ADAPTIVE VARIABILITY AND ABSOLUTE CONSTANCY IN SPEECH SIGNALS: TWO THEMES IN THE QUEST FOR PHONETIC INVARIANCE

Björn Lindblom

Department of Linguistics, University of Stockholm, Sweden

ABSTRACT

Our topic is the classical problem of reconciling the physical and linguistic descriptions of speech: the invariance issue. Evidence is first presented indicating the possibility of defining phonetic invariance at the articulatory, acoustic or auditory levels of the speech signal. However, as we broaden the scope of our review, we find that attempts to define phonetic invariance in terms of absolute physical constancies tend to lose ground to theories that recognize signal variability as an essentially systematic and adaptive consequence of the informational mutuality of natural speaker-listener interactions. We reach this conclusion not only by examining experimental data on on-line speech processes but also by analyzing typological evidence on how the phonetic structure of consonant systems varies with inventory size in a lawful manner.

INTRODUCTION

Traditionally the problem of invariance in phonetics can be said to be that of proposing physical descriptions of linguistic entities that have the characteristic of remaining invariant across the large range of contexts that the communicatively successful real-life speech acts present to us.

Many of us share the conviction that taking steps towards the solution of this problem will be crucial if we are to acquire a deeper theoretical understanding of the behavior of speakers and listeners as well as develop more advanced systems for speech-based man-machine communication (Perkell&Klatt 1986).

The present paper will attempt to address some of the questions that we typically encounter in the search for invariance. We shall do so by summarizing research undertaken mostly in our own laboratory in Stockholm. Although thus deliberately limiting the scope of our review we hope that the issues raised will nevertheless be of sufficient interest to stimulate general discussion.

IS PHONETIC INVARIANCE ARTICULATORY?

A few decades ago phoneticians began to interpret phonetic events by comparing articulators to highly damped oscillatory systems. More recently, such models have acquired an important role within the framework of action theory (Kelso, Saltzman and Tuller 1986). In the sixties it was hoped that a lot of the variability that speech signals typically exhibit - e g reductions and vowel-consonant coarticulation (Öhman 1967) - could be explained in terms of the spatial and temporal overlap of adjacent "motor commands" (MacNeilage 1970). Articulatory movements were seen as sluggish responses to an underlying forcing function which was assumed to change, usually in a step-wise fashion, at the initiation of every new phoneme (Henke 1966). Owing to variations in say stress or speaking tempo, different contexts would give rise to differences in timing for a given sequence of phoneme commands. Articulatory and acoustic goals would not always be reached, the so-called 'undershoot' phenomenon (Stevens and House 1963). But since such undershoot appeared to be lawfully related to the duration and context of the gestures (Lindblom 1963), the underlying articulatory "targets" of any given phoneme - 'die Lautabsicht' - would nevertheless, it was maintained, remain invariant. Accordingly, at that time it seemed possible to argue that phonetic invariance might be articulatory.

Duration-dependent undershoot still seems to to be a phonetically valid notion for biomechanical reasons. But it is clearly not as inevitable a phenomenon as was first thought. Current experimental information indicates that in fast speech articulatory and acoustic goals can be attained despite short segment durations (cf Engstrand 1987, Gay 1978, Kuehn and Moll 1976). Furthermore undershoot has been observed in unstressed Swedish vowels that exhibit long durations owing to 'final lengthening' (Nord 1986). Such deviations from simple duration-dependence appear to highlight the reorganizational abilities of the speech production system. One way of resolving the problem posed by these somewhat contradictory results might be obtained if it were shown that when instructed to speak fast, subjects have a tendency to "overarticulate", thus avoiding undershoot to some extent, whereas when destressing they are more prone to "underarticulate" (cf discussion below of hypo- and hyper-speech). The demonstration of language-specific patterns of vowel reduction

(cf Delattre's 1969 discussion of English, French, German and Spanish) becomes particularly relevant in the context of addressing such questions.

In summary, the original observations of 'undershoot' carried the implication that the invariant correlates of linguistic units were to be found, not in the speech wave nor at an auditory level, but upstream from the level of articulatory movement. Phonetic invariance was accordingly associated with the constancy of underlying "spatial articulatory targets" (for reviews of the target concept see e g MacNeilage 1970, 1980). However, subsequent experimentation - some of which we already hinted at above - has revealed that the notion of segmental target must be given a much more complex interpretation.

This conclusion is reinforced particularly strongly by studies of **compensatory articulation**. Let us summarize some results from an experiment using the so-called "bite-block" paradigm (Lindblom, Lubker, Lyberg, Branderud, Holmgren in press). Native Swedish speakers were asked to pronounce monosyllables and bi- and trisyllabic words under two conditions: normally and with a large bite-block between their teeth. They were instructed to try to produce the bite-block utterances with the same rhythm and stress pattern as the corresponding normal items. Real Swedish words as well as "reiterant" nonsense forms were used: To exemplify, one of the metric patterns was: - '- - . This pattern would occur in the lists as "begabba" and /ba'bab:ab/. Measurements were made of the duration of the consonant and vowel segments of the normal and the bite-block versions of the reiterant speech samples. The question was thus whether subjects would be able achieve the bilabial closure for the /b/ segments in spite of the abnormally low and fixed jaw position and whether they would be able to do so reproducing the normal durational patterns.

We found that the timing in the bite-block words deviated systematically but very little from the normal patterns and concluded that our subjects were indeed capable of compensating. To explain the results we suggested that a representation of the "desired end-product" - the metric pattern of the word - must be available in some form to the the subjects' speech motor systems and that the successful compensations implied a reorganization of articulatory gestures that must have been controlled by such an output-oriented target representation. These results are in agreement with those reported earlier by Netsell, Kent and Abbs (1978). Moreover, they are completely analogous to the previous demonstrations that naive speakers are capable of producing isolated vowels whose formant patterns are normal at the first glottal pulse in spite of an unnatural jaw opening imposed by the use of a "bite-block" (Lindblom, Lubker and Gay 1979, Gay, Lindblom and Lubker 1981).

These results bear on the recent discussion of speech timing as "intrinsically" or "extrinsically" controlled. Proponents of action theory (Fowler, Rubin, Remez and Turvey 1980) approach the physics of the speech motor system from a dynamical perspective with a view to reanalyzing many of the traditional notions that now require explicit representation in extant speech production models such as 'feedback loop', 'target' etc. Their writings convey the expectation that many aspects of the traditional "translation models" will simply fall out as consequences of the dynamic properties intrinsic to the speech motor system. In the terminology of Kelso, Saltzman and Tuller (1986, 55) "...., both **time** and **timing** are deemed to be intrinsic consequences of the system's dynamical organization." Methodologically, action theory is commendable since, being committed to interpreting phonetic phenomena as fortuitous (intrinsic) consequences rather than as controlled (extrinsic) aspects of a speaker's articulatory behavior, it guarantees a maximally thorough examination of speech production processes. However, it is difficult to see how, applying the action theoretic framework to the data on compensatory timing just reviewed, we could possibly avoid postulating some sort of "temporal target" representation which is (i) extrinsic to the particular structures executing the gestures and which is (ii) responsible for extrinsically tuning their dynamics. Speech production is a highly versatile process and sometimes appears strongly listener-oriented.

The plasticity of the speech motor system is further illustrated by an experiment recently done by Schulman (forthcoming) invoking a "natural bite-block" situation. This condition is provided by loud **speech** in which a more open mandible tends to be used than in normally spoken syllables.

Whether rounded or not the vowels of loud test words produced by Schulman's talkers were found to exhibit almost three times as large jaw openings as the corresponding segments in the normal words. In the context of compensatory articulation two observations call for special comments. Why do not speakers compensate for the greater jaw opening in the loud vowels the way they do in the bite-block experiments? Schulman shows that they do not since the fundamental frequency and (as predicted by articulatory-acoustic nomograms) the first formant of the loud vowels are shifted upwards by about one Bark whereas the other formants do not undergo comparable modification. (Below we shall relate the F1 and F0 shift to the results of a perceptual experiment).

The other finding of interest is the fact that loud vowel durations increase whereas loud consonant durations tend to decrease (cf Fonágy and Fonágy 1966). What does that result mean? The normal-loud vowel duration differences look suspiciously similar to the durational differences between normal open and close vowels which have been observed for many languages (Lehiste 1970). Finding that the duration of the EMG recorded from the anterior belly of the digastric correlated with both mandibular displacement and vowel duration Westbury and Keating (1980) suggest that this temporal variation among vowels, although non-distinctive, must be seen as present in the neuromuscular signals controlling their articulation. An alternative interpretation would be to regard the differences as automatic consequences of an interaction between an invariant underlying "vowel duration command" and articulatory inertia (cf Keating 1985 for further discussion). In (Lindblom 1967) we reported some evidence in favor of the latter interpretation, the "extent of movement hypothesis" (Fischer-Jørgensen 1964). We also found that the durational consequences of more extensive articulatory gestures were sometimes actively counteracted.

The question whether the open-close vowel duration difference is an intrinsic or extrinsic phonetic phenomenon is accordingly somewhat controversial. Schulman's findings bear on the problem. He constructed a model of loud speech based on the observation that loud movements appear to be "exaggerated" versions of the normal movements. Assuming that the lips and the jaw are linear mechanical systems and that loud differs from normal speech solely in terms of the amplitudes of the underlying excitation forces he performed a linear scaling of all articulatory parameters recorded for normal syllables (vertical displacements of upper and lower lips and jaw) and combined the scaled curves so as to derive the vertical separation of the lips - the parameter that determines the open-closed state of the mouth opening. By using the value of this parameter at opening and closing in the normal syllables as his criterion he was then able to predict the durations of vowel and consonant segments for loud speech. He found that linear scaling eliminated stop closures entirely or produced much too long vowels.

The implication of this result is that it clearly attributes the durational differences to a superposition effect, that is the interaction arising from the superposition of the lip and the jaw movements. Schulman concludes that, unless the effect of opening and closing of the jaw had been actively counteracted, loud and normal vowel durations would have differed even more than they actually did.

Let us remark in the present context that, while it appears reasonable to suggest, as do Westbury and Keating, that the acoustic vowel duration differences are probably reflected at a level of neuromuscular control, there is also evidence indicating that the function of neural control signals may be a compensatory rather than a positive one, that is a function opposite to that suggested by Westbury and Keating.

The preliminary implication of all work touching the theme of compensatory articulation appears to be that - whether we use "target" with reference to segmental attributes, segment durations or patterns of speech rhythm - the term is better defined, not in terms of any simple articulatory invariants, but with respect to the acoustic output that the talker wants to achieve. If phonetic invariance is not articulatory could it be acoustic then?

IS PHONETIC INVARIANCE ACOUSTIC?

The suggestion that the speech signal contains absolute physical invariants corresponding to phonetic segments and features has received a lot of attention thanks to the work by Stevens and Blumstein (Stevens and Blumstein 1978, 1981; Blumstein and Stevens 1979, 1981). The idea has been favorably received by many, for instance Fowler in her attempts to apply the perspective of **direct perception** to speech (Fowler 1986).

Others have been provoked to emphasize the inadequacy of the non-dynamic nature of the Stevens template notion (Kewley-Port 1983) and the substantial context-dependence that the stop consonants of various languages typically display even in samples of carefully enunciated speech (Öhman 1966).

Recent work by Krull and Lacerda in our Stockholm laboratory uses the method of quantifying the extent of consonant-vowel coarticulation in the form of linear "locus equations". These relationships are obtained by plotting formant frequencies at CV_2- and V_1C-boundaries as a function of the formants for V_2 and V_1 respectively. Acoustic theory indicates that for the consonant-vowel combinations in question near-linear relationships should be expected. Such diagrams show clearly that, although a "locus" pattern can exhibit considerable variation, it is predictable from information on stop consonant identity and adjacent vowel context. Here coarticulation stands out as the salient fact and the lack rather than the presence of absolute acoustic invariance tends to be reinforced.

Incidentally, let us note that, if it exists, acoustic invariance is a strange notion since talkers can only monitor it through their senses and listeners can only access it through their hearing system. Why should sensory and auditory transduction be assumed to have a transfer function of one imposing no transformation? Is it the case that what people really mean when they talk about acoustic invariance is in fact "auditory" invariance? Let us look at some psycho-acoustic results.

IS PHONETIC INVARIANCE AUDITORY?

We mentioned earlier a perceptual result that offers a rather curious parallel to Schulman's findings. It is the "Traunmäller effect" which is a demonstration of the transforms required to preserve the perceptual constancy of vowel quality under changes in (i) vocal effort and (ii) vocal tract size. It is also somewhat reminscent of the findings on F0-Fi interrelationships in soprano vowels (Sundberg 1975).

Effort and vocal tract variations can be dramatically illustrated by synthetically modifying a naturally spoken /i/. When all formants and F0 are shifted equally along a Bark scale an /i/-like vowel is perceived but the voice changes from an adult's to a child's. When both Fi and F0 are varied in such a way that Fi-F0 is kept constant on a Bark scale - and the upper formant complex is left unchanged - an /i/-like vowel is perceived. This is remarkable in view of the fact that Fi reaches a value more typical of a low-pitched /ɑ/. One's impression is that the speaker remains the same but that she "shouts".

Note the parallel between Schulman's and Traunmäller's results. Are the findings causally related? Do we explain the lack of formant compensation in loud speech in terms of the Traunmäller effect? Or do we account for the vowel quality results in terms of the "Schulman" effect?

Of importance for the present discussion is the fact that behavioral constancies have been demonstrated and that they imply that at least in this case phonetic invariance must be defined at a level of auditory representation.

Let us return for a moment to the alleged invariance of the release spectra of stop consonants. Diana Krull collected perceptual responses from Swedish listeners to burst fragments obtained from $V_1C:V_2$ words (Krull 1987). One hundred test words were generated by constructing all possible combinations of V_1 or V_2 = short /i e a o u/ with C: = /b: d: rd: g:/. Confusion matrices for the burst stimuli demonstrate the drastic coarticulation effects. By and large, listener responses can be accounted for in terms of the acoustic properties of the stimuli. This is shown in her attempts to predict the confusions from auditorily based "perceptual distance" computations.

A related study has been carried out by Lacerda (1986). We can characterize one part of his research as variations on the theme struck by Flanagan in his early "difference limen" experiments on vowel formant frequencies (Flanagan 1955). Lacerda's question was: How well can listeners discriminate four-formant stimuli that differ solely in terms of the frequency of F2. His work permits us to compare a psycho-acoustic task: the discrimination of F2 in brief tone bursts with formant patterns static - with a "speech task": the discrimination of the onset of F2-transitions in /da/-stimuli.

The results indicate that the subjects' ability to discriminate on the psycho-acoustic task is in close agreement with Flanagan's findings whereas their performance on the /da/-stimuli is drastically impaired. One interpretation is that the discrimination change is related to the fact that intra-category discrimination is considerably worse than inter-category discrimination (Liberman, Harris, Hoffman and Griffith 1957).

With reference to the invariance issue it is important to note the following. Krull's results on stop perception indicate that the coarticulatory spectral variability of the stop releases is rather accurately reflected in the confusions that her listeners made of such brief sounds. This is fully compatible with Lacerda's results on tone bursts. Note that in Lacerda's speech-task test however, the variability does not seem to be as faithfully mirrored in the listeners' percepts for apparently they treat stimuli easily discriminable in psycho-acoustic tests as "the same". Whether it is the listener invoking the "speech mode" or it is the interaction of the dynamic stimulus properties and speech-independent auditory processing is an issue still worth addressing. However, our main point is this: The invariance that we discern in these findings is not acoustic. It clearly presupposes auditory processing.

IMPLICATIONS OF SPEAKING STYLE: THE HYPER-HYPO DIMENSION

Everyday experience indicates that speaking is a highly flexible process. We are capable of varying our style of speech from fast to slow, soft to loud, casual to clear, intimate to public. We speak in different ways when talking to foreigners, babies, computers and hard of hearing persons. And we change our pronunciation as a function of the social rules that govern speaker-listener interactions (Labov 1972).

Above we considered principally three types of phonetic invariance: articulatory, acoustic and auditory invariance. What are the implications of variations in speaking style for the invariance issue? For the purpose of our discussion let us give phonetic invariance a strong literal interpretation which is rather extreme but nevertheless not too far from working hypotheses explored previously by various investigators: "All the information is in the signal, particularly in its dynamics". For such a view of invariance to be correct - let us call it the strong version of absolute physical invariance - the following must be true: Talkers vary their speaking style and thereby contribute to increasing the variability of the speech wave but in utterances that are intelligible linguistic units will always exhibit a core of invariant physical information that will remain undestroyed so as to be successfully used by a listener.

We recently undertook a literature survey+ in order to systematize the types of speech materials that have been used in acoustic phonetic studies published during the past ten years in J Acoust Soc Am, J of Phonetics, Language and Speech, and Phonetica. A total of over 700 articles were selected as preliminarily relevant. We ended up choosing 216 as meeting our criterion of "descriptive study of speech based on quantitative acoustic phonetic measurements".

Of special interest to us was to ascertain the relative proportions of studies investigating "self-generated" speech (including e g spontaneous conversation) on the one hand and speech samples chosen by the experimenter (e g list readings, nonsense words etc) on the other. Not surprisingly, we found that the majority of studies, over 90%, use experimenter-controlled speech samples. The reason is clear. A satisfactory experimental design presupposes good control of the variables involved. This is less of a problem if the experimenter determines the test items but for "real speech" with its immense number of variables there is no established methodology that will guarantee such control. So rather than drown in an ocean of "unknown factors" our strategy tends naturally to become one of resorting to "given" test materials and read speaking modes.

One way of justifying this widely used procedure is to argue that first we will solve the problem of phonetic invariance in "lab speech". Then we will get to work on "natural speech". Another outlook might be to suggest that, although we lack the supplementary methodology required by "ecological" speech, the excessive use of "lab speech" introduces an undesirable bias in our data bases as well as in our theoretical intuitions about invariance and other key issues - a bias that might make us underestimate the problem of speech variability in spite of the fact that it is readily acknowledged by all workers in the field and has already, it would appear, been rather massively documented. Consequently the situation ought to be balanced.

We have recently been persuaded by the latter point of view and are currently recording (1) "self-generated speech" produced under natural conditions and (2) parallel "citation form" speech based on the syllables, words and phrases that occur in the spontaneous materials. Data are currently being collected by Rolf Lindgren, Diana Krull and myself using this two-pronged approach involving comparisons of reference pronunciations ("citation form" speech) with samples of "self-generated speech". A few preliminary observations can be made that bear on the present discussion (cf also Lindblom and Lindgren 1985).

The reductions that we have found in spontaneous speech - and often escape the trained phonetic ear even after spectrographic evidence has been examined - are sometimes drastic. Speaking style has marked effects on the acoustic patterns of words. The vowel space shrinks in casual style and is expanded in "hyperspeech" modes. The diphthongization of tense Swedish vowels is enhanced and is particularly apparent in clear speech. Contrast in VOT for voiced and voiceless stops increases and decreases as we compare hyper- and hypo-forms respectively. Locus equations show a smaller slope (=less vowel-dependence) for citation form pronunciations than for spontaneous speech which we interpret to indicate that vowel-consonant coarticulation is counteracted in hyperspeech (more invariance) but tolerated in hypospeech (less invariance). Although preliminary the observations made so far suggest that the prospects for any strong version of absolute physical invariance to be substantiated seem most unfavorable.

SPEECH UNDERSTANDING: (IN)DEPENDENCE OF SIGNAL INFORMATION

At the Department of Romance Languages at Stockholm University a test is used to measure how proficient native Swedish students are in understanding spoken French in which the task of the students is to listen to triads of stimuli consisting of two identical sentences and one minimally different and to indicate the odd case.

Montre leur ce chapeau s'il te plaît
Montre leur ce chapeau s'il te plaît
Montre leur ces chapeaux s'il te plaît

Native speakers of French have no problems of course with such sentences whereas Swedish listeners knowing no French have a lot of trouble. However, when the key information - e g the ce/ce/ces triad - is presented as fragments gated from the original sentences the performance of the Swedish subjects improves radically (Dufberg and Stöök forthcoming).

This test can serve to remind us that perception is a product of two things: signal-dependent and signal-independent information. While I am perfectly capable of discriminating the French minimal contrasts as auditory patterns I would quickly lose those patterns in a sentence context unless I have a sufficiently good command of French - that is access to signal-independent 'knowledge' whose interaction with the signal is a part of forming of the final percept.

The speech literature is full of experimental data indicating that processes not primarily driven by the signal play an

---

+I am indebted to Diana Krull for doing the preliminary selections and to Natasha Beery of the Phonology Laboratory, University of California, Berkeley for the statistical analyses.

important role in the perception of speech. There will not be time to do justice to all the research bearing on this issue. Let me just recall some well-known paradigms: Perception of speech in the presence of various disturbances (noise and distortion). The improvement of identification as the signal gets linguistically richer (Miller, Heise and Lichten, Pollack and Pickett 1964 and by Miller&Isard). Detection of deliberate mispronunciations (Cole 1973). Word frequency effects (Howes, Savin). Restoration (Warren 1970, Ohala and Feder 1986). Phoneme monitoring (Foss&Blank). Word recognition from word fragments (Grosjean 1980, Nooteboom 1981). Fluent restorations in shadowing mispronunciations (Marslen-Wilson and Welsh 1978). Verbal transformations (Warren). Intelligibility of lip-reading from video-recordings supplemented by "hummed speech" - an audio signal processed to contain primarily rhythm and intonation cues (Risberg 1979). Inferences from historical sound changes (Ohala 1981).

CONCEPTUALIZING SPEAKER-LISTENER INTERACTIONS

Our review of experimental evidence bearing on the invariance issue has been selective but should nevertheless provide a rough indication of a panoply of alternative positions and their respective pro's and con's. We have considered the suggestion that the invariance of phonetic segments be defined: (i) at an articulatory level (e g the "spatial target" hypothesis); (ii) at an acoustic level (e g spectral properties of stops); (iii) at an auditory level (e g perceptual constancy of vowel quality). Which of these alternatives should we put our money on?

When pursued experimentally articulatory, acoustic or auditory definitions of invariance have the methodological virtue of encouraging a maximally thorough search at these particular levels. But in seeking a broader theoretical understanding of speech communication we would stand little to gain from spending effort on choosing between levels. Such an appoach misreads the evidence which, when viewed in a broader perspective, strongly points to the conclusion that: The invariance problem is not a phonetic issue at all for ultimately invariance can be defined only at the level of listener comprehension.

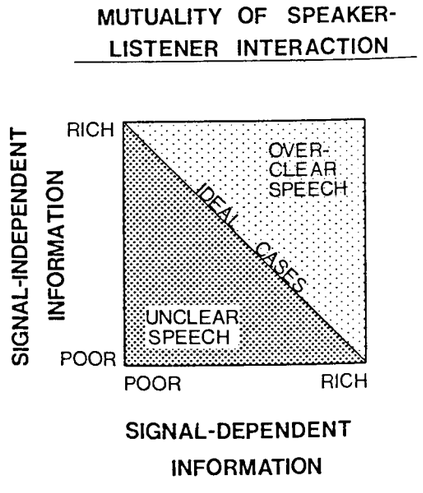
We can convince ourselves of the correctness of that point by considering the following phrase in English: /lesnsevn/. We can hear this utterance either as LESS THAN SEVEN, or as LESSON SEVEN. In the appropriate contexts (say "How many are coming", and "What is our topic to-day?") the listener will not be aware of any ambiguity. At which phonetic level do we find the physical correlates of the initial segments of the word "than"? Needless to say there ARE no such correlates in this particular case. The conclusion seems inescapable: We should not put our money on any of the above alternatives. We must seek a more general theory.

The experimental data on production indicates that the behavior of the speech motor system is shaped primarily by two forces - plasticity (listener-oriented reorganization) and economy (talker-oriented simplification) - which interact on a short-term basis so as to generate signals that may be "rich or poor" in explicit physical information.

The evidence on perception has identified two major sources of information: signal-dependent and signal-independent processes and suggests that on a short-term basis percepts arise from the latter (i e "context") modulating the former in an analogously "rich or poor" manner.

One possible way of schematizing the logical possibilitites of these conceptual simplifications is shown in the diagram of the enclosed figure. This is not a very rigorous scheme but seems useful, at least pedagogically, in contrasting some of the ideas currently entertained in phonetics (cf J of Phonetics, January issue 1986).

This graph states that for speech to be intelligible the sum of explicit physical information and signal-independent information must be above a threshold, that is the 135 degree line. In the ideal case this sum equals a constant the x- and y-values of specific speech samples falling right on that line. Points above the line are associated with what might be termed "over-clear" speech, points below it with "unintelligible" speech.



It appears reasonable to assume that in the real-life situations utterances can vary tremendously with respect to how socially and communicatively successful they prove to be. For our present purposes let us focus on speech samples from hypothetically successful real-life speaker-listener interactions and assume that they produce data points clustering near and above the slant line. What would such a result imply? It would mean that there is a complementary relation between the amounts of information contributed by signal attributes on the one hand and "context" on the other. When speakers come close to the slant line it would indicate first of all that they are capable of varying their speech output in a plastic way (cf evidence on hypo-hyper-speech modes and other instances of reorganization of speech motor control) and secondly that, while perhaps not being perfect 'mind-readers', they are at least capable of adapting their speech on-line to the short-term fluctuations in the listener's access to "context" or signal-independent information (cf experimental documentation of numerous cases showing that listeners are in fact capable of successfully coping with highly context-dependent reduced and coarticulated speech stimuli). The possibility of such complementarity in real speech emerges also from some recent measurements reported by Hunnicutt (1985) as well as from Lieberman's 1963 study.

If we hypothesize that this strategy - let us call it the STRATEGY OF ADAPTIVE VARIABILITY - comes near the way real speakers actually behave when they are communicatively successful, we obtain a natural way of resolving some of the paradoxes that surround the invariance issue. For it follows that intra-speaker phonetic variation - along a hyper-hypo-continuum as well as along other dimensions - is the characteristic that we should expect the units of ecological speech to exhibit - not absolute physical invariance.

The proposed way of thinking about the issue does not, of course, rule out finding physical speech sound invariance in restricted domains of observation but it does explain why our quest for a general concept of phonetic invariance has been largely unsuccessful. And, in a pessimistic vein, it predicts in fact that it will continue to be so.

Our reasoning leads us back to a conclusion already drawn by MacNeilage in his 1970 review of the invariance issue:

> "...the essence of the speech production process is not an inefficient response to invariant central signals, but an elegantly controlled variability of response to the demand for a relatively constant end (p 184)".

If, as suggested here, we take the "relatively constant end" to be defined neither articulatorily, acoustically nor auditorily but specified only with reference to "the level of listener comprehension" MacNeilage's formulation still captures the "essence of the speech production process" satisfactorily.

Let us pause to reflect on some of the implications of the two theories contrasted in our discussion: Absolute Physical Invariance versus Adaptive Variability. The former, if proved correct, would transform what currently looks like instances of massive variability into artefacts. For this theory says in fact that there simply IS NO variability of linguistic units; There seems to be but that is merely a result of our presently inadequate conceptual and experimental tools. Further note that if we push the notion of absolute constancy to its extreme another implication can be noted, namely that the transmission of information by speech - an undeniably biological process - is basically non-adaptive.

The Theory of Adaptive Variability, on the other hand, says exactly the opposite. This is a theory for which it is easier to find support within the general study of the biology of motor control and perception. It is precisely by emphasizing the adaptive nature of speech processes that we obtain a principled way of investigating phonetic variation and its origin.

ON-LINE PROCESSES IN THE LIGHT OF TYPOLOGICAL EVIDENCE ON CONSONANT SYSTEMS

Some time ago Nooteboom did an experiment on word retrieval and was able to show that listeners perform better if presented with the first halves of words than on the corresponding second-half fragments (Nooteboom 1981). For an explanation he suggested that, since word recognition is a real-time left-to-right process, word beginnings are less predictable than word endings. Consequently left-to-right context can be much more easily used than right-to-left context.

He concluded his paper by raising the question whether this asymmetry - that he takes to be a universal feature of the perceptual processing of any language - might have left its imprint on how lexical information is organized in the languages of the world. He predicted (p 422) that: "(1) in the initial position there will be a greater variety of different phonemes and phoneme combinations than in word final position, and (2) word initial phonemes will suffer less than word final phonemes from assimilation and coarticulation rules."

One basic assumption is that variations in perceptual predictability correlate with signal "distinctiveness". Hence "the greater variety of different phonemes and phoneme combinations" in the initial as compared with the final position of words. Restating the

idea we can say that a larger paradigm goes with a RICHER signal inventory. The other side of the coin is of course that a smaller paradigm - such as that attributed to word endings - goes with a POORER signal inventory. In suggesting that the presence of assimilation and coarticulation should vary inversely with the need for keeping items distinct Nooteboom tacitly formulates a hypothesis that comes close to the theory of Adaptive Variability described here. Note that the theory Absolute Physical Invariance does not offer us any basis at all for making predictions about a possible interplay between language structure and on-line processing. Why? As stated earlier according to that theory there IS no phonetic variation, there only seems to be. The idea of language structure adapting to the on-line constraints of speaking and listening only becomes a possibility once we recognize the existence and systematic nature of phonetic variation. Only from that point of departure will we be able to address the question of what feeds the processes of phonological innovation.

We shall not be in a position to present the typological data needed to test Nooteboom's hypothesis. However, we shall conclude our paper by presenting some other data that do bear on it and strongly encourage further examination of the underlying ideas.

In collaboration with Ian Maddieson we recently undertook an analysis of the consonant inventories of 317 languages, carefully selected so as to constitute a reasonable sample of the "languages of the world". Our corpus was that of UPSID, the UCLA Phonetic Segment Inventory Database (Maddieson 1984). The data consists of lists of systems whose elements (allophones of major phonemes) are specified in phonetic transcription.

Inventory sizes range from 6 to 95 consonants per system. The materials lend themselves to testing a paraphrase of Nooteboom's hypothesis: Is the phonetic structure of consonant systems independent of their size? Or is it systematically related to that dimension? If there is a systematic size-dependence what is it?

There is neither time nor space to give the details of the analysis. They will be published elsewhere (Lindblom, MacNeilage and Studdert-Kennedy; Lindblom and Maddieson forthcoming). Fortunately, Nooteboom's perspective provides us with a way of summarizing the main findings.

It turns out that small paradigms statistically favor segments with both phonatory and articulatory properties that can be classified as basic or elementary. Medium-sized paradigms tend to include consonants invoking more elaborated gestures in addition to a core of basic elements. The largest systems use both these types but also combinations of elaborated gestures that we label complex articulations. To exemplify, plain /p t k/ are classfied as "basic" articulations whereas ejective /pˀ tˀ kˀ/ or aspirated /pʰ tʰ kʰ/ invoke "elaborated" mechanisms. A segment such as /t͡sʰ/ is "complex" since it shows more than one elaboration: both of place (retroflexion) and source features (aspiration). Logically a six-consonant system could use the ejective set for its stop series. Small systems never do in our material whereas medium-sized and large systems do. Moreover, the "complex", multiply elaborated segments are most frequent in the large inventories. The basic rule is that a less simple consonant tends not to be recruited without the presence of parallel more simple ("basic" or "elaborated") series (cf the notion of 'implicational hierarchy' of traditional terminology). The claim we make is accordingly that we see a positive correlation between paradigm size and the number of elements that a sound pattern selects from a dimension of "articulatory complexity".

The validity of our analysis naturally hinges on the success with which we can give non-circular, independently motivated definitions of "articulatory complexity". When it comes to the details of the analysis that problem is a topic for future quantitative phonetic theory. For the moment we believe that the major trends are rather gross effects that can be convincingly demonstrated by the force of the examples. They permit us to make the following generalization: Small consonant paradigms invoke 'unmarked' phonetics, large paradigms 'marked' phonetics. That is of course exactly what Nooteboom's hypothesis predicts and it takes a few steps towards an explanation for why seven-consonant systems do not show inventories like the following (Ohala 1980):

[ɗ kʼ tsʲ ɬʲ m r ɬ]

We take the present typological data on consonant systems as providing strong evidence in favor of (a) language structure evolving as an adaptation to the constraints of the on-line processes of speaker-listener interaction; and for (b) the correctness of a theory of Adaptive Variability as an account of those processes.

REFERENCES

Blumstein S and Stevens K N (1979): "Acoustic Invariance in Speech Production: Evidence from Measurement of the Spectral Characteristics of Stop Consonants", J Acoust Soc Am 72, 43-50.

Blumstein S and Stevens K N (1981): "Phonetic Features and Acoustic Invariance in Speech", Cognition 10, 25-32.

Cole R A (1973): "Listening for Mispronunciations: A Measure of What We Hear during Speech", Perception and Psychophysics 13, 153-156.

Delattre, P (1969): "The General Phonetic Characteristics of Languages: An Acoustic and Articulatory Study of Vowel Reduction in Four Languages", Mimeographed Report, University of California, Santa Barbara.

Engstrand, O (1987): "Articulatory Correlates of Stress and Speaking Rate", accepted for publication in J Acoust Soc Am.

Flanagan, J (1955): "A Difference Limen for Vowel Formant Frequency", J Acoust Soc Am 27:613-614.

Fischer-Jørgensen E (1964): "Sound Duration and Place of Articulartion", Zeitschrift für Sprachwissenschaft und Kommunikationsforschung 17:175-207.

Fonágy I and Fonágy J (1966): "Sound Pressure Level and Duration", Phonetica 15:14-21.

Fowler C A, Rubin P, Remez R E and Turvey M T (1980): "Implications for Speech Production of a General Theory of Action", 373-420 in Butterworth, B (ed): Language Production, vol I, London:Academic Press.

Gay, T (1978): "Effect of Speaking Rate on Vowel Formant Movements", J Acoust Soc Am 63(1):223-230.

Gay T, Lindblom B and Lubker J (1981): "Production of Bite-Block Vowels: Acoustic Equivalence by Selective Compensation", J Acoust Soc Am 69(3), 802-810.

Grosjean, F (1980): "Spoken Word Recognition and the Gating Paradigm", Perception and Psychophysics 28, 267-283.

Henke, W J (1966): Dynamic Articulatory Model of Speech Production Using Computer Simulation, Doctoral dissertation, M.I.T.

Hunnicutt, S (1985): "Intelligibility versus Redundancy - Conditions of Dependency", Language and Speech 28(1):47-56.

Keating, P (1985): "Universal Phonetics and the Organization of Grammars", 115-132 in Fromkin, V A (ed): Phonetic Linguistics, Orlando, FL:Academic Press.

Kelso J A S, Saltzman, E L and Tuller, B (1986): "The Dynamical Perspective on Speech Production: Data and Theory", J of Phon 14:1, 29-59.

Kewley-Port, D (1983): "Time-varying Features as Correlates of Place of Articulation in Stop Consonants", J Acoust Soc Am 73:322-355.

Krull, D (1987): "Evaluation of Distance Metrics Using Swedish Stop Consonants", paper submitted to the XIth ICPhS, Tallinn, Estonia.

Kuehn, D P and Moll, K L (1976): "A Cineradiographic Study of VC and CV Articulatory Velocities", J of Phon 4:303-320.

Labov, W (1972): Sociolinguistic Patterns, Philadelphia:University of Pennsylvania.

Lehiste, I (1970): Suprasegmentals, Cambridge, MA:MIT Press.

Lieberman, P (1963): "Some Effects of Semantic and Grammatical Context on the Production and Perception of Speech", Language and Speech 6:172-187.

Liberman A M, Harris K S, Hoffman H S and Griffith B C (1957): "The Discrimination of Speech Sounds within and across Phoneme Boundaries", J of Experimental Psychology 54:358-368.

Lindblom, B (1963): "Spectrographic Study of Vowel Reduction", J Acoust Soc Am 35:1773-1781.

Lindblom, B (1967): "Vowel Duration and a Model of Lip Mandible Coordination", STL-QPSR 4/1967, 1-29 (Dept of Speech Communication, RIT, Stockholm).

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", J of Phonetics 7, 147-161.

Lindblom B, Lubker J, Lyberg B, Branderud P and Holmgren K (in press): "The Concept of Target and Speech Timing", to appear in Festschrift for Ilse Lehiste.

Lindblom, B and Lindgren R (1985): "Speaker-Listener Interaction and Phonetic

Variation", Perilus IV, Dept of Linguistics, University of Stockholm.

Lindblom B, MacNeilage P and Studdert-Kennedy M (forthcoming): Evolution of Spoken Language, Orlando, FL:Academic Press.

Lindblom, B and Maddieson, I (1988): "Phonetic Universals in Consonant Systems", to appear in Hyman, L M and Li, C N (eds): Language, Speech and Mind, Croom Helm.

MacNeilage, P (1970): "Motor Control of Serial Ordering of Speech", Psychological Review 77:182-196.

MacNeilage, P (1980): "Speech Production", Language and Speech 23(1), 3-24.

Maddieson, I (1984): Patterns of Sound, Cambridge:Cambridge University Press.

Marslen-Wilson, W D and Welsh, A (1978): "Processing Interactions and Lexical Access during Word Recognition in Continuous Speech", Cognitive Psychology 10, 29-63.

Netsell R, Kent, R and Abbs J (1978): "Adjustments of the Tongue and Lip to Fixed Jaw Positions during Speech: A Preliminary Report", Conference on Speech Motor Control, Madison, Wisconsin.

Nooteboom, S G (1981): "Lexical Retrieval from Fragments of Spoken Words: Beginnings vs Endings", J of Phonetics 9, 407-424.

Nord, L (1986): "Acoustic Studies of Vowel Reduction in Swedish", STL-QPSR 4/1986, 19-36 (Dept of Speech Communication, RIT, Stockholm).

Ohala, J J (1980): "Chairman's Introduction to Symposium on Phonetic Universals in Phonological Systems and their Explanation", 184-185 in Proceedings of the IXth International Congress of Phonetic Sciences 1979, Institute of Phonetics, University of Copenhagen.

Ohala, J J (1981): "The Listener as a Source of Sound Change", 178-203 in Masek, C S, Hendrick, R A and Miller, M F (eds): Papers from the Parasession on Language and Behavior, Chicago:Chicago Linguistic Society.

Ohala, J J and Feder, D (1986): "Speech Sound Identification Influenced by Adjacent "Restored" Phonemes", J Acoust Soc Am 80.S110.

Öhman, S (1966): "Coarticulation in VCV Utterances: Spectrographic Measurements", J Acoust Soc Am 39:151-168.

Öhman, S (1967): "Numerical Model of Coarticulation", J Acoust Soc Am 41:310-320.

Perkell, J and Klatt, D (1986): Invariance and Variability in Speech Processes, Hillsdale, N J:LEA.

Pollack, I and Pickett, J M (1964): "Intelligibility of Excerpts from Fluent Speech: Auditory vs Structural Context", J Verb Learn and Vert Beh 3:79-84.

Risberg, A (1979): Doctoral dissertation, RIT, Stockholm.

Schulman, R (forthcoming): "Articulatory Dynamics of Loud and Normal Speech", submitted to J Acoust Soc Am.

Stevens, K N and House A S (1963): "Perturbation of Vowel Articulations by Consonantal Context: An Acoustical Study", J Speech & Hearing Res 6:111-128.

Stevens K N and Blumstein S (1978): "Invariant Cues for Place of Articulation in Stop Consonants", J Acoust Soc Am 64, 1358-1368.

Stevens K N and Blumstein S (1981): "The Search for Invariant Correlates Phonetic Features", in Eimas, P and Miller J (eds): Perspectives on the Study of Speech, Hillsdale, N J:LEA.

Sundberg, J (1975): "Formant Technique in a Professional Singer", Acustica 32(2), 89-96.

Traunmüller, H (1981): "Perceptual Dimension of Openness in Vowels", J Acoust Soc Am 69, 1465-1475.

Warren, R (1970): "Perceptual Restoration of Missing Speech Sounds", Science 167, 392-393.

Westbury, J and Keating P (1980): "Central Representation of Vowel Duration", J Acoust Soc Am 67, Suppl 1, S37 (A).