Keynote address

Phonetics and Speech Technology

Gunnar Fant
Stockholm, Sweden

1. Introduction

It is my privilege to address to you on a subject fundamental to our congress - phonetics and speech technology. The close ties and mutual dependencies inherent in the history of speech research and in the last decades of intense developments are apparent: Phonetics has attained a technical profile and speech technology has to rely on phonetics to achieve its advanced goals. This is, of course, an interdisciplinary venture also involving the entire field of speech research indepent of faculty. Instead of speaking about phonetics and speech technology, we could make a distinction between theory and applications and point to the development of handicap aids and new methods of clinical diagnosis and rehabilitation, teaching aids, etc. which add to the specialities represented at this congress. I shall make some general comments about this symbiosis and how it affects speech technology and phonetics. I shall also give my view on the general outlooks for the field, and on some of our problems and current research issues.

In the last decade we have experienced a revolution in computer technology and microelectronics that has paved the way for speech technology. There has been a breakthrough in the data handling capacity allowing very complex processing to be performed in small chips that can be produced at a low price in large quantities. There have also been reasonable advances in speech synthesis and speech recognition techniques which have opened new markets. This has created a boom of industrial expectations, a feeling of surfing on a high wave of technological developments towards the fully automated society where we may converse with computers as freely as with human beings. One expression for this optimistic trend is the Japanese national effort in computing and artificial intelligence which they refer to as the development of the 'Fifth generation of computers' which shall include language translation and speech input and output.

Electronic industry has promoted several large-scale marketing reports with prospects for billion dollars sales at the end of the century.

Will all these expectations come through? I am not the one to judge but there is certainly room for some scepticism or at least caution. The rate of increase of the world market has not progressed at the expected rate. So - the surf on the tidal wave of expectations may end in a brake when we are

confronted with the reefs of the knowledge barrier, Fig. 1. I am referring to our still meager insight in speech as a language code. We need a fifth generation of speech scientists rather than a fifth generation of computers.

A stagnation of advanced speech technology products and the marketing of cheap, lower performance products may discredit the field. You frequently hear comments such as: 'Speech synthesis by rule has now existed for several years but the quality is still questionable and the rate of improvement is low. Will it ever reach an acceptability for public use?' To make speech recognition really useful we must first learn to handle connected speech with relatively large vocabularies in a speaker-independent mode. Indeed, we are far off from such advanced levels of recognition techniques whilst there appear to exist potentialities for reaching a substantial improvement in the quality of synthetic speech within the next few years. The latter optimistic

Figure 1. Speech technology and the knowledge barriers.

opinion is shared by the pioneer in speech synthesis, John Holmes, in his report to this Congress and he also expects significant advances in the handling of connected speech to appear fairly soon.

To the optimistic view we could also add that text-to-speech synthesis already in the present state of the art has opened up new effective means of communication for handicapped, e.g. text-reading aids for the blind and speech protheses for speech handicapped. Also the performance is quite adequate for many special-purpose applications including computer-aided teaching. The Swedish text-to-speech system developed by Carlson and Granström is implemented with a single chi for the terminal synthesis and has an option for operating in six different languages. A similar text-to-speech system developed by Dennis Klatt at MIT has means for changing the speaker type from male to female to child. A flexible choice of speaker type will be quite important in the marketing of synthesizers but this is an area in which we still have much to learn.

There exists a variety of less advanced and cheaper synthesis systems, generally intended for phonetic symbol input but some also capable of handling a proper orthography text input. These devices provide a lower-quality speech. In general, even our best text-to-speech systems are fatiguing to listen to if used for reading long texts.

A substantial part of the speech output market is talking chips which serve as low data-rate recording and play-back systems. They are now introduced in automobiles, household appliances, watches, calculators, and video games. We might even anticipate a sound pollution problem from synthetic voices guiding every step of our daily life.

At present, toy industry and manufacturers of video games have employed phonetic experts to tailor talking chips to simulate special voice types and speaking manners. In the future I believe we can do this more or less by rules. General purpose text-to-speech systems are expected to improve sufficiently in performance to compete with speech coding and concatenating systems, at least when a certain flexibility is desired.

Computer speech input, i.e. speech recognition systems are expected to develop a greater market than speech output systems, at least in terms of sales value. Although we are far off from very advanced speech recognition systems, we might soon expect applications in office automation, e.g. as voice input for word processing systems. An extension of present techniques to handle connected sequences of words would facilitate this application. A speech synthesis monitoring feature could be included.

2. The Computerized Phonetics

The close ties between phonetics and speech technology are apparent. Phonetics has been computerized and has gained new efficient instrumentation and advanced speech processing methods. Of course, computers would have found their way to phonetics anyway but phonetics has now attained some-

what of a technical profile. The more prominent phonetic departments have a staff of engineering assistants and a research budget which was unheard of in former days' humanity faculties but this development has, of course, not come about without an intense engagement of people involved. Phonetics of today has gained a new respect from its vital role in the ever increasing importance of research into human functions. The technical profilation is also apparent in any speech research laboratory whether it is an outgrowth of linguistics, psychology, or a medical department.

This interdisciplinary venture has opened up new channels between formerly isolated faculties. We find young people from a humanities faculty engaged in mathematical problems of signal processing. Conversely, students in electrical engineering and computer science departments make significant contributions to phonetics and linguistics research. Phonetics, within its new profile, takes part in clinical projects and receives funding for basic work in speech recognition and synthesis. This is, indeed, a symbiosis or rather a fusion of research profiles. It is a healthy development much needed in quest of our far reaching goals - but does it not have any negative effects?

Some problems have been apparent all since computer technology penetrated our field. Many phoneticians of an older generation miss the direct contact with their instrumentation which they could handle without engineering support and which gave them an immediate and intimate insight in speech patterns. The old kymograph was indeed valuable in this respect. Even the sound spectrograph, which once revolutionized acoustic phonetics, is in the risk zone of being outdated by multi-function computer analysis programs. However, up till now they have not demonstrated the same temporal resolution as the rotating drum print-out from the ordinary spectrograph, which I still would not be without in spite of access to computer spectrograms with additional synchronized parameters.

At the same time as our appetite grows for more advanced computers systems with analysis and synthesis coordinated in interactive programs, we run into the usual problems of reliability and difficulties in accurately documenting and memorizing complex routines and, as you know, computers have a tendency to break down or to be occupied when you need them most.

Also, if we do not know how to rewrite and expand existing programs, we may become limited by software constraints which are not initially apparent. One example is the widely spread ILS system which, for the benefit of a graphically optimized positioning of curves, has a tendency to discard information on relative intensities comparing successive section frames.

The problem is that neither the software designer nor the user are always aware of needs that emerge from the special properties of speech signals or the research needs. One example is routines for spectrum analysis of unvoiced sounds, for instance of fricatives. Standard FFT routines without additional temporal or spectral averaging retain a random fine structure of

almost the same amplitude as that of true formants. The result is a fuzzy spectral picture in which it is hard to see what is a formant and what is a random peak. Spectral smoothing can be attained in many ways. Cepstrum analysis or LPC are useful but the smoothness of the LPC curve can be deceptive since the location of the formant peaks may vary somewhat from sample to sample.

We all know that computers are fast in operation but that programming can take a long time. It is also apparent that computer programming is an art which possesses a great inherent fascination which may distract the user from his basic scientific problem. An intense love-hate relation may develop. I have stayed away from programming until recently, when I started using a Japanese programmable calculator which gives me the great satisfaction of access to fairly complex modeling at the price of time demanding debugging.

One can also raise the partially philosophical problem: Who is the boss? The user or the computer? Can we leave it to the computer to learn about speech or shall we insist on developing our own insights in the many domensions of the speech code? This is really a matter of strategical importance in speech research.

3. Speech Recognition and Research Needs

There are basically two different approaches possible in automatic speech recognition. Either we start by running the computer in a learning mode to store a number of templates of speech patterns from a single or a few subjects, recognition then simply becomes a best match selection. We learn very little about speech this way and we are generally not aware of why the matching incidentally fails.

The other approach needed for large vocabularies and connected speech is phonetically orientated in the sense that it is based on recognition of minimal units that can range from distinctive feature phonemes, diphones, syllables, and words and which require some kind of segmentation. We now approach the general problem of speech research in quest of the speech code and the relation between message units and their phonetic realization with all the variability induced by contextual factors including language, dialect, speaker specific situational and stylistic variations.

It would be a wishful dream to extract all this knowledge merely by computerized statistics, i.e. to collect a very large material of speech, give the computer some help for segmenting transcription, and look up and then just wait for the results to drop out.

Many institutions are now developing such data banks for their research. This is a necessary involvement to make but satisfies a partial need only. We cannot store all possible patterns with table look-ups. To organize the data bank efficiently, we must rely on a continuing development of a model of speech production and generative rules on all levels up to the linguistic frame and down to an advanced vocal tract model which should include all what we

know of aerodynamics and source filter interaction. Flanagan in his paper to this Congress describes this process as letting the vocal tract model mimick the speech to be analyzed. This is a dynamic realization of analysis by synthesis, which we will be able to handle once we have gained a sufficient understanding of the speech production process.

It has already been proposed by some people to integrate a text-to-speech synthesis system as a part of the top-down arsenal of speech recognition. As pointed out by John Holmes, this general approach of perturbing synthesis parameters for a best match to a natural utterance is also an effective way of improving synthesis by rule. Here lies perhaps the main advantage of interfacing analysis and synthesis. The basic outcome is that we learn more about speech. Once we have a sufficient insight, we may produce short-cut rules for articulatory interpretations of speech patterns to guide further data collection or for recognition of articulatory events to guide the recognition.

This might be a more realistic approach to attempt a complete match which would require a very advanced adaptability to speaker-specific aspects. Again we are confronted with the constraints of pattern matching procedures.

4. Perception

Now you may ask, why all this emphasis on production? What about models of speech perception and feature theory as a guide for recognition?

First of all, it is apparent that the main drawback of present speech recognition schemes is the handling of bottom-up acoustic data. Either we lose a lot of information-bearing elements contained in rapidly varying temporal events or we perform a maximally detailed sampling in which case substantial information may be lost or diluted by distance calculations, performed without insight in the speech code. Frequency and time-domain adjustments by dynamic programming or by some overall normalization procedure are helpful but do not account for the uneven distribution of information.

Would it not be smarter to base the recognition on models of auditory processing including feature detection? Feature detection is, of course, closely related to the search for articulatory events but with the aid of perception models, we could hope to attain a simpler and more direct specification of the relevant attributes.

Formant frequency tracking is often difficult even for non-nasalized sounds and ambiguities have to be solved with reference to specific spectrum shapes. Models of the peripheral auditory system including Bark scaling, masking, lateral inhibition, and short-time adaptation can provide some improvements in portraying essential characteristics but do not immediately suggest a parametrization. The ultimate constraints are to be found at higher levels of auditory perception but here our insight is more limited and speculative, for instance, in questions of what is a general function and what is a speech mode specific mechanism.

There is now emerging a new duplex view of peripheral auditory analysis on the one hand, the basic concept of short-time spectrum transformed to a spatial discharge rate – on the other hand, the tendency of the outputs from a number of adjacent nerve-endings to be synchronized to a dominant stimulus frequency. The equivalent frequency range over which such synchronization takes place becomes a measure of the relative dominance of a spectral component, and the information about the frequency of the component is, at least for lower frequencies, contained in the neural periodicity patttern. The so-called DOMIN modelling of Carlson and Granström (1982) has its support in the neurophysiological studies of Sachs et al. (1982) and those of Delgutte (1982). A consequence of the DOMIN modelling of Carlson and Granström is that the algorithm, based on a Bark scaled filter bank, detects low-frequency harmonics at high F_0, otherwise formants or formant groups. In the earlier experiments of Carlson and Granström, based on the Bekesy-Flanagan auditory filters which are wider than those of the Bark scale, the system produced something that came close to an F_1 and F_2 detection.

We have already noted that models of the peripheral auditory system do not provide you a complete auditory transform. For a more true representation of the neural transform, we would have to inspect the cortical domain. The psychoacoustic experiments of Ludmilla Chistovich and her colleagues in Leningrad suggest some kind of spatial integration to take place above the level of peripheral hearing. They found that two formants interact to provide a joint contribution to the percept when placed closer than a critical distance of about 3.5 Bark and may then be substituted by a single formant of some weighted mean to provide the same categorical effect. On the other hand, when formants come further apart than the critical distance their relative amplitudes can be varied over a wide range without affecting the identification.

These effects are relevant to the discussion of vowel systems and conform with the early studies of Delattre et al. at Haskins Laboratories who found that back vowels can be simulated by a single formant. I may illustrate the categorical boundary between back vowels and more centrally located vowels by reference to Figure 2 which shows Swedish vowel formants arranged in $F_2' - F_1$ versus $F_2' + F_1$ plot with frequencies transformed to equivalent Bark values. The tendency of fairly equal spacing and regular structure has exceptions which can be related to historical sound changes and a combination of contrast enhancement and reductions. Thus, the Swedish long [u] produced with very high degree of liprounding has advanced articulatorily to a front vowel with a tongue location similar to that of [i:], whilst its short counterpart [ə] resembles a back vowel but for a tongue location sufficiently advanced to transcend the 3.5 Bark F_2-F_1 boundary. Perceptually the long [u] and the short [ə] occupy an extreme low F_1+F_2' 'flatness' feature which they share with their historical origin [u:] and [U] in relation to all other vowels, see further Fant (1973; 1983).

Auditory modelling has now penetrated into the domain of speech dyna-

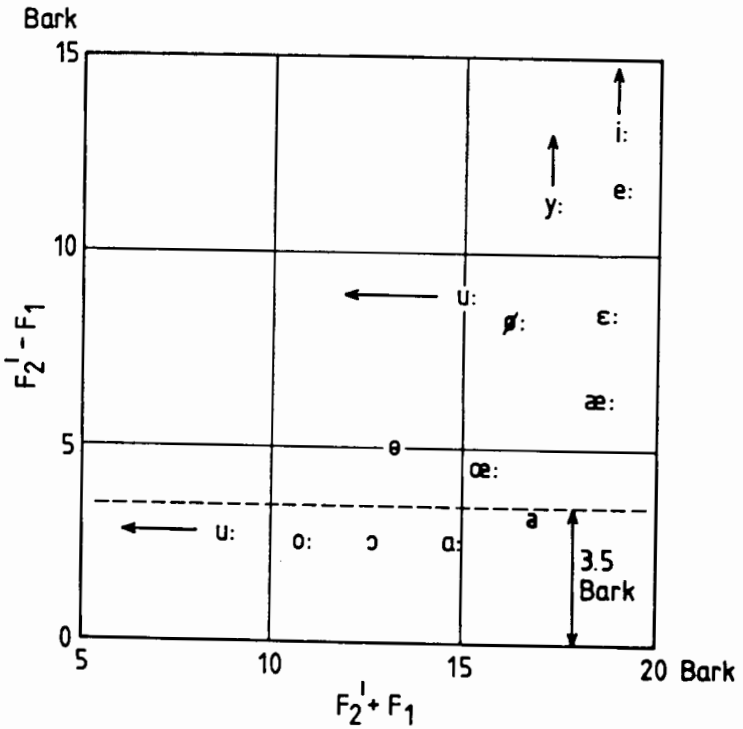


Figure 2. Long and short Swedish vowels in auditory adjusted F_1' and F_2' scales. Arrows indicate diphthongal extensions. Short vowels of F-patterns close to their long mates are omitted.

mics. There are indications that short-time adaptation effects increase the discriminability of rapid onset patterns and that the frequency resolution is enhanced for timevarying formant patterns. There remains much to be learned about these effects and the role of special feature detectors.

Our lack of understanding of the speech perception mechanism may be illustrated by the two spectrograms of one and the same sentence recorded in an auditorium (see Fig. 3) The upper case spectrogram refers to a microphone close to the speaker and in the lower case it originates from a microphone in the middle of the auditorium. The reverbation distortion does not impede intelligibility much but the spectrographic pattern is blurred to the extent that most of the usual visual cues are lost. How does the auditory system combat noise and reverbation?

5. In Quest of the Speech Code. Variability and Invariance

Although there are shortcuts for special purpose speech recognition and synthesis by rule, it is evident that advanced goals can be reached by intensified fundamental research only. The common knowledge needed, the structure of the speech code, is also the central object of phonetics. Models of production and perception constitute a biological frame within which we can

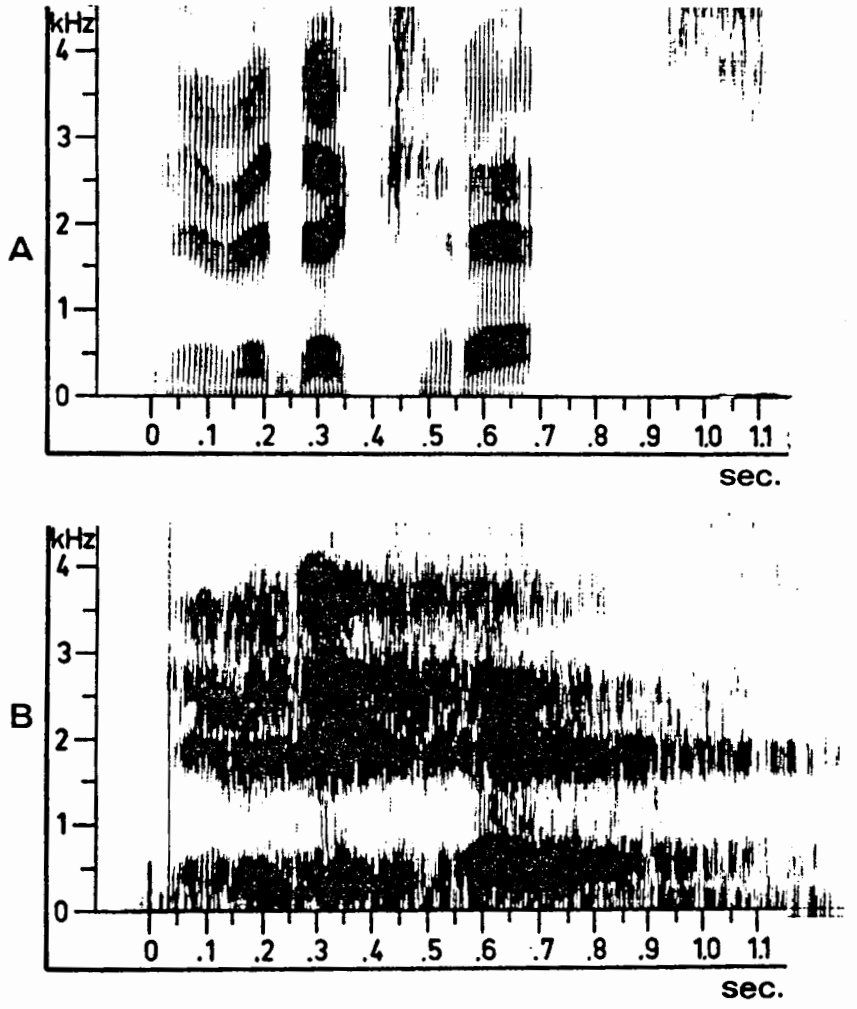


Figure 3. Spectrograms of one and the same utterance from a close talking microphone and in the middle of a reverberant auditorium.

study the speech code. Now even if we possessed perfect general models of production and perception and a maximally effective linguistic framework, we would still have to derive an immense amount of rules and reference data relating message units and speaker categories to observed phonetic sound shapes. Presently available reference data and rules are incomplete and scattered into fractional acoustical phonetic studies. The more complete rule systems are hidden in the software of text-to-speech synthesis systems and are contaminated by elements of ad hoc guess work and by the specific format of the parameter system.

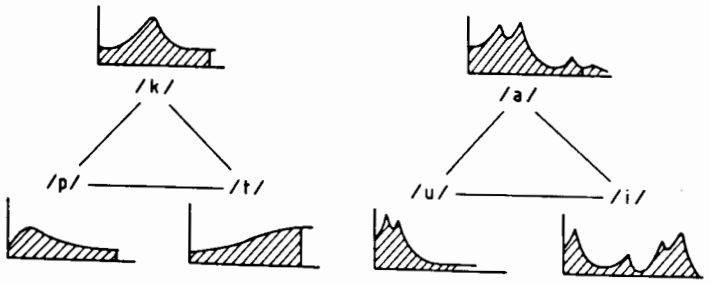
So far, speech technology has relied heavily on linguistic redundancies to ensure an acceptable performance of synthesis as well as recognition, but it is

due time to extend fundamental knowledge by large documentary projects around our data banks. When will we have a new version of the book *Visible Speech* with not only illustrations but with reference data and major contextual rules, in other words, the missing links towards the output of generative grammar? When will we have a complete inventory of rules for generating different voice types and stylistic variations?

For applied work it is of no great concern which distinctive feature system we adopt for addressing phonemes as long as we can properly handle their acoustic-phonetic correlates. Prosodic categories should not be defined by single physical parameters. They should be treated the same way as phonological segmentals, that is, as constituents of the message level with rules for their many phonetic realizations.

The study of coarticulation and reduction is of central importance. There is a need to extend the concept of reduction to variations induced by various degrees of stress emphasis and stylistic factors. A typical example is the variation of vowel formant frequencies with the mode of production. We find a more extreme articulation in citation forms than in connected speech and even more extreme in targets in sustained vowels. Emphasis and de-emphasis affect not only target values but in general all speech parameters and their temporal patterning.

The speech code is a theme about variabilities and invariance. Invariance and manifestation rules are closely connected. How do we define invariance?



Jakobson, Fant, Halle

	Compact	Grave
/k/	+	
/p/	−	+
/t/	−	−
/a/	+	+
/u/	−	+
/i/	−	−

Chomsky and Halle

	Anterior	Coronal	Back	Low	High
/k/	−	−	+	+	−
/p/	+	−	+	−	+
/t/	+	+	−	−	+

Figure 4. Spectral attributes and DF-specifications of /p,t,k/ and /u,i,a/ according to Jakobson, Fant and Halle; Chomsky and Halle

I feel that we should make a distinction between academical and more pragmatic needs. Roman Jakobson's concept of distinctive feature implies in its most general form a relational invariance. Independent of the sequential context and specific combination of other features in a phoneme, there remains 'ceteris paribus', a vectorial difference along the feature dimension comparing the + alternative and the − alternative.

Obviously, we do a better job in recognition if we make use of all conditional factors affecting the sound shapes of the two candidates. However, a research line adopted by Kenneth Stevens is directed towards, what he calls 'absolute' invariance, which conceptually comes close to the common denominator aspect of the distinctive feature theory. Stevens started out by studying spectrum slope properties of the stop burst and extended his descriptions to temporal contrasts, e.g. the intensity of the burst and that of a following vowel in a certain frequency region. I have suggested an extension of the concept of absolute invariance to employ any description which does not imply a prior phonological identification of the context. In this sense, positional allophones of /k/ and /g/ may be identified by both the degree of spectral concentration and by the location of energy with respect to the format pattern after the release.

Returning to academical issues we find that the use of one and the same feature, such as compactness in both consonant and vowel systems, complicates and dilutes the common denominator whilst there still remains an interesting parallelism, in the Jakobson-Fant-Halle system brought out by the identification of the [k] [p] [t] relations with those within [a] [u] [i]. The Chomsky-Halle system operating with independent consonant place features has its shortcomings in the roundabout labeling of labials as [+anterior [-coronal]. I prefer the output oriented acoustic-perceptual basis. Major spectral attributes are preserved in neurophysiological studies as those of Sachs et al. (1982: 121) see their figure of [i] versus [a] emphasizing the compactness feature.

I am now approaching the more philosophical aspects of phonetics. We are all more or less engaged in studies of the speech code but this is a painstaking slow process. Meanwhile we can make general remarks about the code, e.g. that it has developed with a major concern for the final stage of the speech chain. Roman Jakobson's theme 'we speak to be heard in order to be understood' has had a great impact. This principle is referred to by Bjorn Lindblom as *teleological*. With a slight deletion in this exclusive term, we end up with the word *teology* which has some bearing on issues such as motor theory of speech perception, 'speech is specially handled in perception', the speech code is innate, speech production is a chain process or is preplanned etc.

I am personally in favor of a both-and principle. No single statement is sufficient. Speech is both precise and sloppy. Speech perception involves many parallel processings and may rely on both phonemes, syllables, and words as minimal recognition units. The statement that the truth about

segmentation is that you cannot need modification. You both can and cannot. The common denominator of distinctive feature is sometimes easier to describe with reference to articulation than to perception and the reverse is often true. Motor theory of speech perception as well as auditory theory of speech production both have something to contribute to our perspective.

The most absolute statement I can make is that speech research is a remarkable, exciting venture. Most people take speech for granted. A small child can do what 700 wise men and women at this congress do not quite understand. I wish you all an exciting continuation of the congress.

References

Carlson, R. and Granström, B. (1982). Towards an auditory spectrograph. In: *The Representation of Speech in the Peripheral Auditory System*, 109-114. Amsterdam: Elsevier Biomedical Press.

Delgutte, B. (1982). Some correlates of phonetic distinctions at the level of the auditory nerve. In; *The Representation of Speech in the Peripheral Auditory System*, 131-149. Amsterdam: Elsevier Biomedical Press.

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

Fant, G. (1983). Feature analysis of Swedish vowels – a revisit. *STL-QPSR* **2-3**/1983.

Sachs M.B., Young, E.D., and Miller, M.I. (1982). Encoding of speech features in the auditory nerve. In: *The Representation of Speech in the Peripheral Auditory System*, 115-30. Amsterdam: Elsevier Biomedical Press.