communities than within any single linguistic community. The F2/F1 ratio, an index of vowel compactness, was used to give information on the relation between infant and adult vowel spaces. The adults' data were taken from the existing literature on each language (frequency of occurrence and F1, F2 mean values). The same trends were found in the compactness index for F2/F1 in the infants' and adults' data: English < French < Algerian < Cantonese.

Table 1. F1/F2 ratios for infants and adult vowels by language communities.

	Infants	Adults
English	3.00	3.68
French	2.80	3.28
Algerian	2.40	3.03
Cantonese	2.24	2.71

In the production of their vowels, infants have begun to approximate favored values in the adult language even before production of the first words.

2. CONSONANTS

In the second study (*) we made an analysis of the consonantal repertoire of five infants in each linguistic group: French, English, Swedish and Japanese. As for the vowel study the method to be used emphasizes the relative distribution of consonants found in infants' productions and its relation to adult language targets. However in this study (1) the analysis of consonantal production is based on a perceptual analysis of infant vocalizations. (2) We use a longitudinal analysis of vocalizations spanning the period from babbling only to early words with concurrent babbling (first 25 words) (3). Phonetic characterizations of infant productions were directly compared with statistical properties of the actual target words attempted by the children in each group. We took these words to offer the most representative adult-language sample available.

The consonants produced by infants were classified according to place and manner categories: labials, dentals and velars for place, and stops, fricatives, nasals and liquids for manner.

2.1 Overall productions show good deal of stability in mean frequency of distribution for place and manner categories throughout the period studied within each group. The infants' groups differ significantly in distribution according to the place of articulation. The interaction effect is mainly due to the difference in the distribution of labials. French infants produce significantly more labials than Swedish and Japanese. French infants' velar productions differ also from Swedish and Japanese.

There is also a main effect of the factor of manner. Swedish infants' productions differ significantly from French for stops and nasals.

2.2. Comparison betwen babbling productions and infants' words. No significant difference is found in any group for place categories, the distribution in infants' words is similar to the distribution in babbling (Fig.1).

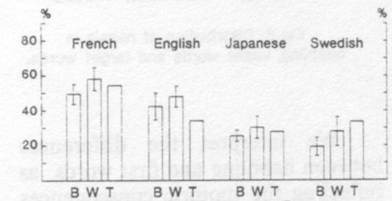


Fig 1: Distribution of labials in babbling, infant words and target words.

The distribution of manner categories for words is parallel to the babbling distribution for French, American and Swedish. There is a significant difference between babbling productions and words for stops in Japanese infant production (see Fig.2).

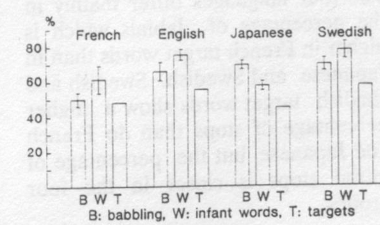


Fig 2: Distribution of stops in babbling, infant words and target words.

In words, there is a general tendency to produce more labials, more stops, fewer fricatives and the inter-group differences found in the percentage of nasals are small (see Fig.3).

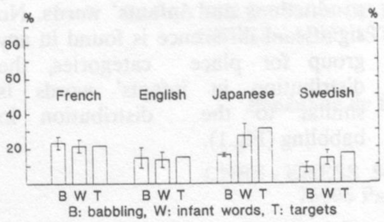


Fig 3: Distribution of nasals in babbling, infant words and target words.

We interpret the differences between babbling and first words as reflecting the motoric consequences arising from the obligatory sequentiality of segments and syllables for word production. A limited ressource assumption (Kent 1991) may account for a tendency to return to more basic adjustments.

2.3. Adult reference sample. The distribution of consonants in the adult reference sample shows significant differences for place and manner categories among the four language groups (see Figs 1- 2-3). The four languages differ mainly in the percentage of labials which is higher in French target words than in Japanese and Swedish. Swedish and English target words show a higher percentage of stops than do French and Japanese, but the percentage of initial stops is closer in the four groups.

2.4. Comparison of infant productions with the adult reference sample (target words).

If there is an effect of linguistic environment, the distribution of labials, based on either the overall or the initial-position distribution in target words should yield the same predicted ranking that is found between infants' groups. And indeed we found the predicted ranking French > English > Japanese = Swedish (see fig 1). For velars the expected ranking is Swedish > Japanese = American > French. We find that French infants produce, as expected, a lower percentage of velars, while the other three groups have a closely similar percentage of velars.

The patterns of manner distribution in infant productions agree with those of manner categories in target words. For stops, the distribution in target words predicts the ranking Swedish = American > Japanese = French. The percentage of stops is indeed higher in Swedish and English data than in Japanese and French (see fig 2). For nasals the expected ranking would be Japanese > French > American = Swedish; this ranking is indeed found in the infants' word production (see fig 3). The four groups do not differ significantly in fricative or liquid distribution and the percentages are too small to be informative for group comparisons.

The differences found in the distribution of consonants for place and manner categories in infant groups are predictable from the distribution of place and manner categories in target-words. Already we can see language specific tendencies in consonants as was demontrated in vowels.

3. REFERENCES

1.BOYSSON-BARDIES, B. de, HALLE, P., SAGART, L., DURAND, C. (1989), "A cross-linguistic investigation of vowel formants in babbling", *Journal of Child Language*, *16*, 1-17.

2.BOYSSON-BARDIES, B. de, VIHMAN, M., (to appear in 1991) "Adaptation to language: Evidence from babbling and early words in four languages", *Language*.

3.BOYSSON-BARDIES, B. de, VIHMAN, M., ROUG-HELLICHIUS, L., DURAND, C., LANDBERG, I., ARAO, F., (to appear in 1991), "Material evidence of infant selection from target language:A cross-linguistic phonetic study", in C. Ferguson, L. Menn, C. Stoel-Gammon (Eds.), *Phonological Development: Models, Research and Application*, Parkton, MA, York Press Inc.

4.KENT, R.D., (to appear in 1991), "The developmental biology of phonological acquisition" in C. A. Ferguson, L.Menn and C. Stoel-Gammon (Eds.), *Phonological Development: Models, research and application*, Parkton: York Press.

*** Note We gratefully acknowledge support from NSF grant BNS 85-20048 to C.A. Ferguson, with the participation of M.M. Vihman, B. de Boysson- Bardies, B. Lindblom, L. Roug O. Engstrand. For their assistance we thank M.A. Macken, R. Miller, and H. Simmons for English, and I. Landberg for Swedish, C. Durand for French, and F. Arao for Japanese.**

Acquisition of the Swedish tonal word accent contrast

Olle Engstrand and Karen Williams
Institute of Linguistics, University of Stockholm, Sweden,
and
Sven Strömqvist
Institute of Linguistics, University of Gothenburgh, Sweden

ABSTRACT

Measurements of F0 contours in disyllabic vocalizations show that, at 17 months, Swedish children are beginning to produce contours typical of the grave word accent and to mark the appropriate words with these contours.

1 INTRODUCTION

Studies of phonological acquisition of tone languages frequently state that language-specific features of tone are acquired early; in particular, tonal features are said to be acquired earlier and to show less variability than segmental features (e.g., Chao, 1951; Li and Thompson, 1977; Tuaycharoen, 1977; Tse, 1978; Clumeck, 1980; Crystal, 1986). Proposed explanations have claimed that pitch, the primary auditory correlate of tone, is easy to control in both production (Li & Thompson, 1977) and perception (Tse, 1978). Independent experimental evidence corroborating child phonologists' conclusions that early acquisition of tone has an auditory-motor basis comes from studies showing that infants are able to imitate pitch before the middle of the first year of life (Kessen et al., 1979; Kuhl and Meltzoff, 1988) and discriminate pitch as early as 4 weeks of age (Kuhl and Miller, 1982).

Existing data which might be relevant to tonal development in young children come from small-scale investigations using very small groups of informants and utterance samples. Thus, there is a need to use direct controlled cross-language comparisons to broaden the empirical foundation for further theoretical studies in the area. The general purpose of our current research program is to use large-scale experimental-phonetic techniques, based on cross-language observations, to investigate the development of the Swedish word accent distinction. The present study represents a first step in this direction, focussing on the so called grave accent, which is the marked member of the word accent distinction.

Functionally, the grave accent marks lexical contiguity by connecting a primary stressed syllable with a later secondary stressed syllable. Most disyllabic and polysyllabic words, particularly compounds, have the grave accent. The characteristic F0 correlates of the grave accent in Stockholm Swedish have been shown to include a two-peaked contour resulting from (1) a falling sequence High-Low, associated with the primary stressed syllable, and (2) a subsequent rising sequence Low-High, associated with the secondary stressed syllable (Bruce, 1977; Engstrand, 1989), the latter rise being optional, however, in that it primarily marks sentence stress (Bruce, 1977). The F0 fall has turned out to be an extremely reliable grave accent criterion. In a recent study (Engstrand, 1989), we examined F0 contours correlating with the word accents in spontaneous speech. The results showed that the grave accent was consistently marked by a falling F0 contour in the primary stressed syllable, frequently followed by a rise in the secondary stressed syllable.

It can be concluded for these data that Swedish children are heavily exposed to grave-like pitch patterns, and that it is, therefore, reasonable to expect an influence of this pattern on early imitative vocalizations. The specific purpose of the present study was to investigate: a) whether such an influence is present at the developmental stage where children have a vocabulary of approximately 50 words, i.e. at about 17 months of age, and produce both meaningful and pre-meaningful utterances; and b) whether there is a beginning functional mastery of the system in the sense that children associate grave-like F0 contours with words that have the grave accent in the adult language.

2 METHODS

The question of the presence of grave-like contours in Swedish children's vocalizations was tested by a cross-language comparison, using a comparable group of children acquiring American English. The question of possible functional use of the grave accent was tested by a within-language comparison of grave word candidates as opposed to non-candidates as produced by the Swedish children.

The data were based on audio and video recordings of five monolingual Swedish children and five monolingual American English children. The recordings of the two groups of children were made using the same protocol in the children's homes in Stockholm and Stanford, respectively, biweekly from the age of nine months until the child had an approximate vocabulary of at least 50 words. The results presented here are drawn from that last stage. The data are based on an examination of all disyllabic vocalizations produced by the respective children during one session, i.e., vocalizations that were unambiguously heard as disyllabic by the experimenter, irrespective of whether or not they had been identified as approximations to real words. The material was digitized at 2C kHz. Extracted F0-contours were displayed in synchrony with spectrograms on the computer screen. Vocalizations lacking measurable F0 contours due to voice irregularities were discarded from further analysis. Spectrographic segmentation was made at acoustic discontinuities marking vowel-consonant and consonant-vowel boundaries. F0 values were sampled at the following points in time: a) the acoustic onset of the first spectrographic vowel segment (V1), b) the F0 turning-point, if any, during V1 (if the F0 contour was monotonous throughout V1, the turning-point was assigned the value of the onset), c) the acoustic offset of V1, d) the acoustic onset of V2, and e) maximum F0 during V2 (if F0 declined throughout the V2 segment, maximum F0 was assigned the value of the onset). The Fall parameter was defined as the F0 difference between V1 turning-point and offset, and the Rise parameter was defined as the F0 difference between V2 maximum and V1 offset. Both parameters can thus take positive or negative values; for example, a negative Fall value means that F0 rises in V1, and for a grave-like Fall-Rise sequence, both parameters take positive values.

3 RESULTS

Table 1 shows the mean value in Hz of all measurable utterances for the Fall and Rise parameters for each of the subjects. It is evident that the individual variation is considerable. There is a considerable overlap between the language groups for all parameters. However, an unpaired one-tailed t-test reveals a statistically significant difference between the language group means for the Rise parameter ($df=8$, $t=2.31$, $p<0.05$). There was no statistically significant difference in the Fall parameter.

Table 2 shows the Fall and Rise data a) for the children's grave word approximations, i.e. vocalizations that were judged to represent these children's attempts to say words that have the grave accent in the adult language, and b) for all remaining vocalizations. A paired one-tailed t-test of the difference between the two sets of vocalizations

Table 1. Individual values (Hz) and grand means by ambient language of the Fall and Rise parameters in the complete set of vocalizations.

SWEDISH					AMERICAN ENGLISH				
		$\bar{x}$	s	n			$\bar{x}$	s	n
Didrik	Fall	2	28	99	Emily	Fall	-12	30	43
	Rise	25	27	94		Rise	13	39	31
Hanna	Fall	17	44	76	Deborah	Fall	-14	43	36
	Rise	18	53	76		Rise	26	43	24
Kurt	Fall	-8	28	31	Sean	Fall	8	63	33
	Rise	75	46	31		Rise	8	53	24
Lina	Fall	43	36	63	Molly	Fall	28	29	132
	Rise	52	70	61		Rise	-42	40	132
Stig	Fall	1	32	83	Timmy	Fall	44	41	19
	Rise	51	55	83		Rise	28	43	19
Grand mean	Fall	11	20	5	Grand mean	Fall	11	25	5
	Rise	44	23	5		Rise	7	28	5

Table 2. Values (Hz) of the Fall and Rise parameters in the Swedish grave word candidates and in the remaining vocalizations.

Child	Parameter	Grave word candidates			Remaining vocalizations		
		$\bar{x}$	s	n	$\bar{x}$	s	n
Didrik	Fall	-1	28	42	5	28	57
	Rise	31	29	39	21	24	55
Hanna	Fall	16	45	19	17	45	57
	Rise	25	73	19	15	45	57
Kurt	Fall	-8	28	23	-8	28	8
	Rise	79	48	23	62	40	8
Lina	Fall	39	42	29	47	30	34
	Rise	64	57	27	45	78	34
Stig	Fall	11	27	9	-1	33	74
	Rise	85	40	9	46	72	74
Grand mean	Fall	11	18	5	12	22	5
	Rise	57	27	5	38	19	5

shows a statistically significant difference in the Rise parameter ($df=4$, $t=3.57$, $p<0.05$). Again, there was no statistically significant difference in the Fall parameter.

4 DISCUSSION

In summary, then, an evaluation of the Rise parameter provides
evidence that the Swedish children are beginning to produce grave-like F0 contours at 17 months and to mark the appropriate words with these contours. However, the absence of an effect for the Fall parameter, both between language groups and within the Swedish group is unexpected in view of Engstrand's (1989) previous finding that an early F0 fall is an extremely stable feature of grave disyllables in adult-directed speech. It is possible, however, that the Rise and Fall parameter characteristics are different in speech directed to children. In adult to adult speech the rise component of grave F0 contours is known to be sensitive to stress, the height of the rise increasing with the salience of the word. It is conceivable that this becomes more exaggerated in speech directed to children. Phonetic studies of speech directed to children are presently being carried out in our lab to shed light on this question.

REFERENCES

BRUCE, G. 1977. "Swedish word accents in sentence perspective." *Travaux de l'institut de linguistique de Lund 12*, Lund: Gleerup.

CHAO, Y.R. 1951. "The Cantian idiolect: an analysis of the Chinese spoken by a twenty-eight-months-old child". In C.A. Ferguson and D.I. Slobin (eds.): *Studies of child language development.* New York: Holt, Rinehart and Winston, 1973.

CLUMECK, H. 1980. "The acquisition of tone". In G.H. Yeni-Komshian, J.F. Kavanagh and C.A. Ferguson (eds.): *Child phonology, Vol. 1.* New York: Academic Press, pp. 257-275.

CRYSTAL, D. 1986. "Prosodic development". In P. Fletcher and M. Garman (eds.): *Language acquisition.* Cambridge: Cambridge University Press.

ENGSTRAND, O. 1989. "F0 correlates of tonal word accents in spontaneous speech: range and systematicity of variation". *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS) 10*, 1-12.

ENGSTRAND, O. 1989. "Phonetic features of the acute and grave word accents: data from spontaneous speech". *Phonetic Experimental Research, Institute of Linguistics, University of Stockholm (PERILUS) 10*, 13-37.

KESSEN, W., LEVINE, J. AND WENDRICH, K.A. 1979. "The imitation of pitch in infants". *Infant Behavior and Development 2*, 93-99.

KUHL, P.K. AND MILLER, J.D. 1982. "Discrimination of auditory target dimensions in the presence or absence of variation in a second dimension by infants". *Perception and Psychophysics 31*, 279-292.

KUHL, P.K. AND MELTZOFF, A.N. 1988. "Speech as an intermodal object of perception". In A. Yonas (ed.): *Perceptual development in infancy*. Hillsdale, NJ: Erlbaum.

LI, C.N. AND THOMPSON, S.A. 1977. "The acquisition of tone in Mandarin-speaking children". *Journal of Child Language 4*, 185-200.

TSE, J. K.-P. 1978. "Tone acquisition in Cantonese: a longitudinal case study". *Journal of Child Language 5*, 191-204.

TUAYCHAROEN, P. 1977. *"The phonetic and phonological development of a Thai baby: from early communicative interaction to speech"*. Doctoral Dissertation, University of London.

THE EMERGENCE OF INTONATION AND STRESS IN HUNGARIAN: A CASE STUDY

Ilona Kassai

Linguistics Institute,
Hungarian Academy of Sciences, Budapest, Hungary

ABSTRACT

The paper discusses the prosodic achievements of the one-unit stage as they appear in the comparison of coexisting communicative and non-communicative utterances.

1. INTRODUCTION

I started research on child prosody by asking the following questions: (1) How does the conventional prosodic system of the adult language emerge out of the physiologically controlled, therefore highly symptomatic vocalizations of the child? I am particularly interested in the emergence and evolution of intonation and stress; (2) How does the child make use of prosodic features in performing different functions and what kind of functions does it perform through these features? For the purposes of a longitudinal study I regularly recorded the spontaneous productions of my dauther from the moment when she was 1 year old up to 6 years.

In the present paper I shall outline the child´s achievements in the one-unit period.

2. THE SYSTEM TO BE ACQUIRED

Hungarian, a "free" word order language, has fixed, first-syllable stress. Sentences may have several, equally strong primary stresses in their main part (comment). The rightmost primary-stressed syllable initiates a character tone (=terminal contour) which can be: falling, falling-rising, rising, descending and rising-falling. The character tones actually appear in phonetic variations conditioned by the number of syllables on which they are spread out. The one-syllable, two-syllable and three-or-more syllable variants (=allotones) are in complementary distribution. If there are any primary-stressed syllables before the terminal pattern, each of them initiates a half-falling tone, i.e. a steep fall not reaching the base line. These primary stressed sequences are subject tro downdrift. If there is only one primary stress in the sentence, it is most often located on the focus position, i.e. on the position immediately preceding the verb or, if the F-position is vacant, utterances 8 items show an overall rising contour. 19 items are expressed by level contour while the remaining 36 items, i.e. the majority, display falling intonation. Tokens of the same type make the consistency in the use of falling contour apparent (Fig. 1).

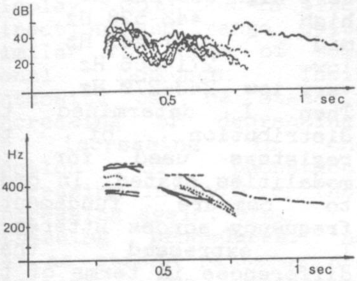


Fig. 1. Intensity and Fo curves of seven occurrences of the utterance *Cica* [t͡sit͡sɔ] *'cat'*

The tendency for fall is also evidenced by the distribution of the registers used in the first and the second syllables of disyllabic utterances (Table 1). The average extent of the fall turned out to be a **third**, the same as in adult language use [1]. The intensity loss of second syllables is 6 dB on the average.

Table 1

Reg.	1st syll.	2nd syll.
high	20,0	11,0
mid	65,0	39,0
low	12,7	47,6

Interrogatives. This modality is not really used yet by the child but the acoustic shape of monosyllabic yes/no questions is already set up: all four instances display a sharp rise of about a large fourth up to the end of the phonation from mid to high register which turns out to be the dominant one.

Imperatives. Imperative utterances, as a rule, are characterized by a strong fall (5-8 semitones) from high to low register and a strong decrease in intensity.

Calls or vocatives. In calls both the intensity curve and the pitch curve are level, the latter being kept in mid register. Therefore, the overall pitch range is as narrow as 50 Hz (371-421).

Sound play. A general characteristic of playful sequences, in contrast to communicative utterances, is their much longer duration along with abrupt and rapid changes both in F_o and intensity values within and across syllables. The magnitude of F_o movement within syllables goes from a third to a large seventh. Sound play is realized in mid, high and very high registers within a range given between 243 and 629 Hz. In monosyllabic items durational values range between 280-1157 ms, the average duration being 600 ms, while that of the corresponding comunicative utterances is 170 ms. If we analyze F_o changes as a function of duration we can state that the longer the utterance, the smaller the F_o changes are. This fact suggests that the child always performs the same underlying pattern displaying a constant difference between starting on the verb itself. However, it may happen that sentence stress falls on some other constituent within the comment. In final analysis, both word order and stress placement seem to be governed by the speaker's communicative needs reflected in topic-comment structure. (For details see [1] [3] [4].)

3. THE ONSET OF ACQUISITION

The results reported concern the period from 1;0 to 1;7.From the recorded material I selected 123 utterances for instrumental analysis. On the basis of their primary function 94 of these utterances were identifified as **communicative**, i.e. aiming at communication with the environment, while 29 items were regarded as non-communicative, **informative** [2] utterances aiming at practising skills in voice production and also at playing with sounds. Therefore, they are taken as late babbling and referred to as **sound play**.

Within the communicative utterances, on pragmatic grounds, I distinguished the following modalities or intention-types: (1) declarative (80 items), (2) interrogative (4 items), (3) imperative (6 items), and (4) call or vocative (4 items).

The instrumental part of the investigation consisted of fundamental frequency, intensity and durational measurements. Data processing was completed by a perceptual test for stress patterns made on 20 adult listeners.

In the data first I established the overall pitch range used by the child in both communicative and non communicative utterances. The respective values are:

Play	243-629 (=386 Hz)
Decl	271-528 (=257 Hz)
Int	357-500 (=143 Hz)
Imp	314-443 (=129 Hz)
Call	371-421 (= 50 Hz)

Within the overall pitch range I defined five registers(=pitches):

very high	529-629 Hz
high	443-528 Hz
mid	357-442 Hz
low	271-356 Hz
very low	243-270 Hz

Then I determined the distribution of the registers used for the modalities stated. In order to compare fundamental frequency across utterances I expressed their differences in terms of the musical scale, i.e. in semitones. For the intensity I only measured peak values. On the basis of the measurements the following general statement may be made: the overall intensity curve and pitch curve go hand-in-hand, i.e. both peaks and valleys coincide at some point of the utterance. It follows then that all that will be stated about pitch contours holds for intensity curves as well. The methodological framework thus established, I analyzed each utterance for the pitch of its syllables, the difference in the registers of subsequent syllables and the pitch movement occurring within syllables. The results underly the overall description that follows.

Declaratives. Monosyllabic utterances are all characterized by level contour. From among the 63 disyllabic declarative and end points of the phonation. Thus, when the pattern is realized in a longer time, F_0 changes become even. The underlying pattern itself is likely to be determined by the physiological capacity of the child for voice production. Sequences built up of repeated syllables, e.g. [pipipipipi], often display variations in pitch direction and range quite similar to those of some tonal language. Their intensity can be steadily increasing or decreasing, or increasing in one section of phonation, decreasing in another, then increasing again, etc.

Stressing procedures. As far as stress patterns are concerned, the perceptual test has yielded the following results. In many cases one-unit utterances display **more than one** stress and this does not agree with the stress rules of adult Hungarian assigning, if at all, a single stress to the first syllable of a word. The options are:

- There is one stress which can fall on any syllable; usually, however, it falls on the first or the last one. This variation can even be observed in different occurrences of the same word, like in ATléta/atLÉta/atléTA ´athlete´(capital letters refer to stressed syllables).
- There are two stresses, one placed on the first and one on the last syllable as in BAbakoCSI ´baby carriage´.
- There are more than two stresses as each syllable of the word has its own stress, e.g. PINGVINEKET ´pinguins (acc)´ OLLÓVAL ´with scissors´.

These procedures are present simultaneously during the period examined. Among stresses assigned by the child to more than one syllable one can discern a "primary stress", i.e. the strongest one. Last syllable stress occurs mostly when the child wants to maintain the contact already established with the partner.

In playful sequences stress falls either on every syllable or on every other syllable that can be the even-numbered as well as the odd-numbered. All it looks as if the child used stressing for the rhythmical structuring of these playful sequences.

4. CONCLUSION

Communicative utterances seem to be from the start under **linguistic control** manifested in each acoustic parameter, especially in the use of pitch patterns while non-communicative, playful utterances are under **physiological control**. All kinds of differences between the two categories are ultimately due to this fundamental difference. In the category of communicative utterances intonation serves to actualize abstract linguistic entities in different speech acts by signalling modalities. On the other hand, the wild variety of stress patterns indicates that stress, at this early stage of the acquisition process, does not reliably perform yet its linguistic functions.

5. PROSPECTS

Later prosodic development

applies the same trial-and-error principle that operates in segmental development. In complex structures the child may be (and is) mistaken both in the number of stresses to assigne and their placement. Intonation errors occur mainly in yes/no intonation questions whose patterning is intimately related to semantic focus. The elimination of errors takes places through learning the complex interplay of prosody, syntax and semantics. Therefore, these errors constitute a major challenge to the linguist for they can tell how far the child's grammatical knowledge actually extends.

5. REFERENCES

[1] Fónagy, I. & Magdics, K. (1967), "A magyar beszéd dallama [The Melody of Speech in Hungarian]", Budapest: Akadémiai Kiadó.
[2] Lyons, J. (1977), "Semantics" I, Cambridge: Cambridge University Press.
[3] Varga, L. (1983), "Hungarian Sentence Prosody: An Outline", *Folia Linguistica 17*, 117--151.
[4] Varga, L.(1984),"The Syntactic Structure and Intonation Segmentation of Hungarian Sentences", *Acta Linguistica Acad. Sci. Hung. 34*, 197--250.

ACQUISITION DE LA PROEMINENCE DANS LE LANGAGE EMERGENT

Gabrielle KONOPCZYNSKI

Laboratoire de Phonétique, Besançon,
France

ABSTRACT

This is a preliminary report on the acquisition of prominence by young children in European languages.From the description of our population (12 French children) it can be seen that final lengthening, typical of French trailor timed rhythm, appeared towards 16;. but the same has been found in English and in Comanche which are stress-timed languages. Other languages, like Portuguese, Hungarian, Italian, German, are discussed. An operationnal prominence acquisition model is put forward.

On sait la difficulté de travailler sur la notion de proéminence, ainsi que sur les structures rythmiques du langage adulte, en raison de l'interférence des divers paramètres prosodiques et même segmentaux. Dans le langage enfantin, étudier ce point, dans l'état actuel de nos connaissances, relève pratiquement d'une mission impossible. Cette communication est donc plutôt un **rapport préliminaire**, destiné à poser les problèmes et à montrer comment cette question a été abordée dans les études sur le langage émergent à travers diverses langues européennes. Un point nous a particulièrement intriguée au cours de nos recherches : il semblerait que l'enfant acquière d'abord un rythme de type "universel" avant d'acquérir la structuration accentuo-temporelle propre à sa langue maternelle. C'est du moins ce qui ressort des rares études dont nous disposons, concernant notamment des langues aussi différentes du point de vue rythmique que le français, l'anglais, le portugais et le hongrois. Rares sont en outre les études instrumentales qui font une analyse pluriparamétrique; c'est pourquoi nous nous limiterons également dans cet article au seul paramètre de durée; les paramètres de hauteur et d'intensité seront réservés pour une recherche ultérieure. Enfin, signalons que la plupart des travaux, anciens, sont de type auditif, et la notion de proéminence n'y est pas -ou mal- définie. En ce qui nous concerne, nous définirons la *proéminence* comme une mise en relief quelconque, sans valeur linguistique, alors que *l'accent* a une fonction linguistique déterminée. Nous ferons dans une première partie état de nos propres travaux, qui consistent en une triple analyse : auditive, linguistique et acoustique des paramètres rythmiques de 12 bébés français entre 9 et 24 mois, puis nous tenterons pour plusieurs autres langues d'établir un bilan à partir de la littérature.

1. LE RYTHME DU LANGAGE EMERGENT FRANCAIS.

Nous avons pu montrer dans des travaux antérieurs [14,16] que le babillage, loin d'être monolithique, contient des "discours" différenciés selon le contexte situationnel. Ainsi, lorsque le bébé est seul et qu'il n'a donc pas à communiquer avec son entourage, il émet des productions floues et instables, difficiles à décrire auditivement et auxquelles des auditeurs extérieurs ne peuvent attribuer aucune signification. Nous les avons appelées Jasis. Au contraire, en situation d'interaction, les productions sonores sont plus stables, souvent récurrentes et des auditeurs leur attribuent, avec une large unanimité, des fonctions ou des modalités linguistiques précises, telles que catégorisation des énoncés en énonciatives, interrogatives, appels, impératives... Nous les avons appelées Proto- ou Pseudo-Langage(abrégé en PL).

L'unité rythmique sur laquelle nous avons choisi de travailler est la syllabe qui est peu sensible aux différences de débit et qui semble être une unité de perception et de production, tout au moins pour le bébé (MEHLER); par ailleurs, à 9/10 mois, l'enfant , du moins l'enfant français, ne sait pas encore produire de proéminences et son discours ne semble pas structuré du point de vue accentuel.

A 9 mois, Jasis et PL s'opposent totalement de par leur structuration rythmique. Du point de vue de l'organisation syllabique, le Jasis comporte majoritairement (71%) des vocoïdes de durée très variable, allant d'éléments très brefs à des éléments extra-longs; l'enfant explore par ce biais ses capacités respiratoires et phonatoires. Au contraire, le PL est constitué majoritairement de structures CV de 2 (28,5%) ou 3 (28,5%) syllaboïdes, les énoncés multisyllabiques plus longs formant 29% des productions. Ces syllaboïdes n'ont rien de commun avec les vocoïdes du Jasis: leur durée est assez stable, proche de celle du langage adulte

	JASIS	PL
Moyenne	1007	251
E.T.	1211	105
Mediane	530	220
min.	60	80
Max	8530	660

durées exprimées en ms.

L'évolution de ces deux catégories de syllaboïdes jusqu'à 24 mois est tout à fait remarquable. Même quand l'enfant parle déjà, vers 2 ans, lorsqu'il émet ce qu'il est convenu d'appeler du babillage tardif, ses vocoïdes - qui diminuent certes en nombre - gardent les mêmes caractérisques qu'antérieurement. En revanche, les durées des syllaboïdes du PL accusent un net changement : d'abord isochrones, les syllabes se diversifient progressivement selon leur position dans l'énoncé. Les syllabes non finales (SNF) se raccourcissent notablement et régulièrement (coefficient de corrélation r entre âge/ durée significatif) ; les syllabes finales (SF) , quant à elles, restent instables assez longtemps (r non significatif) , mais l'abrègement des SNF donne une impression globale d'allongement final, le rapport SF/SNF devenant supérieur à 1,30; vers 16 mois, les SF s'allongent, prenant finalement presque le double de la valeur des SNF. Ainsi, le rythme "trailor timed" [27] avec son point d'orgue en finale d'énoncé, est mis en place vers la milieu de la seconde année. Certes, dans le détail, l'évolution est quelquefois plus complexe. L'apparition de cibles segmentales précises, de mots articulés, la fréquence d'occurrence de ceux-ci, la longueur totale de l'énoncé ... sont autant de facteurs qui interviennent et qui amènent l'enfant à modifier son programme moteur de timing ; nous avons étudié ces facteurs dans le détail [14,16] mais nous n'avons pas le temps de les présenter ici. En conclusion, on peut souligner que le débat au sujet de l'organisation du rythme dans le langage établi - à savoir les schèmes accentuels sont-ils prédéterminés par l'organisation temporelle ou au contraire la structure accentuelle prime-t-elle sur la structuration temporelle - trouve une réponse relativement plus simple dans le langage émergent.

2. ACQUISITION DU RYTHME DANS D'AUTRES LANGUES.

La langue pour laquelle nous disposons du plus d'informations est l'**anglais**; malgré cela tous les renseignements sont à déduire des travaux, le problème qui nous intéresse n'ayant pas été directement traité. Il semblerait, d'après les travaux de SMITH [24] et d'OLLER ([21] 16 sujets) qu'il y aurait au début isochronie, tout comme en français. Un allongement final (A.F.) serait en place vers 30 mois.Mais quand commence-t-il à apparaître? La période charnière, capitale pour découvrir les processus de mise en place des premiers éléments langagiers a été négligée dans les travaux. SMITH trouve, dans deux groupes de dix enfants de 30 mois et dix de 4 ans,prononçant des logatomes, des rapports SF/SNF comparables à ceux des adultes d'un groupe de référence. Il cite également des résultats plus partiels: chez un enfant de 13 mois, existence d'A.F. dans quelques mots significatifs de structure CVCV, où la voyelle finale, pourtant inaccentuée, dépasse de 32% en durée la voyelle non finale accentuée; chez un autre enfant de 18 mois, toujours dans des mots de même structure, 80% des voyelles finales sont plus longues que les autres. En revanche, chez le seul sujet étudié par KEATING & KUBASKA [12]pas encore d'A.F. à 28 mois dans les énoncés redupliqués; mais légèrement plus tard, lorsque cet enfant commence à combiner deux mots, les SF sont plus longues que les SNF. L'acquisition de l'accent en anglais est étudiée entre l'âge de 2-3 ans, parallèlement à l'acquisition du lexique. Les travaux de Mc NEILL, GRUNWELL[9], H.KLEIN [13] montrent que sa place est bien repérée perceptuellement et ceux d'ALLEN [2] précisent qu'elle est bien perçue et reproduite dans des situations expérimentales, surtout lorsqu'il s'agit d'un accent d'emphase, notamment sur les mots nouvellement acquis [6,28]; (signalons qu'en letton, d'après RUKE-DRAVINA, les mots outils nouvellement acquis sont accentués par l'enfant, alors qu'ils sont toujours atones dans le modèle adulte); mais en expression spontanée, les syllabes toniques ne sont pas en place à 2 ans; en revanche, des exemples ponctuels, analysés seulement auditivement, laissent à penser que les prétoniques faibles commencent à disparaître et que les atones se raccourcissent dans les premiers mots stables. D'après BELLUGI, BROWN, FRASER, les mots acccentués semblent mieux retenus que les autres dans un discours (selon PACESOVA [22] ceci est vrai aussi pour le tchèque où les mots ont un accent initial). Mais l'accent de type grammatical, à valeur phonologique contrastive, n'est acquis que tardivement, après 3 ans, et la plupart de temps seulement vers 6 ans, voire 12. On peut expliquer ces phénomènes en posant l'hypothèse que très souvent dans le langage adressé à l'enfant (baby talk) l'accent "normal" est remplacé par un fort accent d'emphase, et que les mots contrastés par la place de l'accent sont relativement rares. Par ailleurs, il semblerait que l'enfant anglais soit obligé de mémoriser le patron accentuel de chaque mot séparément, puis, lorsqu'il connaît le patron de beaucoup de mots, il pourrait progressivement découvrir les règles d'accentuation. Il ressort de l'ensemble des remarques éparpillées dans la littérature que l'enfant repère vers 2 ans, et probablement bien avant, dans le signal acoustique des changements qui correspondent à ce qu'il est convenu d'appeler accent et qu'il est capable de le reproduire en situation expérimentale. Mais il n'utilise pas encore les patrons accentuels comme marques linguistiques. C'est la présence d'une proéminence, par opposition à son absence, qui semblerait acquise, mais pas encore la hiérarchie, ni les règles de structuration accentuelle.Pourtant, le problème est loin d'être éclairci, et bien des points restent à étudier. Notamment l'ensemble des travaux voit l'accent comme un phénomène global, sans chercher à en connaître les paramètres; Seules les études d'ALLEN [1] et de ses collaborateurs [3]tentent de préciser la part relative des divers facteurs; ainsi, lorsque la proéminence frappe la dernière syllabe, les enfants se servent essentiellement d'un indice temporel; mais lorsqu'elle frappe une syllabe non finale,la préférence est donnée à une mélodie montante; cependant il règne une très grande variabilité intra- et interlocuteur.

Nous effectuons actuellement à Besançon, en laison avec la Faculté de Médecine de Lisbonne, dans le cadre d'un programme Erasmus, une étude encore réduite, concernant le **portugais** et portant pour le moment sur 4 bébés entre 9 et 12 mois. Il convient tout d'abord de rappeler que le portugais est une langue à accent lexical, qui porte majoritairement sur la pénultième. L'intensité y joue un rôle non négligeable [7]. Or nos premiers résultats montrent une quasi isochronie à 9 mois, et un début d'A.F., tout juste perceptible, à 12 mois (SF/SNF = 1.30). Etant donné l'importance du parmètre intensité dans cette langue, et ne pouvant l'analyser instrumentalement dans le corpus pour des raisons techniques (distance au micro trop variable), un test auditif a été mené. Les auditeurs, étudiants en phonétique de niveau maîtrise, repéraient une proéminence sur la syllabe finale à 9 mois dans 73% des énoncés, mais à 12 mois, ces proportions s'inversent, et une proéminence est repérée sur la pénultième dans 71% des énoncés. Le rythme typique du portugais se mettrait -il en place dès ce moment là? Ces résultats, surprenants de par la précocité de la mise en place de la structure accentuelle, peuvent s'expliquer partiellement par le fait que les bébés étaient en interaction très directe avec leur mère (bébés sur les genoux, souvent tentant de répéter les modèles maternels). D'autre part, nous les donnons pour le moment avec beaucoup de prudence, vu le nombre trop restreint de sujets et d'énoncés traités.

Des données sont également disponibles pour le **hongrois** où I. KASSAI [11] a travaillé sur sa fille entre 14 et 20 mois, avec une technique proche de la nôtre. Elle parvient à dégager que l'accent est réalisé différemment selon que l'enfant produit des énoncés à fonction communicative (en PL donc) ou des énoncés de type ludique (notre Jasis). Dans les premiers, des accents sont systématiquement présents dans les énoncés de type énonciatif, mais les règles accentuelles du hongrois ne sont pas respectées, car le hongrois présente un seul accent par énoncé, dont la place est strictement règlementée, alors que l'enfant met plusieurs accents, et ce, sur quasiment n'importe quelle syllabe. Par ailleurs, la hiérarchie des accents est encore absente. Enfin , les paramètres accentuels sont beaucoup plus variables que chez l'adulte, le paramètre premier étant un accroissement de hauteur. Dans les énoncés de type ludique, plus il y a de syllabes, plus il y a d'accents, marqués essentiellement par l'intensité; ceux-ci semblent utilisés pour des besoins de structuration rythmique de l'énoncé. KASSAI émet l'hypothèse que dans les énoncés à fonction communicative l'enfant teste à la fois ses capacités phonatoires et la valeur des divers paramètres.

Pour les **autres langues**, les travaux sont quasiment inexistants. Les seuls renseignements que nous ayons pu trouver sont de type anecdotique. Ainsi, pour l'**italien**, RAFFLER-ENGEL [23] signale que son enfant oppose dès 9 mois [ˈpapa](nourriture) à [paˈpa] (papa). VELTEN[26] qui suit sa fille entre 11 et 36 mois note, non pas une réelle proéminence, mais un substitut. L'enfant utiliserait en effet certaines voyelles dans les syllabes accentuées et d'autres dans les atones. Ceci est confirmé par ERVIN-TRIPP , mais à un âge plus avancé.

Pour l'**allemand**, MOSKOWITZ [20] entend une sylllabe accentuée suivie d'une inaccentuée avec mélodie descendante, c'est-à-dire un rythme de type trochaïque. Il en irait de même pour le zuñi [17]. En revanche, les bébés **lusophones** du Brésil (25) ainsi que les comanches [5] auraient, en babil redupliqué, un rythme de type iambique, avec allongement final. Ce fait, prédictible pour le français, est plus étonnant en comanche qui, selon HYMAN [10] est une langue à accentuation initiale.

Notons que ces renseignements sont tous sujets à caution, la notion de rythme ou d'accent étant rarement définie par les auteurs, encore moins mesurée instrumentalement. Par ailleurs, on relève toujours une très grande variabilité intra- et interlocuteurs, qui diminue avec l'âge, selon ALLEN[1, 3].

DISCUSSION

Il apparaît que l'acquisition des phénomènes de proéminence n'est pas chose simple pour le bébé. Il semblerait cependant que le passage par une isochronie initiale, suivie par l'apparition d'un net allongement final serait commun, dans le langage émergent, à des langues de structuration rythmique totalement différente. On pourrait alors, avec ALLEN [1], poser l'hypothèse d'une contrainte rythmique très généralisée due à l'existence d'une horloge neurale, au rythme régulier, contrôlant la production de la parole à sa base, mais dont le fonctionnement serait contrarié par toute une série de facteurs résultant des contraintes phonétiques, mais aussi phonologiques, lexicales, syntaxiques, prosodiques, caractéristiques de chaque langue, et donc acquises. Il existerait donc une structure rythmique sous-jacente régulière dans le tout premier Proto-Langage, et ce, dans des langues à rythmicité totalement opposée telles que l'anglais et le français.C'est du moins ce qui ressort de nos travaux dans leur état actuel, mais d'autres langues sont à étudier de près pour infirmer ou confirmer ces hypothèses. On notera qu' ainsi se trouverait validée, d'un point de vue diachronique, la "structure temporelle simple" posée comme modèle par FRAISSE [8] et reprise par LLORCA [19]. Cette structure est soumise aux seules lois biologiques.Ses limites, ainsi que son organisation interne, corespondent à celles des capacités motrices de l'enfant. La mise en place de l'allongement final, quoique fait très fréquent dans de nombreuses langues, est donc un phénomène acquis, et non inné, ni contraint physiologiquement, contrairement à ce que pensent de nombreux chercheurs (cf. Bilan in [16] p. 270 et suivantes).S'il n'est pas inné, il nous semble en revanche intéressant de considérer l'A.F., à la suite de LINDBLOM [18], comme un processus naturel, qu'on retrouve en musique, en danse,etc.. D'où son acquisition précoce et son utilisation dans de nombreuses langues. Par ailleurs, l'étude constrastive du langage émergent français, anglais, hongrois, portugais nous amène à poser une seconde hypothèse: il semblerait que, dans les langues où la proéminence a une place stable (français, portugais) l'enfant l'acquière relativement rapidement , car il est en présence d'un modèle présentant peu de variabilité; en revanche, dans les langues où la proéminence accentuelle est répartie à des places très variables, l'enfant, qui ne dispose pas d'un modèle stable, a plus de difficultés. Ceci aussi reste à vérifier, mais pourrait être une hypothèse de travail intéressante pour le choix de langues à étudier en priorité.

1 ALLEN, G.D. (1973), " Segmental timing control in speech production" *Journal of Phonetics* 1, 219-237.

2 ALLEN, G.D. (1983), "Linguistic experience modifies lexical stress perception", *Journal of Child Language* 10/3, 535-549.

3 ALLEN, G.D., HAWKINS, S.(1980), "Phonological rhythm : definition and development", in

4 YENI-KOMSHIAN & al (Eds) : Child Phonology, t.I, 227-256.

5 CASAGRANDE, J.B. (1948), " Comanche baby language", *Intern. Journal of American Linguistics* 14, 11-14.

6 CHAFFE, W. (1970), " *Meaning and Structure of Language* ", Chicago University Press

7 DELGADO MARTINS, M.R. (1982), " *Aspects de l'Accent en Portugais*", Hambourg : Buske Verlag, 206p.

8 FRAISSE, P. (1956), " *Les Structures Rythmiques. Etude Psychologique*". Paris : Editions Erasme, 125p.

9 GRUNWELL, P. (1975), " Phonological development and phonological disability in children ", Londres : *Third Intern. Child Language Symposium*, 74-76.

10 HYMAN, L.M. (1975), " On the nature of linguistic stress", Los Angeles : *Studies on Stress and Accent*, 37-82.

11 KASSAI, I. (1981), " The acquisition of stress" , Budapest : *Magyar Fonetikai Füzetek* 3, 80-83.

12 KEATING, P., KUBASKA, K. (1978), "Variation in the duration of words", *J.A.S.A.* 63, Suppl. 1, S56 (Abstract)

13 KLEIN, H. (1984), " Learning to stress : a case study", *Journal of Child Language* 11/2, 375-390.

14 KONOPCZYNSKI, G. (1986), " *Du Prélangage au Langage : Acquisition de la Structuration Prosodique*.Thèse d'Etat, Université de Strasbourg II, 1120 pages.

15 KONOPCZYNSKI, G. (1989), *"Prosodie du Langage Enfantin : une Bibliographie Thématique"*, Hambourg : Buske Verlag, 186p.

16 KONOPCZYNSKI, G. (1990), *"Le Langage Emergent : Caractéristiques Rythmiques"*, Hambourg : Buske Verlag, 362p.

17 KROEBER, A.L. (1916), " The speech of a Zuñi Child ", *American Anthropologist* 18, 520-534.

18 LINDBLOM, B. (1978), " Final lenghthening in speech and music", Lund : GARDING & al (Eds), *Nordic prosody*, 85-101.

19 LLORCA, R. (1988), " *Eléments d'Analyse du Rythme de la Parole en Français*", Besançon : Thèse Régime 1984.

20 MOSKOWITZ, A. (1970), " The two-year-old stage in the acquisition of English phonology'", *Language* 46,426-441.

21 OLLER, K.D. (1980), " The emergence of speech sounds in infancy", YENI-KOMSHIAN (Ed): *Child Phonology* , T.I: *Production*, 93-111.

22 PACESOVA, J. (1959), " *Contribution à l'Etude de la Phonétique du Langage Enfantin* ", Brno : Sbornk Praci Filosofocke Fakulty, A719-30.

23 RAFFLER-ENGEL, W. (1965-73), "The development from sound to phoneme in child language", FERGUSON, SLOBIN (Eds) : *Studies of Child Language Development*, 9-12.

24 SMITH, B.L. (1978), " Temporal aspects of English speech production : a developmental perspective", *Journal of Phonetics*, 6/1,37-67.

25 STOEL-GAMMON, C. (1976) : "Baby talk in Brazilian Portuguese", Stanford University, *Papers and Reports.in Child Language Development*, 11,83-86.

26 VELTEN, H.V.(1943), " The growth of phonemic and lexical patterns in infant language", *Language* 19/4, 281-292.

27 WENK, B., WIOLAND F. (1982), "Is French really syllable timed?" *Journal of Phonetics*, 10,193-216.

28 WIEMAN, L.A. (1976), "Stress patterns of early child language", *Journal of Child Language* 3, 283-285.

VOWEL-CONSONANT RELATIONS IN BABBLING

Peter F. MacNeilage and Barbara L. Davis

University of Texas, Austin, Tx. U.S.A.

ABSTRACT

Early babbling is apparently based on syllabic "Frames" produced by mandibular oscillation. Thus, reduplicated babbling with cooccurring labials and central vowels is "Pure Frames" produced <u>only</u> by mandibular oscillation. Reduplicated babbling with cooccurring tongue front consonants and front vowels is "Fronted Frames". Variegated babbling may initially be produced primarily by "Frame Modulation" - variations in the amplitude of mandibular oscillation (related to stress variation in English).

1. INTRODUCTION

At an earlier Congress of this society, the suspicion was voiced [6] that, counter to the then-prevailing views of Jakobson [2] and Lenneberg, [4] there may be a close relation between the production of babbling and the production of the first words. In the intervening years this suspicion has been abundantly confirmed (e.g. [5], [15]). Both segmental preferences and syllable structure preferences have been found to be virtually identical in first words and concurrent babbling. Consequently, babbling is of prime importance in attempts to understand the acquisition of speech. The question raised in this paper is, what are the organizational principles underlying babbling, and, therefore underlying early speech production? The answer will include the novel conclusion that there are extremely close relationships between early consonantal and vocalic components of babbling. (We will call these two components consonants and vowels for convenience, though this should not be taken to mean that we regard them as independent control units in babbling.)

2. THE CONCEPT OF "FRAMES" FOR BABBLING

Babbling is defined as a relatively rhythmic alternation between an open and a closed vocal tract configuration accompanied by normal phonation. This alternation is produced primarily by oscillation of the mandible. The main thesis of the present paper is that these mandibular oscillations literally provide a "Frame" for the earliest attempts at true speech. It is claimed that much of what happens in the articulatory domain during the entire prespeech babbling period -typically a 6 month period beginning at about 7 months of age - can be understood in terms of cycles of mandibular oscillation <u>alone</u>. If this is true, then it follows that the best way to understand the eventual iacquisition of speech production is in terms of modifications of this basic frame structure. Speech acquisition may be largely a matter of "Frames, then Content" [8].

Multisyllabic babbling can be divided into reduplicated babbling, in which the same syllable-like component (typically consonant followed by vowel) is repeated, and variegated babbling, in which changes occur from syllable to syllable. The present claim is that there are two main kinds of reduplicated babbling. In one, particular labial closing phases - most often phases heard as labial stops - alternate with particular opening phases heard as central vowels. Both closing and opening phases are considered to be entirely produced by mandibular movement. Consequently these are termed "<u>Pure</u> Frames". The labial closures, and tongue positions during opening are considered to be purely passive effects of mandibular oscillation. It is suggested that the vowel configurations are passive because there is no other reason for these <u>particular</u> vowels to preferentially cooccur with these consonants. In the other main type of reduplicated babbling, closing phases involving the tongue front region alternate with front vowels. These are termed "Fronted Frames". Manibular action in these cases is considered to be basically identical to that in pure frames, but superimposed on it <u>throughout</u> <u>an</u> <u>utterance</u> is a fronted tongue position that is adopted <u>before</u> <u>the</u> <u>utterance</u> <u>becomes</u> <u>audible</u> ("Pre-Utterance Fronting"). In this case, cooccurrence constraints between vowels and consonants result from the presence of a single non- resting tongue configuration.

It is also claimed that much <u>variegated</u> babbling is produced by the frame component. It may be produced by variation in the amplitude of the mandibular cycle, with the consequence that most vowel variation within a single utterance is in tongue height, and most consonant variation is in amount of vocal tract constriction, not place of constriction. This process is termed "Frame Modulation". It is also claimed that, in languages in which stress is prominent (e.g. English), attempts to produce differences in stress play an important role in frame modulation because the amplitude of the mandibular oscillation cycle will be positively related to stress level. In particular, more stressed syllables will have more open vowels. In the remainder of this paper, the evidence that led to these claims will be summarized and the validity of the claims will be assessed in a quantitative case study of the babbling of one infant in an English-speaking community.

3. EVIDENCE FOR A FRAME/CONTENT VIEW OF SPEECH

The necessity for considering the ontogeny of speech in an English-speaking community.

3. EVIDENCE FOR A FRAME/CONTENT VIEW OF SPEECH

The necessity for considering the ontogeny of speech in terms of Frame structures with developing Content elements arises when the organization of adult speech is investigated. Segmental serial ordering errors, such as spoonerisms, show that segments (Content elements) are independent units. And syllable position constraints on segmental movement, whereby misplaced segments must retain the position in the syllable which they should have had, if produced correctly, show the existence of syllable Frames at a premotor level of organization [7], [12]. Thus, an important question about speech acquisition becomes, how does this Frame/Content structure develop.

The mandibular oscillation cycle of babbling tells us that the basic temporospatial structure underlying the syllable, the unit which plays a Frame role in adults, is typically present from the age of 7 months onward. The rhythmic continuity of multisyllabic babbling tells us that this embryonic syllable places tight constraints on the structure of speechlike vocalization from the moment of its inception. What are these constraints? Studies of relative frequencies of consonant-like sounds (summarized in [5]) and vowel-like sounds (summarized in [9]) are in good agreement about the features of the closed and open phases of the mandibular cycle of babbling in a handful of languages. Sounds accompanying the mandibular closing phase fall into two main categories; 1. Labials, particularly labial stops, and, to a lesser degree nasals and glides, and 2. Tongue front sounds, particularly stops, and, to a lesser degree, nasals and glides. Tongue back closures tend to develop later. Favored sounds of the opening phase are in the mid and low, front and central regions of the vowel space. High and back vowels are rare.

These patterns, together with the evidence now to be reviewed on consonant-vowel cooccurrence constraints led us to the claims about the organization of babbling made earlier. In a study of the babbling and speech of a child during the period from 14 to 20 months of age,[1] evidence was found for three sets of consonant-vowel cooccurrence constraints; labial consonants with central vowels, tongue front consonants with high front vowels, and tongue back consonants with high back vowels. Vihman [14] for the most part confirmed these trends in an analysis of groups of children around one year of age from four language communities (English, French, Japanese and Swedish) the same groups discussed in another paper in this symposium by de Boysson Bardies. There is also evidence that these patterns may be universal in languages (e.g. [4]). These are the findings

which led us to suggest that there may be cooccurrence constraints between labials and central vowels (Pure Frames), and between tongue front consonants and front vowels (Fronted Frames) in the earliest babbling. One other finding of Vihman [14] added support for one of our claims. She found a cooccurrence constraint between glottal [h] and central vowels. The choice of central vowels to accompany this nonarticulatory segment, as well as the choice of these vowels to accompany labial (nonlingual) segments suggests that these vowels may not be produced with iactive tongue movement.

An additional finding from our case study seemed to have implications for the organization of babbling. A strong tendency was found for more stressed syllables to be accompanied by more open vowels, whether or not these vowels were the ones in the English target word being attempted. Subsequent analysis of English words [1] and words of New Zealand Maori (unpublished obser- vations) showed a strong tendency for more stressed syllables to have more open vowels. This evidence led to our claim that a considerable amount of the variation in variegated babbling may result from variation in the amplitude of the mandibular cycle (Frame Modulation).

4. A CASE STUDY OF BABBLING

To evaluate our claims about the organization of babbling, we conducted a case study (in preparation) in which we recorded the babbling of a child in an English language enironment on 12 occasions evenly spaced from the age of 7 to 12 months. A total of over 400 utterances were phonetically transcribed. Closing and opening phase preferences were, with one exception, quite typical, featuring labial and tongue front stops, with some homorganic nasals and glides in closing phases and mid and low front vowels and mid but not low central vowels in opening phases. In both monosyllables (mostly CV) and disyllabic and multisyllabic reduplicated babbling, there was statistically significant confirmation of the prediction of cooccurrence constraints between labials and central vowels and between tongue front consonants and front vowels. About 75% of vowels in labial environments were central vowels, while about 60% of the vowels in the environment of tongue front consonants were front vowels.

Until recently it had been thought that a phase of reduplicated babbling preceded a phase of variegated babbling in the developmental sequence (e.g. [11]) implying a natural progression in complexity of control. However more recent studies [10], [13] including our own, show instead that variegated babbling often coexists with reduplicated babbling from the very beginning. In the present case study, most instances of variegated babbling only involved variegation in vowels. There was extremely strong confirmation of the prediction that Frame

Modulation would be an important source of variegation, in the case of vowels. In approximately 90% of instances of vowel variegation involving change in stress (assessed perceptually) the more stressed vowel was more open. The pattern in the observed instances of consonant variegation was also consistent with the Frame Modulation prediction, though there were only 11 instances of such variegation. Six instances involved the predicted change in manner of articulation, while only two involved change in place of articulation. The remaining 3 involved use of the glottal [h] as a variant on articulated closures.

In an unexpected finding, much variegation resulted from alternations between front and central vowels. In alveolar environments, front vowels were even more favored in stressed vowels than they were in reduplicated babbling, but not favored at all in unstressed syllables. Conversely in labial environments central vowels were not favored at all in stressed syllables, but more favored in unstressed syllables than in reduplicated babbling. These vowel classes may have had these different roles because front vowels are actively produced, while central vowels are passive resultants of mandibular oscillation.

5. CONCLUSIONS

By use of the concepts of Pure Frames, Fronted Frames and Frame Modulation, a relatively complete account can be provided of the main properties of both reduplicated and variegated babbling in the infant studied here. These concepts are offerred as an attempt to delimit a possible set of core mechanisms of babbling, for evaluation in other infants and languages. It is important to note that this analysis does not include an a priori commitment to any linguistic unit, other than the syllable, as an independent control unit in babbling. The concepts of Segment, Feature, or Gesture are not required. Even the independence of the syllable is limited at this stage. The two most freqently used CV sequences in both reduplicated and variegated babbling were [pʌ] and [tæ], but there was not a single instance of either the [pʌ tæ] sequence or the [tæpʌ] sequence in variegated babbling. The main functional unit of babbling seems to be a cycle of mandibular oscillation, but with severe limitations on its detailed articulatory accompanyments, both when occurring singly and when reiterated.

REFERENCES

[1] DAVIS, B.L. & MACNEILAGE, P.F. (1990), Acquisition of correct vowel production, Journal of Speech & Hearing Research, 33, 16- 27.

[2] JAKOBSON, R. (1968) Child language, aphasia and phonological universals, The Hague: Mouton & Co.

[3] JANSON, T. (1986) Cross-linguistic trends in the frequency of CV sequences, Phonology Yearbook, 3, 179-195.

[4] LENNEBERG, E.H. (1967), Biological foundations of language, New York: Wiley.

[5] LOCKE, J. (1983) Phonological acquisition and change, New York: Academic Press.

[6] MACNEILAGE, P.F. (1980), Speech production, Language & Speech, 23, 3-23.

[7] MACNEILAGE, P.F, STUDDERT-KENNEDY, M.G. & LINDBLOM, B. (1985), Planning and production of speech: an overview, In J. Lauter (Ed.) Proceedings of the conference on planning and production of speech by normally hearing and deaf people, ìAmerican Speech and Hearing Association Reports.

[8] MACNEILAGE, P.F. & Davis, B.L. (1990a), Acquisition of speech production: Frames, then content. In Jeannerod, M. (Ed.) Attention & performance X111: Motor representation and control, Hillsdale, N.J.: Lawrence Erlbaum.

[9] MACNEILAGE, P.F. & DAVIS, B.L. Acquisition of speech production: The achievement of segmental independence, In Hardcastle, W.J. & Marchal, A. (Eds) Speech production and speech modelling, Kluwer: Dordrecht.

[10] MITCHELL, P. & KENT, R.D. (1991), Phonetic variation in multisyllabic babbling (submitted for publication).

[11] OLLER, D.K. (1980), The emergence of the sounds of speech in infancy. In G.H. Yeni-Komshian, J.F. Kavanagh & C.A. Ferguson (Eds.) Child Phonology, Vol. 1, New York: Academic Press.

[12] SHATTUCK-HUFNAGEL, S. (1979), Speech errors as evidence for a serial ordering mechanism in speech production. In W.E. Cooper, & E.C. Walker (Eds) Sentence processing: Psycholinguistic studies presented to Merrill Garrett, Hillsdale, N.J.: Lawrence Erlbaum.

[13] SMITH, B.L., BROWN-SWEENEY, S. & STOEL-GAMMON, C. (1989), A quantitative analysis of reduplicated and variegated babbling, A First Language, 9, 147-153.

[14] VIHMAN, M. (1991), Motor control and the evolution of phonological patterns, An analysis of syllables at the onset of speech, Proceedings of the conference on phonological development, Stanford University, September, 1989.

[15] Vihman, M., Macken, M., Miller, R., Simmons, H. & Miller, J. (1985), From babbling to speech: a reassessment of the continuity issue, Language, 60, 397-445.

INVESTIGATING LINGUISTIC UNIVERSALS

Ian Maddieson

Department of Linguistics, UCLA, Los Angeles, U.S.A.

ABSTRACT

A property that is widely distributed across the world's languages is often taken to be a necessary or desirable feature of language design. This paper reviews ways that biological limits or physical laws may constrain language, and proposes that avoidance of extremes is a desirable trait. The problem of distinguish-ing between 'universals' (shared features deriving from response to such pressures) and inherited similarities is discussed in the light of growing evidence of wider language relationships and new models of human prehistory.

1. INTRODUCTION

For centuries linguists have tried to understand what is essential to the nature of Language, as opposed to something that is particular to a given language or group of languages, by studying language universals. The reasoning is essentially as follows: if a feature is distributed widely enough in the world's languages to be labeled universal then it is either a necessary property of human language, or it is in some sense a desirable one. A similar arg-ument applies when clusters of co-occurr-ing linguistic properties are described under the heading of linguistic typology. If certain sets of properties repeatedly occur together in languages, then it can be argued that their co-occurrence is a necessary or a desirable property. That is, typology includes the study of contingent universals.

It is usually because of this line of thinking that people are interested in universals of language. The study of prevalent patterns in languages, of universals, is a window to examine the question of why language is the way it is. By this hypothesis, universals arise because of biological limits and environmental pressures that are at work on all languages simply by virtue of the fact that they are in use by members of the same species of mammal. Despite the great geographic dispersion of our species and a good deal of individual and group variability between its members, all humans make use of basically the same equipment of brain, vocal tract and auditory system. Studying universals is therefore not so much a goal in its own right as a challenge to the linguist to come up with explanatory accounts of what these pressures are and how they affect human language in general. The goal is to produce models of these pressures that predict the universals that have been observed.

However, if we are interested in universals we have two great problems to face. One is the problem of obtaining knowledge about them. How do we decide what is universal? That is, how do we go about finding what is prevalent enough in the languages of the world to count as a possible universal? The second problem is how to distinguish those properties that we wish to consider 'universal' (in the particular sense that they arise from design considerations that apply to human language in general) from prevalent patterns that arise from other sources of uniformity? It is a hypothesis that important properties of human languages are common because they are based on inherent characteristics of the human species and of the environment in which we as a species employ our linguistic abilities. This hyp-othesis must be compared with alternative hypotheses that might explain the data in better or equally satisfactory ways.

Since the concern of this conference

is with the phonetic sciences, the discussion of these issues which follows will be directed to and illustrated with examples from the domains of phonological and phonetic universals, based in part on my own work with the UCLA Phonological Segment Inventory Database or UPSID [25, 28], but there is nothing in the general principles concerned that would be any different if the field of enquiry was in some other area of linguistics.

2. HOW TO FIND MEANINGFUL PREVALENT PATTERNS?

As has been pointed out before, but is worth stressing again, some kind of structured systematic sampling of the universe of known languages is essential if we want to know what linguistic patterns are prevalent [2, 12, 18, 25]. Prevalence is an essentially statistical concept. We need to be able to say with some confidence that the set of languages within which some property is said to be prevalent (or more common than some other pattern) represents the larger universe that we are really interested in studying, ultimately that of all possible human languages. Above all, if we are looking at patterns of co-occurrence of properties, at typological patterns, we must be able to evaluate the independent distribution of these properties, in order to be able to say if they are significantly associated with each other.

An obvious way to know how widely distributed a particular feature is would be to count the frequency of that feature in *all* human languages. Even if we limit ourselves to languages still spoken at this time, there are two straightforward practical problems which prevent us from attempting this. First, linguists have not yet got around to examining all of the world's living languages, and, second, even if they had, surveying descriptions of all languages would be impossibly time-consuming.

There are also theoretical objections to making this the goal; these concern the need to survey the data in a way that gives appropriate weight to each language. First, there is no unambiguous principle to define the borderline between the degree of difference between two speech varieties that warrants assigning them to different languages and the degree of difference which can be accomodated within the construct of a single language. Different linguists will give various classifications of the same speech varieties. Without an answer to this problem we might include only one dialect of one language but many varieties of another, giving it undue weight in the survey. This makes it impossible to be certain that we have assigned equal weight to each language.

Secondly, we know that where we find close-knit families of languages existing today this reflects an evolution from an earlier stage at which the precursors of these now-distinct languages were dialects of a single language. Separately counting all members of such close-knit groups transparently gives undue influence to the group, just as separately counting all dialects of a language does for that language. This is because members of the group will have so many features in common that are simply inherited. Few of their shared features are likely to be due to *independent* response to the pressures shaping human language that it is our ultimate objective to investigate. We may take the North Germanic languages as an example. There are perhaps five living languages in this group, Icelandic, Faroese, Swedish, Norwegian and Danish, and we know that they go back to a common Norse parent language that was quite uniform as recently as five or six hundred years ago. The descendent languages, unlike their next nearest relatives in the West Germanic group, share some elements of a pattern relating consonant quantity to vowel quality and quantity features. We might overestimate the global prevalence of such a pattern, that is, the number of independent occurr-ences of the pattern, by counting each of these languages separately. In contrast, a language such as Albanian in the same period of time has not fragmented into a number of daughter languages. We would under-represent features that might have been shared by the daughter languages it never had. The problem is just the same as if we were to count each of a large number of modern dialects of English but only to count one variety of modern French. In that case, our survey might show an inflated number of interdental fricatives, and a correspondingly depress-ed number of front rounded vowels.

It is therefore necessary for both

practical and theoretical reasons to develop some strategy so as to make a selection of languages such that each contributes an appropriately equal weight to the sample. One needs to create a sample that can be trusted to represent in a fair way the overall frequency of the properties of interest in the world's languages. For the UPSID project the decision made was to aim for a sample that includes one and only one language from each group of languages that is separated from its closest relative by a genetic distance similar to that separating the North Gemanic languages from the West Germanic languages. In terms of time depth this might translate into about 1500 years of separation, a long enough period for substantial independent developments to occur in the phonological patterns of any two lang-uages belonging to the same larger family. Related languages will, of course, have certain elements of their phonological patterns in common, or we would hardly be able to recognize their relatedness. But at the same time they will have a degree of independence. Languages with *no* closer relatives are also included, as they too represent the outcome of certain lines of independent development. The current UPSID sample size is 451 languages, probably between 5% and 8% of the world's existing languages.

However, despite the restriction built in to constructing the UPSID sample, problems concerning whether the selected languages can be considered truly independent samples do not go away. I will return to this question when it comes to discussing the interpretation of prevalent patterns. But first I will provide a simple illustration of the use of this database to derive estimates of the frequency of phonological patterns.

It is generally agreed that there are more languages with a voicing contrast in stops than languages with a voicing contrast in fricatives [16, 25]. But let us suppose that we want to investigate the claim that voicing contrasts in fricatives *preferentially* occur in languages which have a voicing contrast in stops, that is, there is an implicational universal involved. To do this, it is not good enough simply to point to a large number of languages that do indeed share both types of voicing contrasts and then list a number of languages that have a voicing contrast in stops but not in fricatives. It has to be shown that there are *fewer than expected* cases of languages that have a fricative voicing contrasts but lack stop voicing contrasts. The frequency of fricative voicing and stop voicing independently can be estimated from our language sample, and each number can be viewed as the probability that a given language will have the property in question. We can then multiply these independent probabilities together to obtain estimates of how frequently fricative and stop voicing might be expected to co-occur if there was no contingent relationship between them. The expected value can then be compared with observed frequencies of co-occurrence and singular occurrence, and the significance of the association of the voicing contrasts with each other can be statistically evaluated.

In our UPSID database, about 72% of the languages included have voicing contrasts in stops (i.e. have voiced and voiceless plosives) and about 47% have voicing contrasts in fricatives. That is, the probability of one of the individual languages in the database having a stop voicing contrast is .72, and the probability of an individual language having a fricative voicing contrast is .47. If we multiply these two probabilities together, the result is about .34. That is, if there is no connection between the occurrence of these two things, we may expect 34% of these languages to show both stop and fricative voicing, leaving about 13% that have fricative voicing without stop voicing. The observed figures are in fact 38% and 9% respectively. A simple χ-square test can then be applied to compare the observed with the expected distributions, yielding the answer that there is about a one in five chance that these results are accidental. Since there is only one degree of freedom in this problem the level of significance should perhaps not be taken too seriously, but for what it is worth, the result suggests that the connection between the occurrence of fricative and stop voicing is not all that strong.

As a final note in this section, a word should be said about the care required in drawing conclusions from any assemblage of data about a set of languages. Typically, when a large number of descriptions of languages are brought together to get a view of the variety of language, a wide range of explicit or implicit linguistic theories are represented. Scholars of different language families and from different parts of the world are trained in different traditions, so that different facts are observed and the same facts will be reported in different ways. Also as theor-etical models in a given tradition evolve, what are considered to be the significant properties of a language change. We need to be sure that descriptions are comparable before generalizations are drawn, and to be sure that the inferences made are responsive to the particular nature of the data that is represented.

3. HOW TO INTERPRET PREVALENT PATTERNS?

As remarked above, universals in themselves are not objects of ultimate interest. It is the theory that will account for universals that is the focus. Compiling a sample such as UPSID provides a basis for stating which types of patterns might be justly interpreted as prevalent. For example, since over 98% of languages in the UPSID sample have stops at bilabial, anterior coronal (dental or alveolar) and velar places of articulation, we can say that it is a valid generalization about languages that they are overwhelmingly likely to have stops at these three places. What this means is that we would expect to find this to be true of some different but representative sample of extant languages or if we could travel 2000 years forward or backward in time and sample the lang-uages spoken at that time. In this case, as in any other, once it has been established that some pattern is prevalent or that there is a certain set of properties that tend to co-occur in the world's languages, we are challenged to look for the explanation that might be responsible for that pattern.

There are several types of explanations that may be entertained. They fall into two basic groups. The first type posits that prevalent patterns reflect necessary or desirable properties of language. The second group takes more account of the extent to which prevalent patterns might be due to inherited similarities between languages or to the spread of traits due to contact. These two types of explanations reflect on the one hand the fact that the faculty of language is a basic part of our human make-up and on the other hand the fact that the particular languages that survive and spread result from accidents of history shaped by many socio-political and environmental factors.

The first type of explanation includes the possibility that certain universals are inevitable. Some universals may be due to species-specific biological constraints; these at least set limits to the range of variation that languages may exhibit. However, the absolute biolog-ical constraints that can be stated at this time do not seem to be very interesting. This is perhaps because we know relatively little about what our language-related biological limitations actually are, and hence are restricted mostly to stating the obvious, such as that in their speech mode languages must use articulations that are possible human gestures that leave some acoustic signature of their presence. Thus, although languages make use of various gestures involving the lower lip, such as bilabial, labiodental and linguo-labial articulations, labio-uvulars are universally absent. This is so for the rather uninteresting reason that the lower lip and the uvula cannot meet. This tells us why labial-uvulars are absent but does not tell us why bilabials are universal and linguo-labials very rare.

Aside from articulatory impossibilities, we can also point to certain inevitabilities in speech production of the sort that have been the focus of research by John Ohala and some of his associates (e.g. Ohala [33]). These are effects that arise from the operation of physical laws applicable to the functioning of the vocal apparatus. They are not species-specific in any sense, but since the physical laws apply to all individuals, these effects are also inevitable. Ohala points out how physical laws produce assymetrical results. For example, given the higher resistance to air-flow in high vowels, there is a certain level of subglottal driving pressure at which voicing of high vowels will fail to occur but voicing of low vowels will be sustained. The consequence is that voice-less high vowels are a little more likely to occur than voiceless low vowels. This addresses the observation that there are languages in which all vowels devoice and languages in which only high vowels devoice, but no language is known that has devoicing of low vowels only.

The possibility that there are innate categorical classifications of certain sound types due to the way the perceptual system works remains uncertain, but further biologically-determined limits could arise from such a cause.

Other universals may reflect desirable design attributes of languages rather than inevitable properties. Let us think about one class of desirable properties. For a variety of reasons, humans do not wish to operate *near* the limits of their capabilities. In any mode of activity, errors increase when performance is pushed towards the limits. The nearer an approach is made to an operating limit the greater the difficulty of learning becomes, the more variable individual levels of success become, the greater the degradation of performance under stressful conditions, the greater the difficulties resulting from effects of age, tiredness, etc, and so on. For spoken language, the relevant limits would include limits on the range and speed of movement of the constituent parts of the vocal tract mechanism, limits on the ability of the auditory system to resolve distinctions between sounds, and limits related to the capacity for storage of linguistic knowledge in the brain.

Without knowing exactly where any of these limits lie, we can understand what represents movement *towards* these limits. It seems safe to assert that it is a desirable property of language that it should avoid any approach to the performance limits. This is at once a more inclusive and more cautious formulation of old observations that are usually phrased in terms of lang-uages maximizing ease of articulation and auditory distinctiveness. These two principles have been appealed to in selective ways to account for particular synchronic or diachronic patterns in languages, but the implications of proposing these principles as ones that affect languages across the board have rarely been taken seriously.

An exception is the ambitious phonetic model of phonological origins being developed by Björn Lindblom. The aims of this theory, the Theory of Adaptive Dispersion (summarized in [24]), include being able to account for the ontogeny of segments and the structure of segment inventories. Lindblom's presentations of his theory include a model of the way we might envisage a language developing phonological patterns through selecting an optimal set of syllables. The optimal set is the one that minimizes the value of aggregate articulatory effort, expressed as the sum of deviations from a neutral vocal tract position plus the magnitude of articulatory trajectories in transitions between onset and offset of syllables, and, at the same time maximizes the value of overall auditory contrastiveness, expressed as the sum of differences over time in the auditory spectrum across the set of syllables. This model has at present been developed more as a demonstration that it is possible to predict the optimal set of syllables from any set of input candidates using the very general principles described, and it is not intended that the particular set of selected syllables has any special standing. So it is not appropriate to analyze the set of selected syllables to see if they reflect the preference patterns seen in actual languages. However, we can see in principle how such a model might explain the relative frequencies of bilabials and linguo-labials. Linguo-labial contact requires a tongue gesture of much greater magnitude than the lip-rasing gesture used for a bilabial.

But it is possible that this part of Lindblom's theory is too deterministic. The articulatory and auditory components produce a single optimal solution for a given input.[1] Our impression is that languages are more variable than this. Considering just segment inventories, the evidence from surveys such as UPSID provides no clear evidence that languages are tending towards unique solutions. Consonantal and vocalic systems show certain similarities in their common core but the ways that they are elaborated beyond this common core are quite variable, and reduced systems with less than the common core are not unusual. A cross-linguistic study of syllable patterns currently under way at UCLA [29] shows that most of the languages studied do not have the strong dependencies between adjacent consonants and vowels that might be expected if ease of articulation and auditory distinctiveness, evaluated at the syllable level, play dominant roles in selecting preferred sound patterns. And, after all, languages with linguo-labials do occur [26].

Rather than leaving this variability to be accounted for by social factors, as Lindblom provides, two directions for developing the more strictly phonetic elements of the model seem to merit consideration. One is to add further parameters, reflecting costs and benefits of other aspects of sound patterning, such as rules of word formation and phonological alternations. The 'cost' of the higher degree of articulatory difficulty of, say, consonant clusters may be mitigated when these result from morphological processes such as affixation (as English *move, moved*). Typically, affixes form a closed set and articulatory precision can be relaxed. Similarly, the cost of reduced auditory distinctiveness associated with an increased number of vowel contrasts might be mitigated by the presence of a rule of vowel harmony that limits the free distribution of these vowels at the word level. Recognition of whatever phonetic parameter forms the basis of the vowel harmony distinction is only required once per word, rather than for each syllable. The possibility of an association between larger vowel inventories and vowel harmony is suggested by the fact that for languages in Africa the modal size of the vowel inventory is 7, whereas on the other continents it is 5. Vowel harmony systems are more prevalent in the language families of Africa than in most other areas [27].

The second direction for taking the development of such a model is to relax the constraint that it seeks a single, optimal, solution, so that it produces a variety of possible solutions that cluster around the optimum. That is, accepting our restatement of the desirable design requirements and modelling avoidance of extremes rather than maximization of ease of articulation and auditory distinct-iveness. Of course, here the problem would be to determine how close an approach to the limits should be modeled as acceptable.

If we cannot be satisfied that universals arise from inevitable causes or result from shared pressures towards desirability, our other alternative is to consider that they may result from inherited similarities (or at least transmitted similarities). That is, we may see prevalent patterns that are not the result of innate limits or pressures to select desirable traits independently applying to many separate languages, but are the result of preservation of traits, possibly quite accidental ones, of a parent language which is ancestral to many or even all of the surviving languages (or a lang-uage which influenced surviving lan-g-uages at an early stage). For this reason, it is important for universalists to be very concerned with the issue of how closely related the surviving languages are. Otherwise our conclusions may be little better than the ones we would draw from a sample consisting only of modern English dialects. At one time it seemed that our understanding of the story of human evolution might have allowed for the likelihood that language evolved in parallel in several different areas and over a long period of time. Early diffusion of hominids through the Old World seemed to be followed by a long period of somewhat separate but parallel development (see, e.g. [7]). Present-day populations in East Asia, Africa and Europe were believed to reveal traces of characteristics seen in ancient fossils found in those areas. This would have allowed for the interpretation that when two language families were said to be unrelated, it meant more than that the relationship could not at present be demonstrated by traditional historical-comparative methods. They could actually be of independent origin. The picture now seems more confused.

First of all, there seems to be increasing evidence that many of the groupings of languages that linguists were once content to say were "not related" can be shown to have genetic relationships demonstrable by traditional methods [19]. Reorganizations of the familiar major language families are disruptive to the scholarly communities involved and tend to be met with resistance, or ignored. But even conservative scholars are beginning to concede that the data being assembled in favor of relating (at least) Indo-European, Dravidian, Ural-Altaic, Afro-Asiatic and Kartvelian together has merit [32]. Sagart [35] has recently provided strong evidence that Chinese may be more closely related to Austronesian than to Tibeto-Burman. Since Benedict [3] has shown Austronesian and Thai-Kadai languages to be related, and Sino-Tibetan comparisons still seem valid, a macrogrouping of languages in Asia seems to be emerging. Benedict has further claimed Japanese as a relative of Austronesian [4], whereas Miller [30,31] has shown strong reasons for linking it with Ural-Altaic. If both connections are valid, then a huge number of the languages of the Old World are genetically linked. Missing so far from this agglomeration are the three other major language families of Africa. While there is no shortage of fanciful speculation on their wider relationships [11,17,37], at least the data assembled by Gregerson [14] and Boyd [5] seems to indicate the serious possibility that the Niger-Kordofanian and Nilo-Saharan families are related. As for the New World, many Americanists reject Greenberg's [13] grouping of most American languages into a single Amerind family [8,21] as being, at best, premature. However, cautious scholars continue to show how parts of the picture relate together (e.g. Payne [34]) and eventual demonstration that many of these languages are related seems probable. The late twentieth century is thus a period during which we are recognizing more and more of the world's languages as related to each other, and pushing back the time depth at which relationships can be recognized.

Secondly, our picture of human origins is shifting as modeling of the past based on studies of genetic markers in present-day populations is added to the tools of paleontology. A plausible account has been offered that the surviving human population may trace back to a single African origin of a considerably more recent period than earlier models suggested [6,9,38]; but see Spuhler [36] for a more cautious view). This would suggest that all languages also have a single origin of the same time depth (not more than 200,000 years, or about 4 to 6 times as long as humans have colonized areas such as Australia and the Americas, and perhaps as little as 100,000 years). The recognition of language relatedness among larger groupings tends to support this possibility of a single parent language at a not impossibly remote time period.[2] Since this language doubt-less had its share of arbitrary and idio-syncratic features, we must be concerned that at least some of the properties that we see as prevalent in the world's languages trace back to the idiosyncratic features of this postulated parent language. Such features would be misleading testimony concerning which properties are necessary or desirable in human languages.

Of course, we know that languages change their phonetic and phonological structures over time and much diversity would have evolved from any ancient parent language. Historical studies show that, for example, vowel systems tend to be quite changeable. But there are certain other properties that tend to remain quite stable [15]. Nasals tend to remain nasals in syllable-initial position, for example. Another diachronic pattern is that stops tend to remain stops, at least in pre-stress syllable-onset position, and to retain their place of articulation, especially in low vowel environments. As noted above, stop systems including three major places (bilabial, anterior coronal, and velar) are nearly universal in languages. This seems a candidate for a trait that might be a conservative, inherited, feature. All reconstructed languages at the greatest time-depth that linguists go back to have stops at these places and the great majority of the daughter languages have retained them. There seems to be no necessity for languages to have a stop system with these particular places in contrast. Some languages, such as Ahtna [20], get along quite well with no bilabials, and it is easy to imagine languages that would have no contrast between front and back tongue articulations, with a rule-governed distribution like that of the second articulation in the so-called labial-velar stops of Nzema and Dagbani—alveolar with front vowels and velar with back vowels.

If the minimal three-place structure of stop systems is not necessary, can we show that it is universal because it is desirable? The answer, at least at present, is that we probably can't. This is because our most effective tools for attempting to understand the issue of desirability depend on having variability to analyze and on being able to look at the co-occurrences of particular properties. The many fruitful investigations of the structure of vowel systems in the last several decades—see [22, 10, 23, 1, 40]—illustrate this point. All of these studies analyze the covariation between the size and the content of vowel inventories, and draw their conclusions in the main from comparing changes in the modal structure of vowel inventories as the number of contrasting vowels varies. Without such variability, our ability to create models is impaired, and there is a shortage of data points with which to test the success of the predictions of any model.

The perhaps paradoxical conclusion is that we study the effect and nature of the ambient pressures on language with the most confidence when studying those aspects in which languages display the greatest variability, rather than in studying aspects in which they show the most conformity. Where universal or near-universal conformity is found, and we cannot explain it as due to biological factors or physical laws, it is difficult to reject the hypothesis that the trait in question is inherited.

FOOTNOTES

1. Lindblom's model provides for cross-language variability in two ways; the number of distinct syllables can vary and the output of the articulatory and auditory components can be modified by a matrix of sociolinguistically determined functions. These are not specified in any detail but would presumably include such things as the role of linguistic markers in identifying group membership.

2. Thomason and Kaufman [39] argue for multiple language origins, making the point that creole languages have no single parent and hence their descendents are unrelated to other languages. They also argue that it is impossible to know how often the social circumstances that lead to formation of a creole may have occurred in the distant past. These may be valid points, although it is uncertain how often conditions for long-term survival of creole languages are likely to arise. However, in the creole languages of which we know the recent histories, the sound patterns are constructed out of material that is present in one or more of the 'input' languages. There is no reason to believe this would have been different at earlier times. Ancient creole languages would not represent independent language development in the sense we are concerned with here. They would reflect continuity of traits such as three-place stop systems.

ACKNOWLEDGMENTS

The work reported in this paper was supported in part by grant ROI DC00642 from the National Institutes of Health and grant BNS 8720098 from the National Science Foundation.

REFERENCES

[1] ABRY, C., L-J. BOE, & J-L. SCHWARTZ (1989), "Plateaus, catastrophes and the structuring of vowel systems," *Journal of Phonetics,* 17, 47-54.

[2] BELL, A. (1978), "Language samples." In J. H. Greenberg et al (eds.) *Universals of Human Language, Vol 1, Method and Theory,* Stanford: Stanford University Press, 123-156.

[3] BENEDICT, P. (1975), *Austro-Thai Language and Culture,* New Haven: HRAF Press.

[4] BENEDICT, P. (1988), *Japanese/Austro-Tai,* Ann Arbor: Karoma.

[5] BOYD, R. G. (1978), "À propos des ressemblances lexicales entre langues Niger-Congo et Nilo-Sahariennes." In *Études Comparatives (Bulletin de la SELAF,* 65), 43-94.

[6] BROWN, W. M. (1980), "Polymorphism of mitochondrial DNA of humans as revealed by restriction endonuclease analysis," *Proceedings of the National Academy of Sciences of the USA*, 77, 3605-3609.

[7] CAMPBELL, B. (1966), *Human Evolution,* Chicago: Aldine Press.

[8] CAMPBELL, L. (1988), "Review article: Language in the Americas, by Joseph H. Greeenberg," *Language,* 64, 591-615.

[9] CANN, R. L., M. STONEKING & A. C. WILSON (1987), "Mitochondrial DNA and human evolution," *Nature,* 325, 31-36.

[10] CROTHERS, J. (1978), "Typology and universals of vowel systems." In J. H. Greenberg et al (eds.) *Universals of Human Languages, Volume 2, Phonology,* Stanford: University of Stanford Press, 93-152.

[11] DIOP, C. A. (1988), *Nouvelles recherches sur l'egyptien ancien et les langues négro-africaines modernes,* Paris & Dakar: Presence Africaine.

[12] DRYER, M. (1989), "Large linguistic areas and language sampling," *Studies in Language, 13*, 257-292.

[13] GREENBERG, J. H. (1987), *Language in the Americas,* Stanford: Stanford University Press.