Dept. for Speech, Music and Hearing
Quarterly Progress and Status Report

Auditory models and isolated word recognition

Blomberg, M. and Carlson, R. and Elenius, K. O. E. and Granström, B.

journal: STL-QPSR
volume: 24
number: 4
year: 1983
pages: 001-015

KTH Computer Science and Communication

http://www.speech.kth.se/qpsr

I. SPEECH RECOGNITION

AUDITORY MODELS AND ISOLATED WORD RECOGNITION*

Mats Blomberg, Rolf Carlson, Kjell Elenius, and Björn Granström

Abstract

A straightforward isolated word recognition system has been used to test different auditory models in acoustic front end processing. The models include BARK, PHON, and SONE. The PHONTEMP model is based on PHON but also includes temporal forward masking. We also introduce a model, DOMIN, which is intended to measure the dominating frequency at each point along the 'basilar membrane.' All the above models were derived from an FFT-analysis, and the FFT processing is also used as a reference model. One male and one female speaker were used to test the recognition performance of the different models on a difficult vocabulary consisting of 18 Swedish consonants and 9 Swedish vowels. The results indicate that the performance of the models decreases as they become more complex. The overall recognition accuracy of FFT is 97% while it is 87% for SONE. However, the DOMIN model which is sensitive to dominant frequencies (formants) performs very well for vowels. Three different metrics for measuring the distance between speech frames have been tested: city-block, Euclidean, and squared (Euclidean without taking the square root). The Euclidean seems to give a slightly better performance. Reducing the number of channels in the FFT processing clearly shows that performance increases with the number of channels.

Introduction

The use of auditory models as speech recognition front ends has recently attracted a great deal of interest. The underlying assumption is that a good model of the auditory system should generate a more natural and efficient representation of speech compared to ordinary spectum analysis. However, we have to keep in mind that only some of the peripheral processes of sound perception are included in most existing models. In this paper we will discuss some standard spectral transformations of the acoustic information and also a model based on the dominant frequency concept. We will also evaluate the performance of the different representations in the context of a standard speech recognition system.

Auditory models

Basic research has resulted in several models of the peripheral auditory processing. The elaboration of Zwicker and Feldtkeller (1967) of the loudness and the Bark concept has become more or less standard in psychacoustics. Other models, including lateral inhibition and time dependent mechanisms, have been created elsewhere. The development of methods for similarity rating of speech spectra has been of interest in

---

* *This is an expanded version of a paper presented at the Symposium on 'Invariance and Variability of Speech Processes', MIT, Oct. 8-10, 1983.*

many research groups, Plomp (1970), Pols (1970), Bladon and Lindblom (1977), and Carlson and Granström (1979). At the same time, efforts have been made to include knowledge of the auditory system in practical applications, Schroeder et al. (1979) and Lyon (1982). Klatt (1979, 1982a, 1982b) has discussed physiologically related spectral representations in models for lexical access and speech recognition systems. New models of the peripheral auditory system are developed based on neurophysiological results, Chistovich et al (1979, 1982), Sachs and Young (1980), Sachs et al. (1982), Delgutte (1980, 1982), Dolmazon (1982), Goldhor (1983), and Seneff (1983). This positive development make us believe that in the future we could use these kinds of models as the first analyzing steps in a speech recognition system.

In the present paper we try to elaborate some of the basic facts of the auditory mechanisms in the context of such a system. In Fig. 1 we present the different models/transformations that we have used in the current experiment. A pure sinusoid and a vowel are used as test stimuli to illustrate some alternative representations in the amplitude/frequency domain. Fig. 2 gives examples of computer-generated spectrograms based on some of these models.

The speech signal is first filtered through a sampling filter of 6.3 kHz and digitized at 16 kHz. An FFT-spectrum is calculated every 20 ms using a 25 ms Hamming window. The line spectrum is then transformed in the frequency domain by adding energies to get 300 Hz wide channels. This will reduce the influence of the fundamental frequency on the spectrum and is done in 74 overlapping channels from 0 to 7.7 kHz using a linear frequency scale. The result of this processing is seen in Figs. 1a and 2a (FFT).

If we use a Bark scale and a bandwidth of one Bark, we will have a psychoacoustically more relevant representation (BARK, Fig. 1b). The Hamming window is set to 10 ms with a sampling frequency of 16 kHz to facilitate a fast response for frequencies higher than 2 kHz, and to 20 ms with a sampling frequency of 4 kHz for frequencies lower than two kHz. The reduction to lower sampling frequency gives a better frequency resolution for the following transformation into the Bark scale. This transformation is used in this and all the following models, while the summation into one Bark bands is only done in this model.

A psychoacoustic masking filter (Schroeder et al., 1979), rather than a sharp bandpass filter, together with equal loudness curves (phone curves), has been used to derive a phon/Bark plot (PHON, Figs. 1c, 2b, and 4a). We argue that the visual impression of Fig. 2b has a much closer relation to the perceived sound than the FFT representation. Note the reduced emphasis on the fricative and the position of the very important second formant in the middle of the spectrogram. The perceptually prominent lowest formant is also visually enhanced.

The phon/Bark representation has been transformed to a sone/Bark representation which often is claimed to give a better description of the percieved loudness (SONE, Fig. 1d).

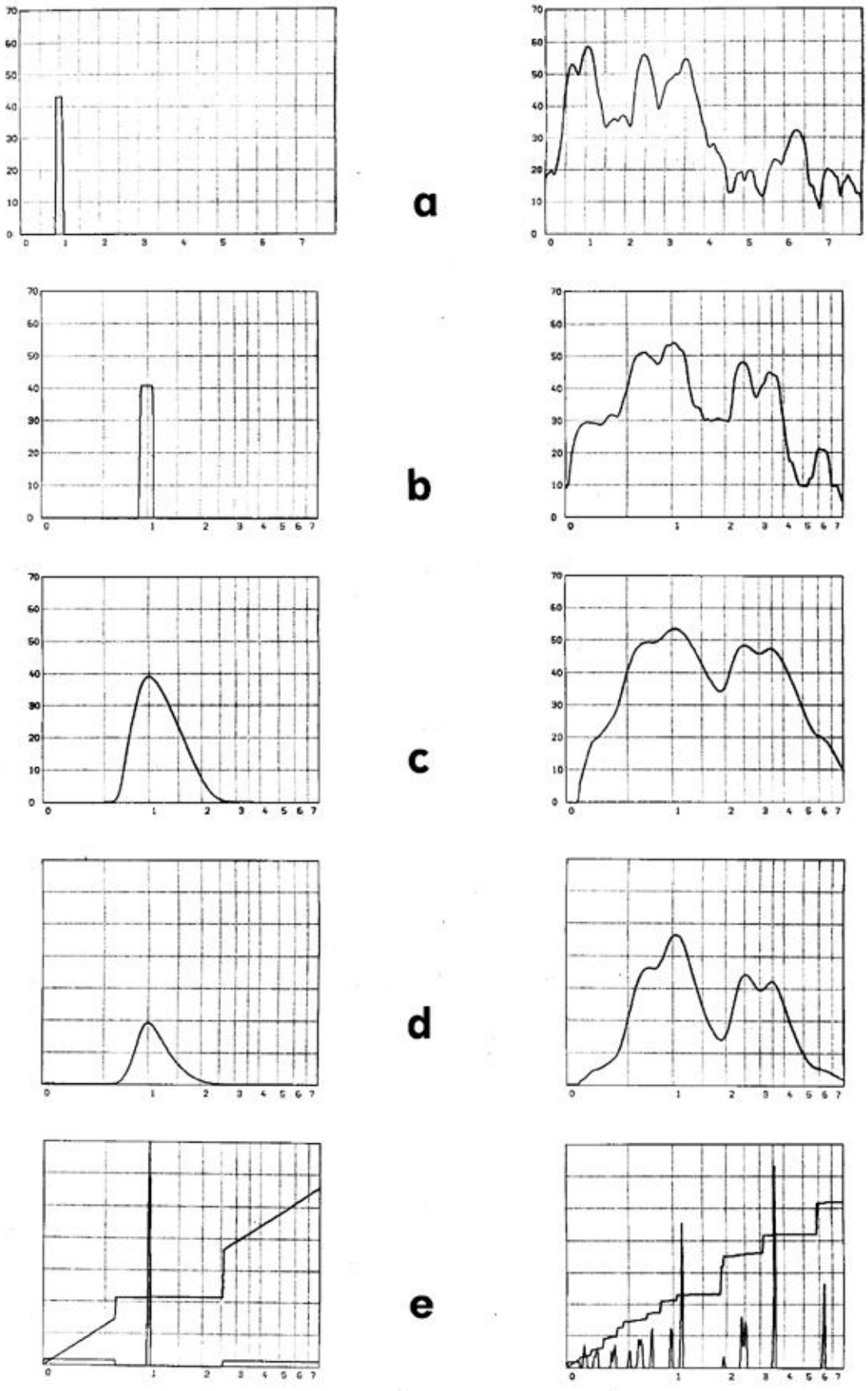

Fig. 1. A sinusoid of 1 kHz in the different representations explained in the text.

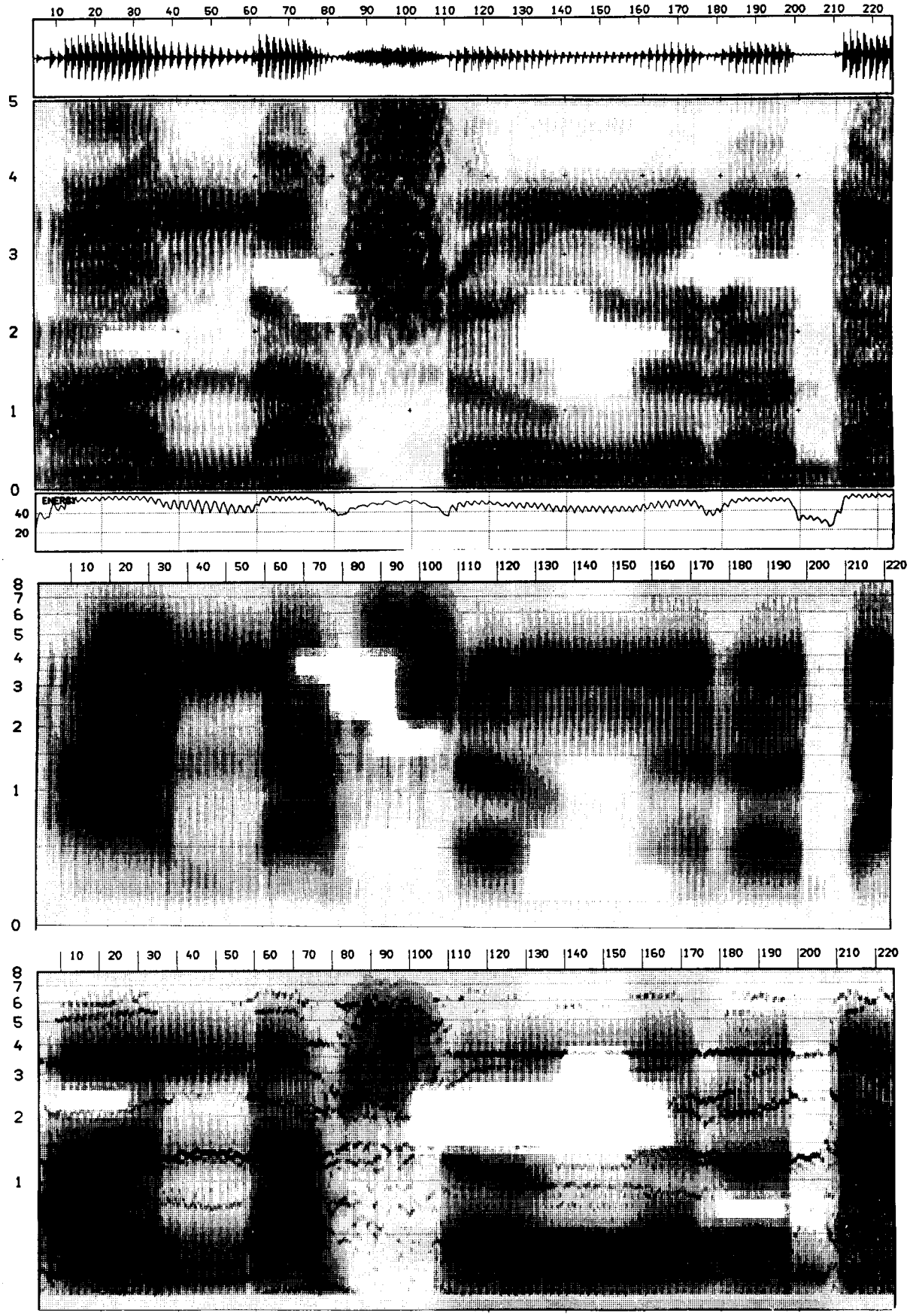

Fig. 2. Three alternative spectrograms of the sentence /ala ʃUŋər ɔda.../. See text for details.

In Fig. 1e, an alternative model (DOMIN) is introduced. It is based on our earlier work on vowel perception (Carlson and Granström, 1975, 1979) and explores the possibility of temporal analysis in the auditory system (Sachs et al., 1982). The model uses the masking filter introduced in Figure 1c to find which frequency dominates each point along the "basilar membrane." The dominant frequency is plotted along the y-axis while the x-axis still corresponds to the Bark scale. It may be seen that the sinusoid generates a step in the curve. If the stimulus consists of a number of resonances or cut-off frequencies, they will all generate steps in the curve. The width of the step will be dependent on the amount of masking or dominance.

The superimposed narrow peak in Fig. 1e presents the same information in histogram form, i.e., the y-axis is now the interval along the Bark scale that is dominated by a certain frequency. Intuitively, this could be regarded as the number of neurons that respond to the same dominant frequency.

Fig. 2c incorporates the histogram representation of Fig. 1e with the phon/Bark analysis. The formants are emphasized and the resonances during the occlusion of /b/ may be observed. Since the frequency-dependent analysis makes the bandwidth narrow at low frequencies, the first harmonics are well marked while they disappear at higher frequencies in favor of formants. The intonation and the formant pattern can be studied in the same representation.

The auditory models described so far in this paper are static even if time is used as a parameter. No temporal masking effects have been taken into account. This kind of effects are obvious candidates for future developments of speech recognition front ends. Zagoruiko and Lebedev (1975) have indicated positive results by including such effects in a speech recognition system. Hence, the last model, PHONTEMP, is introduced. It is based on the PHON model and includes also a forward masking function with a time constant of 40 ms. At the onset of a signal in a PHON analysis channel we get an extra excitation of up to 15 phon depending on the size of the step. The effect is illustrated in Fig. 3. The sinusoid in Fig. 3a is processed either by the BARK model, Fig. 3b or by the PHONTEMP model, Fig. 3c. It could be seen how the masking filter broadens the response, especially in the beginning where the onset effects take place. After the onset the response is adapted to a constant excitation. When the signal ends, a forward masking effect takes place. This is simulated by a simple low pass filter with a time constant of 40 ms. Fig. 3d gives an alternative picture where only unmasked frequencies are plotted. The white area in the time and frequency domain shows what is masked by the stimulus. The grey area is uneffected and has a base level corresponding to spontaneous activity. Fig. 4 gives an example of natural speech processed by the different models: PHON, Fig. 4a, PHONTEMP, Fig. 4b and in Fig. 4c with a processing corresponding to Fig. 3d. The same sentence as in Fig. 2 has been used. It could be seen how some of the transitions are emphasized and

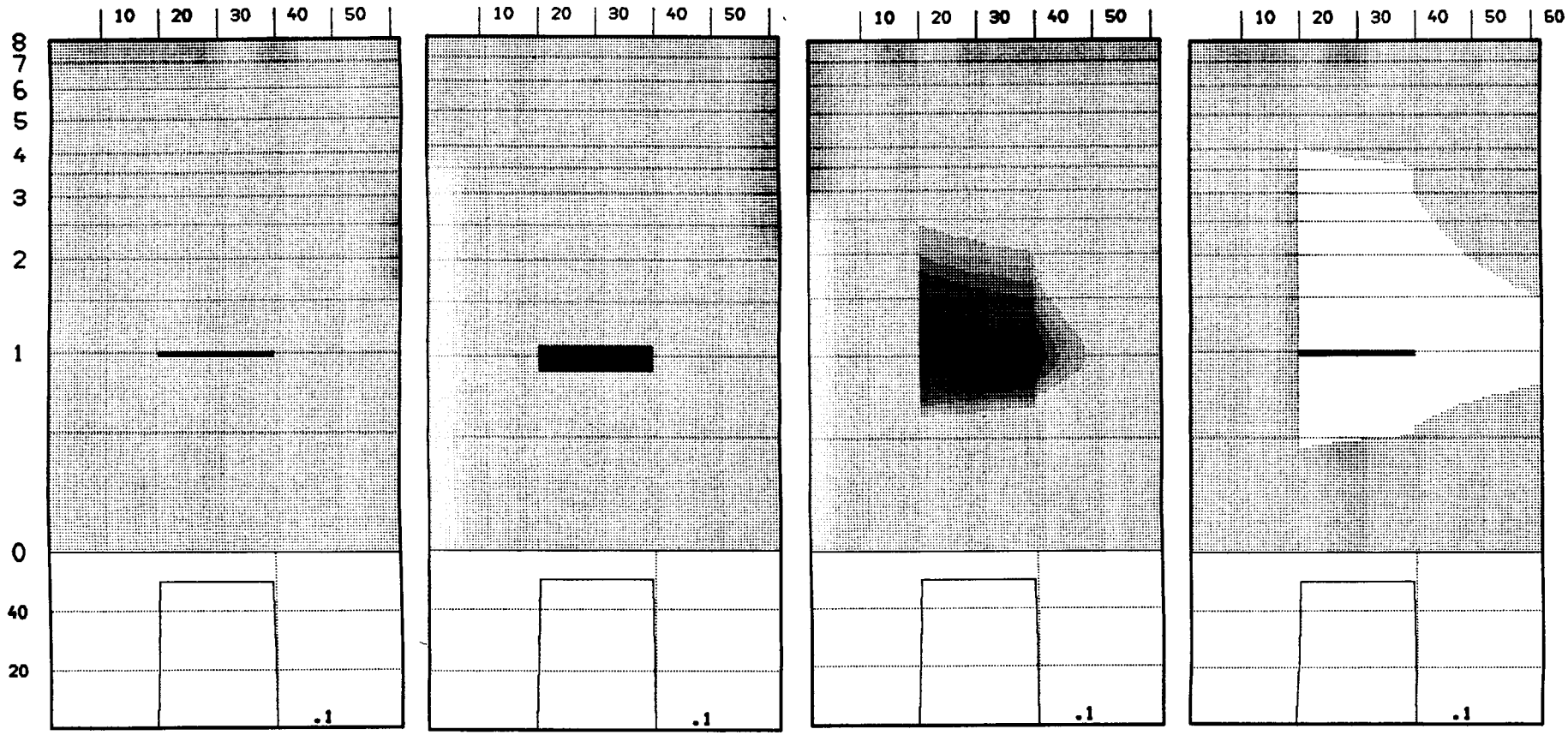

Fig. 3. Four alternative spectrograms of a sinusoid at 1 kHz. See text for details.

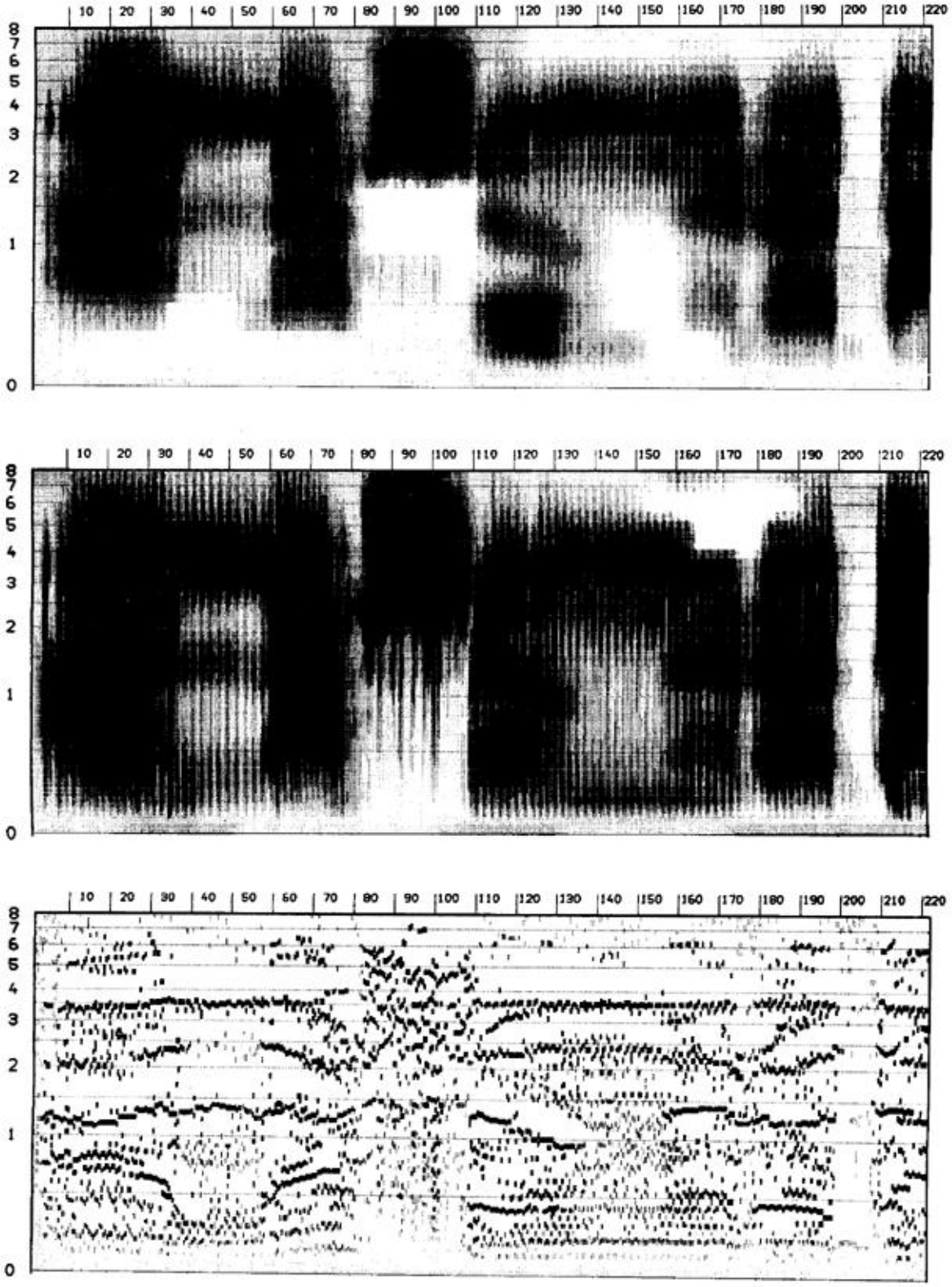

Fig. 4. Three alternative spectrograms of the sentence / ala ɧUŋər ɔ da.../. See text for details.

how the formants dominate the surrounding. The forward masking effect could be seen in the stop gap at the end of the spectrogram.

The recognition system

The recognition system is based on ordinary isolated word recognition techniques using pattern matching and dynamic programing (Elenius and Blomberg, 1982). The Euclidean metric is normally used for calculation of distances between word patterns. The intention with this recognition system is not to get the best possible performance; what we intend is a flexible, standard framework to compare different recognition preprocessors and recognition strategies.

All the auditory models are derived from FFT processing. In each representation, 74 parameters are used, which are evenly sampled along the different frequency scales. When a word is detected, it is linearly normalized to 40 sample points in time. This 'normalize and warp' technique was used because it reduces the processing needed for dynamic programing (Myers et al., 1980). A word is, hence, described by 40 parameter vectors, each having 74 elements. The time normalization violates what we know about speech perception but the effects on the recognition performance are small as may be seen below under "results". A more correct time normalization should probably reduce the influence of stable segments while emphasizing the transitional segments. It would be tempting to apply a perceptual time dimension rather than a normalized physical dimension. This has not been explored in this experiment but has been dealt with earlier, Elenius and Blomberg (1982). The method used by Kuhn et al. (1981), where the sampling points are chosen so that they all contain the same amount of spectral change, may also be seen as a step towards a perceptual time representation.

In the experiment, each reference word is built from seven utterances during a learning phase. The first learning sample of each word is used for a dynamic time alignment of the second one and a mean is calculated. The third utterance is warped to this mean and then added to the reference and so on. Each word added to the reference is weighted so that all learning words have equal influence on the resulting template. The reason for using several words for each template is that using only one word gives a rather poor and varying recognition result as is shown later under "results".

Vocabulary

Two vocabularies were tested. One consisted of the nine long Swedish vowels in an hVl-context and the other of 18 Swedish consonants in an aCa-context. What really was tested, since the context is constant, was the consonant and vowel discrimination ability of the system, though there was no segmentation into phonemes of the words used. The recordings were made in a relatively quiet office room and reduced the problem in detecting the end points of the utterances.

Each vocabulary was read 31 times by one male and one female speak-

er. Seven repetitions of each word were read in a normal way and were used to build the reference templates. The test utterances were read in four different ways: normally, slowly, rapidly, and emphatically. Six utterances for each speaking style made a total of 24x18=432 test words for the consonants and 24x9=216 test words for the vowels. The reason for varying the manner of speaking was to increase the error rate in order to make the differences between the models more pronounced.

Experiment

The recognition performance of the different auditory models has been tested and compared to the FFT processing using the vocabulary presented above. We have also explored some specific characteristics of the recognition system; the method of building each reference template from several words and the effects of the linear time normalization. We have also tested three different metrics for calculating the difference between speech frames; city-block, Euclidean, and squared metric. Finally we have reduced the number of parameters of the FFT processing by increasing the bandwidth of each channel. This is analogous to using a filter bank with a decreasing number of channels.

Results

Before presenting the recognition performance of the different auditory models we give the results concerning the details of the recognition system.

Different number of words per reference.

In this section we present results of varying the words to build each reference template. Using the seven learning words as references directly, one at a time, results in recognition rates between 78% and 85% for the FFT processing of the consonants of the male speaker. Thus, the variation is quite large as argued earlier. The mean is 81%. Increasing the number of words per template from one to seven gives accuracies according to Fig. 5.

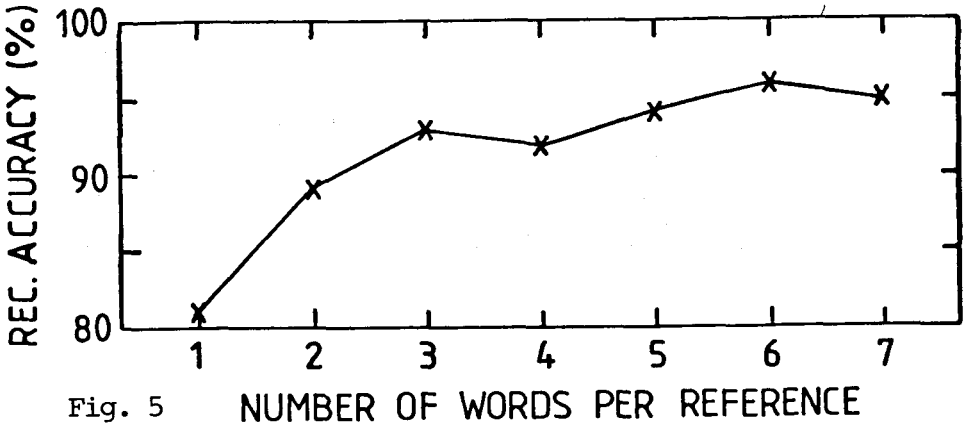

Fig. 5

It may be seen that increasing the number of words per reference gives a substantial increase in performance, especilally when going from one to three words. The results were achieved by permuting only seven

different learning lists but the effect is quite pronounced and also in agreement with our experience.

Time normalization and speaking style.

We will present some statistics on variations due to the different kinds of speaking style. The 'slow' words were about 25% longer than the normal ones, the 'rapid' words were about 25% shorter, and the 'emphatic' words were about 10% longer. Changing the manner of speaking may also influence the speech spectrum. The error distribution according to speaking style over all models (except PHONTEMP) of the male consonants were; normal 21% of the errors, slow 20%, rapid 30%, and emphatic 29%. Thus, the rapid and emphatic readings caused about 50% more errors than the normal and slow readings. The fact that the slow readings perform as well as the normal ones indicates that the effect of the linear time normalization on the recognition results is small and probably negligible.

Auditory models.

The recognition results of the different auditory models are displayed in Table I. The recognition performance of each model is shown separately for consonants and vowels for each speaker. The mean of each model is also given.

MODEL	VOWELS		CONSONANTS		MEAN
	MALE	FEMALE	MALE	FEMALE	
FFT	**99**	96	95	97	97
BARK	**99**	95	92	95	95
PHON	**96**	91	88	91	92
PHONTEMP	95	85	82	88	88
SONE	**91**	93	82	83	87
DOMIN	99	99	90	90	94

Table I. Recognition accuracy of FFT processing and different auditory models in per cent.

The FFT based recognition is surprisingly accurate especially for consonants, 97.5% for vowels and 96% for consonants. The number of parameters (74), however, is greater than what is normally afforded in speech recognition systems. Furthermore, the intervocalic position of the consonants by-passes the end point detection problem.

When the auditory-inspired transformations of the speech spectra grow more complicated (BARK-PHON-SONE-PHONTEMP), the recognition results deteriorate progressively for both consonants and vowels. The DOMIN, which is a model of a rather different kind, does not follow this general tendency. The performance for vowels is excellent, but for

consonants it is no better than the PHON representation. This is understandable in the light of the function of DOMIN. In this model, all emphasis is put on the frequency location of prominent regions (formants) which are known to form a good basis for vowel identity decisions. This point is strongly supported by our earlier experiments on vowel perception and model predictions of perceptual distance, Carlson and Granström (1979). In the DOMIN model, all loudness information is disregarded, which makes the discrimination between consonants problematic, especially if they have the same place of articulation. For the SONE model, the difference between performance on vowels and consonants is considerable. The low score for consonants is possibly due to the amplitude transformation involved. The relatively weak consonant segments tend to be disregarded in the distance calculation. In a pilot experiment, not reported here, an ad hoc combination of the PHON and DOMIN models was used with quite promising results.

Distance calculations.

Three different kinds of metrics have been explored; city-block, Euclidean, and squared distances. City-block means adding the absolute values of the difference between the parameters of two compared speech frames. Squared means adding the square of the differences. By taking the square root of the squared sum, we get the Eucliden metric. The squared metric punishes large differences much more than small ones. The metrics have been tested for the consonants of the male speaker for all processings except PHONTEMP. The experiment is done with only every third of the calculated parameters for each model since the spacing between the parameters is one third of their bandwidth. For the DOMIN model we have taken the arithmetic average over three parameters instead of just sampling every third. This will smooth the effects of steps in the DOMIN representation which otherwise would cause problems.

MODEL	VOWELS			CONSONANTS		
	CB	EU	SQ	CB	EU	SQ
FFT	99	99	99	92	95	96
BARK	97	99	98	89	92	91
PHON	95	96	95	88	88	85
SONE	87	92	87	80	81	78
DOMIN	99	99	97	93	92	94

Table II. Recognition accuracy for different distance metrics; city-block (CB), Euclidean (EU), and squared (SQ). Male speaker, consonants, 25 parameters.

Comparing the Euclidean results with those of Table I shows that reducing the number of parameters from 74 to 25 does not impede the recognition performance. This is not very surprising taking into ac-

count the redundancy due to the overlap between channels in the 74 parameter case. The differences between the different metrics are not very pronounced. However, the Euclidean metric seems to perform somewhat better. It is the best, or among the best, in eight cases of ten while the corresponding count is three for both other methods.

Reducing the number of parameters.

If we reduce the number of channels in the FFT filter bank by increasing the bandwidth of the filters, the overall recognition rate decreases, as may be seen in Fig. 6. he new parameters are calculated as the energy sum over 3, 6, 12, and 25 'original' parameters resulting in a total of 25, 13, 7, and 3 new parameters.

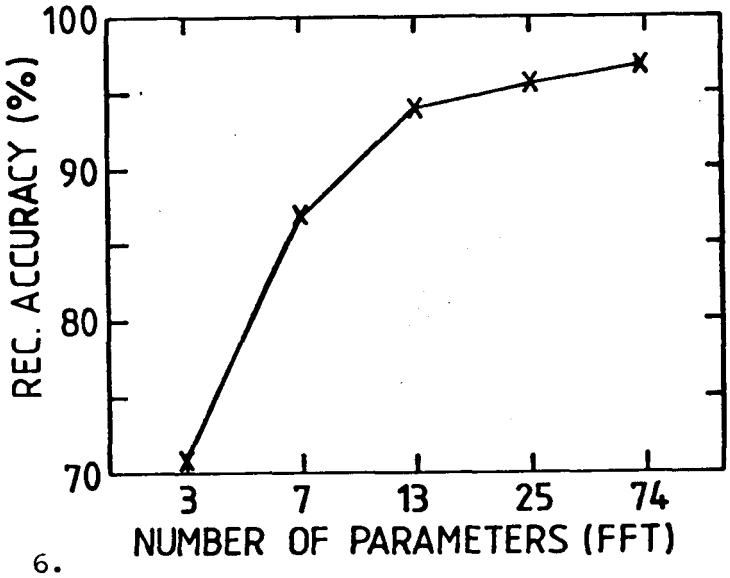

Fig. 6.

These results do not conform with those of Dautrich et al. (1983) who reported decreasing accuracy for increasing number of filters. However, they used non-overlapping filters, and as the number of channels increased, the bandwidth decreased, making the filter outputs sensitive to variations in fundamental frequency, especially for high pitched female voices.

Conclusion

We have found that the difference between the different kinds of metrics used is very small though the Euclidean seems to be slightly better. The results achieved by varying the number of channels in the FFT model clearly show that increasing the number of parameters also increases the recognition performance.

It is obvious from our experiment concerning the auditory models that the unqualified assumption does not hold - auditory models used as speech recognition front ends will not consistently improve performance. Several plausible explanations for our results could be mentioned.

- All models are based on the FFT analysis. The data will be smeared

in frequency and time depending on the chosen approach.

- The models describe only a few selected ways in which the human auditory system processes data. They may be based on too specific experiments, capturing ways of processing the signal that is not very important for speech processing and missing those that are.

- There is no match between the human-modeled primary analysis and the rest of the recognition system. If this match is required, the modeling could pay off only if the decision making part of the program models the way the central nervous system looks at the sensory data.

All these explanations may contain some truth. This should, of course, not hold us back from the interesting fields of auditory modeling and speech perception. It might, however, be premature to include our fragmentary knowledge of the auditory system in today's speech recognizers.

References

Bladon, R.A.W. and Lindblom, B. (1981): "Modelling the Judgement of Vowel Quality Differences", J.Acoust.Soc.Am. 69, pp. 1414-1422.

Blomberg, M., Carlson, R., Elenius, K., and Granström, B. (1982): "Experiments with Auditory Models in Speech Recognition", pp. 109-115 in The Representation of Speech in the Peripheral Auditory System (eds. R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam.

Carlson, R. Fant, G., and Granström, B. (1975): "Two Formant Models, Pitch and Vowel Perception", pp. 55-82 in Auditory Analysis and Perception of Speech, (eds. G.Fant & M.A.A. Tatham), Academic Press, London.

Carlson, R. and Granström, B. (1979): "Model Predictions of Vowel Dissimilarity", STL-QPSR 3-4/1979, pp. 84-104.

Carlson, R. and Granström, B. (1982): "Towards an Auditory Spectograph", pp. 109-115 in The Representation of Speech in the Peripheral Auditory System, (eds. R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam,

Chistovich, L.A., Sheikin, R.L., and Lublinskaya, V.V. (1979): "Centres of Gravity and Spectral Peaks as the Determinants of Vowel Quality", pp. 143-158 in Frontiers of Speech Communication Research, (eds. B.Lindblom & S.Öhman), Academic Press, London.

Chistovich, L.A., Lublinskaya, V.V., Malinnikova, T.G., Ogorodnikova, E.A., Stoljarova, E.I., and Zhukov, S.JA. (1982): "Temporal Processing of Peripheral Auditory Patterns", pp. 165-181 in The Representation of Speech in the Peripheral Auditory System, (eds. R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam.

Dautrich, A., Rabiner, L.R., and Martin, T.B. (1983): "On the Effect of Varying Filter Bank Parameters on Isolated Word Recognition", IEEE Trans. Acoust., Speech and Signal Proc., ASSP-31, pp. 793-806.

Delgutte, B. (1980): "Representation of Speech-like Sounds in the Discharge Patterns of Auditory-nerve Fibers", J.Acoust.Soc.Am. 68, pp. 843-857.

Delgutte, B. (1982): "Some Correlates of Phonetic Distinctions at the Level of the Auditory Nerve", pp. 131-151 in The Representation of Speech in the Peripheral Auditory System, (eds. R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam.

Dolmazon, J-M. (1982): "Representation of Speech-like Sounds in the Peripheral Auditory System in the Light of a Model", pp. 151-165 in The Representation of Speech in the Peripheral Auditory System, (eds.R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam.

Elenius, K. and Blomberg, M. (1982): "Effects of Emphasizing Transitional or Stationary Parts of the Speech Signal in a Discrete Utterance Recognition System", Proc. IEEE ICASSP-82, Paris, pp. 535-538.

Goldhor, R. (1983): "A Speech Signal Processing System Based on a Peripheral Auditory Model", Proc. IEEE ICASSP-83, Boston, pp. 1368-1371.

Klatt, D.K. (1979): "Speech Perception: a Model of Acoustic-phonetic Analysis and Lexical Access", J. Phonetics 7, pp. 279-312.

Klatt, D.K. (1982a): "Speech Processing Strategies Based on Auditory Models", pp. 181-197 in The Representation of Speech in the Peripheral Auditory System, (eds- R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam.

Klatt, D.K., Seneff, S., and Zue, V.W. (1982b): "Design Considerations for Optimizining the Intelligibility of a DFT-based, Pitch-excited, Critical-band-spectrum Speech Analysis-resynthesis System", Speech communication group working papers, MIT research lab of electronics, Cambridge, MA, USA.

Kuhn, M.H., Tomaschewski, H., and Ney, H. (1981): "Fast Nonlinear Time Alignment for Isolated Word Recognition", Proc. IEEE ICASSP-81, Atlanta, Georgia, pp. 736-740.

Lyon, R.F. (1982): "A Computational Model of Filtering, Detection, and Compression in the Cochlea", Proc IEEE ICASSP-82, Paris, pp. 1282-1285.

Myers, C.S., Rabiner, L.R., and Rosenberg, A.E. (1980): "Performance Tradeoffs in Dynamic Time Warping Algoritms for Isolated Word Recognition", IEEE Trans. Acoust., Speech and Signal Proc. ASSP-28, pp. 623-635.

Plomp, R. (1970): "Timbre as a Multidimensional Attribute of Comlex Tones", in Frequency Analysis and Periodicity Detection in Hearing, Sijthoff, Leiden.

Pols, L.C.W. (1970): "Perceptual Space of Vowel-like Sounds and its Correlation with Frequency Spectrum", in Frequency Analysis and Periodicity Detection in Hearing, Sijthoff, Leiden.

Sachs, M.B. and Young, E.D. (1980): "Effects of Nonlinearities on Speech Encoding in the Auditory Nerve", J.Acoust.Soc.Am. 68 (3), pp. 858-875.

Sachs, M.B., Young E.D., and Miller M.I. (1982): "Encoding of Speech Features in the Auditory Nerve", pp. 115-131 in The Representation of Speech in the Peripheral Auditory System, (eds. R.Carlson & B.Granström), Elsevier/North-Holland Biomedical Press, Amsterdam.

Schroeder, M.R., Atal, B.S., and Hall, J.L. (1979): "Objective Measure of Certain Speech Signal Degradation Based on Masking, pp. 217-229 in Frontiers of Speech Communication Research, (eds. B.Lindblom & S.Öhman), Academic Press, London.

Seneff, S. (1983): Thesis work to be published. Personal communication.

Zagoruiko, N.G. and Lebedev, V.G. (1975): "Models for Speech Signal Analysis Taking into Account the Effect of Masking", Acustica 31, pp. 346-348.

Zwicker, E. and Feldtkeller, R. (1967): Das Ohr als Nachrichtenempfänger, S. Hirtzel Verlag, Stuttgart.