On the Use of Auditory Models in Speech Technology

Piero COSI

Centro di Studio per le Ricerche di Fonetica, C.N.R.
P.zza G. Salvemini, 13 - 35131 PADOVA, ITALY

Abstract. A joint Synchrony/Mean-Rate model of Auditory Speech Processing (ASP) is described, and its application in speech technology is considered. As for automatic segmentation and recognition, few examples are illustrated in which the superiority of the ASP scheme over other methods is enphasized, especially considering speech in adverse conditions.

1. Introduction.

Acoustic analysis front-end of almost all presently commercialized Automatic Speech Recognition (ASR) systems is built using speech "prodution-based" processing schemes. In other words, Short-Time Fourier Transform (STFT), Cepstrum, and other related Speech Processing (SP) [1] schemes were all developed strictly considering physical phenomena that characterise the speech waveform obtained by the electrical transduction of the sound pressure wave. Moreover LPC tecnique [2] and all its variants were developed directly by modelling the human speech production mechanism. In the last years, almost all these analysis schemes have been modified by incorporating, at least at a very general stage, various perceptual-related phenomena. Linear prediction on a warped frequency scale [3], STFT-derived auditory models [4], perceptually based linear predictive analysis of speech [5],[6] are few simple examples of how human auditory perceptual behaviour is now taken into account while designing new signal representation algorithms. Furthermore, the most significant example of attempting to improve acoustic front-end with perceptual related knowledges, is given by the Mel-frequency cepstrum analysis of speech [7], which transforms the linear frequency domain into a logarithmic one resembling that of human auditory sensation of tone height. In fact, Mel Frequency Cepstrum Coefficients (MFCC) are almost universally used in the speech community to build acoustic front-end for ASR systems.

All these speech processing schemes make use of the "short-time" analysis framework [1]. Short segments of speech are isolated and processed as if they were short segments from a sustained sound with fixed properties. In order to better track dynamical changes of speech properties, these short segments which are called *analysis frames*, overlap one another. This framework is based on the underlying assumption that, due to human articulatory characteristics, the properties of the speech signal change relatively slowly with time. Even if overlapped analysis windows are used, important fine dynamic characteristics of speech signal are discarded. Just for that reason, but without solving completely the problem of correctly taking into account the dynamic properties of speech, "velocity"-type parameters (simple

differences among parameters of successive frames) and "acceleration"-type parameters (differences of differences) [8] have been recently included in acoustic front end of almost all ASR systems found on the market. The use of these temporal changes in speech spectral representation (i.e. ΔMFCC, ΔΔMFCC) has given rise to one of the greatest improvements in ASR systems. In some of the best ASR systems, the incorporation of transitional information has reduced errors by as much as 50%. [9], [10].

Moreover, in order to overcome the resolution limitation of the STFT (due to the fact that once the analysis window has been chosen , the time frequency resolution is fixed over the entire time-frequency plane, since the same window is used at all frequencies), a new technique called Continuous Wavelet Transform (CWT), characterized by the capability of implementing multiresolution analysis, has been recently introduced [11]. With this new speech processing scheme, if the analysis is viewed as a filter bank, the time resolution increases with the central frequency of the analysis filters. In other word, different analysis windows are simultaneously considered in order to more closely simulate the frequency response of the human cochlea. As with the preceeding processing schemes, this new auditory-based tecnique, even if it is surely more adequate than STFT analysis to represent a model of human auditory speech processing, it is still based on a mathematical framework built around a transformation of the speech waveform, from which it tries directly to extrapolate a more realistic perceptual behaviour .

Cochlear transformations of speech signals result in an auditory neural firing pattern significantly different from the spectral pattern obtained from the speech waveform by using one of the above mentioned techniques. In other words, speech spectral representations such as the *spectrogram*, a popular time-frequency-energy representation of speech, or either the *wavelet spectrogram*, or *scalogram*, obtained using the above described multiresolution analysis technique are quite different from the true *neurogram*. In recent years, basilar membrane, inner cell and nerve fiber behaviour have been extensively studied by auditory physiologists and neurophysiologists and knowledge about the human auditory pathway has become more accurate. A number of studies have been accomplished and a considerable amount of data has been gathered in order to characterize the responses of nerve fibers in the eighth nerve of the mammalian auditory system using tone, tone complexes and synthetic speech stimuli [12-21]. Phonetic features probably correspond in a rather straightforward manner to the neural discharge pattern with which speech is coded by the auditory nerve.

Various auditory models which try to *physiologically* reproduce the human auditory system have been developed in the past [22], and, even if they must be considered as only an approximation of physical reality, they appear to be a suitable system for identifying those aspects of the speech signal that are relevant for automatic speech analysis and recognition. Furthermore, with these models of auditory processing, perceptual properties can be re-discovered starting not from the sound pressure wave, characterising speech, but from a more internal representation which is intended to represent the true information available at the eighth acoustic nerve of the human auditory system.

Advanced Auditory Modelling (AM) techniques not only follow "perception-based" criteria instead of "production-based" ones, but also overcome "short-term" analysis limitations, because they implicitely retain dynamic and nonlinear speech characteristics. For example, the dynamics of the response to non-steady-state signals, as also "forward masking" phenomena, which occur when the response to a particular sound is diminished as a consequence of a preceeding, usually considerably more intense signal, are important aspects captured by efficient auditory models[23]. Various evidences can be found in the litterature [24-27] suggesting the use of AM techniques, instead of more classical ones, in building speech analysis and recognition systems. Especially when speech is greatly corrupted by noise [27-28], the effective power of AM techniques seems much more evident than that of classical digital signal processing schemes.

2. Joint Synchrony/Mean-Rate Auditory Speech Processing.

The computational scheme proposed in this paper for modelling the human auditory system, apart from small differences regarding the filter bank designing strategy, refers essentially to the joint *Synchrony/Mean-Rate* (S/M-R) model of *Auditory Speech Processing* (ASP), recently proposed by S. Seneff [23], resulting from her important studies on this matter [29-31]. The overall system structure, whose block diagram is illustrated in Fig. 1, includes three stages: the first two deal with peripheral transformations occurring in the early stages of the hearing process while the third one attempts to extract information relevant to perception. The first two blocks represent the *periphery* of the auditory system. They are designed using knowledge of the rather well known responses of the corresponding human auditory stages [20-21]. The third unit attempts to apply a useful processing strategy for the extraction of important speech properties like an efficient representation for locating transitions between phonemes useful for speech segmentation, or spectral lines related to formants useful for phonetic identification.

The speech signal, band-limited and sampled at 16 kHz, is first pre-filtered through a set of four complex zero pairs to eliminate the very high and very low frequency components. The signal is then analyzed by the first block, a *40-channel critical-band linear filter bank*. Fig 2 shows the block diagram of the filter bank which was implemented as a cascade of complex high frequency zero pairs with taps after each zero pair to individual tuned resonators. Filter resonators consist of a double complex pole pair corresponding to the filter center frequency (CF) and a double complex zero pair at half its CF. Although a larger number of channels would provide superior spatial resolution of the cochlear output, the amount of computation time required would be increased significantly. The bandwidth of the channels is approximately 0.5 Bark, which corresponds to the width of one critical band that is a unit of frequency resolution and energy integration derived from psychophysical experiments [32]. Filters, whose transfer functions are illustrated in Fig. 3, were designed in order to optimally fit physiological data like those observed by N.Y.S. Kiang et al. [21]. Frequencies and bandwidths for zeros and poles of each filter were

designed almost automatically by an interactive technique developed by S. Seneff and described in her Thesis [30]. As for the mathematical implementation of the 40-channnel critical-band filter bank, it is described on the top of Fig. 4, where serial (FIR) and parallel (IIR) branches are illustrated in detail.

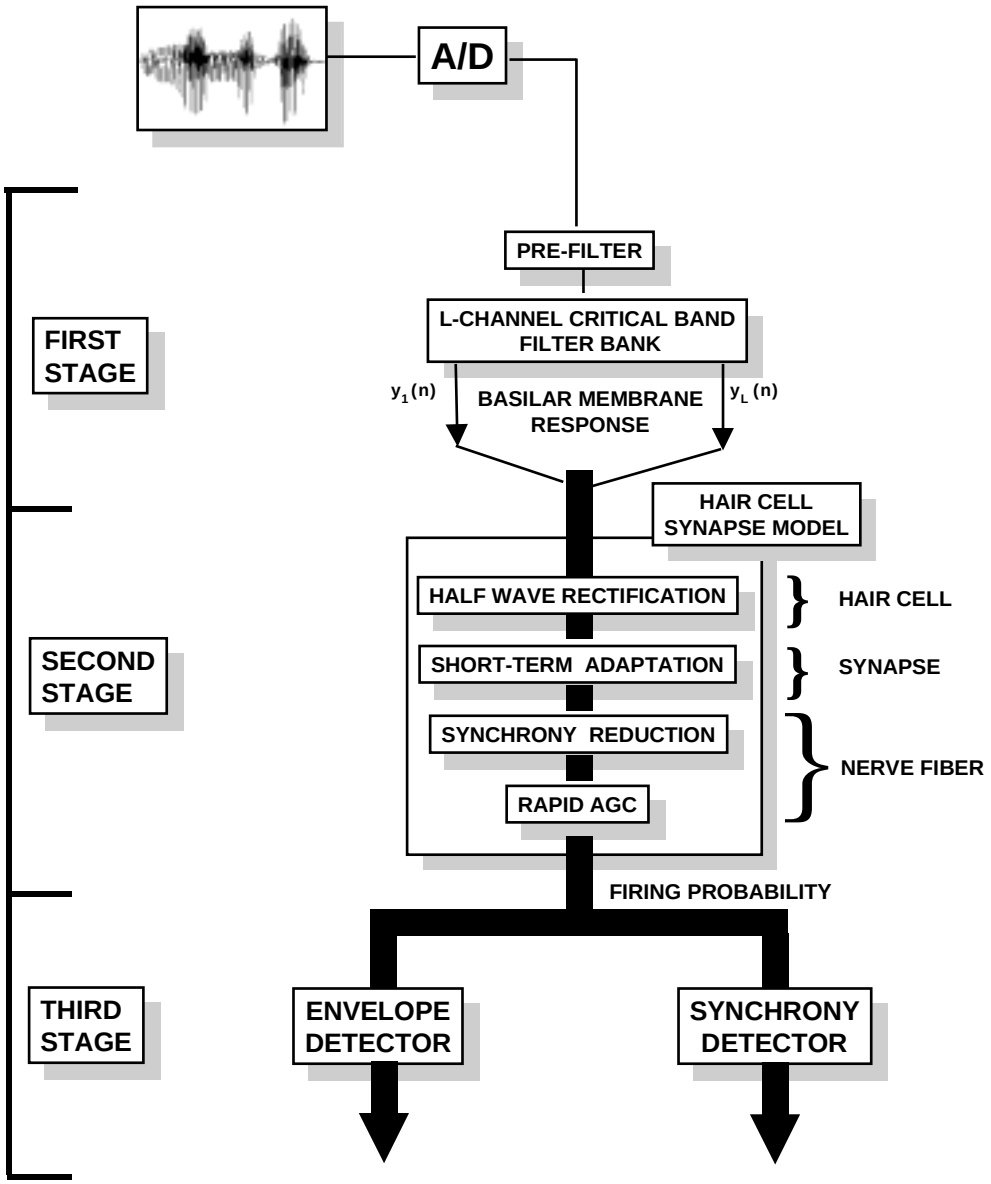

Fig. 1. Block diagram of the joint Synchrony/Mean-Rate model of Auditory Speech Processing.

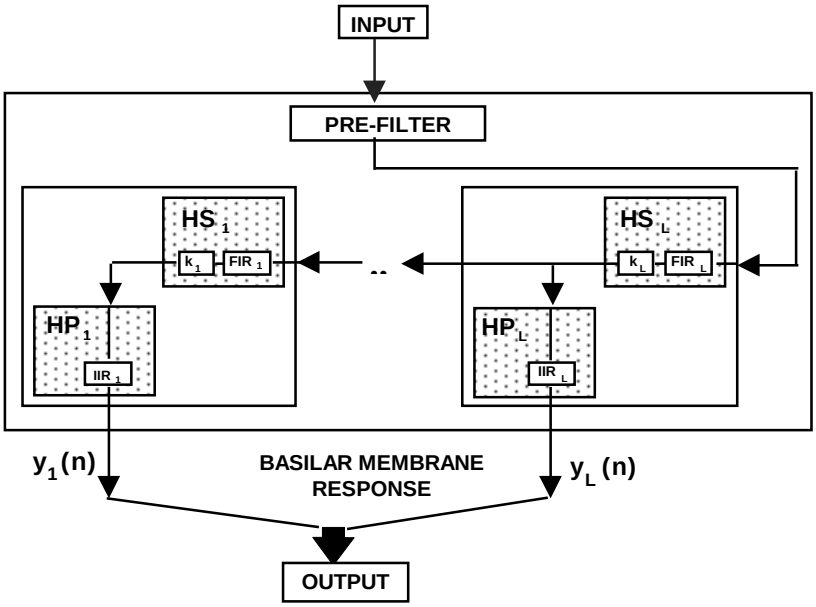

Fig. 2. Block diagram of the 40-channel critical-band linear filter bank.

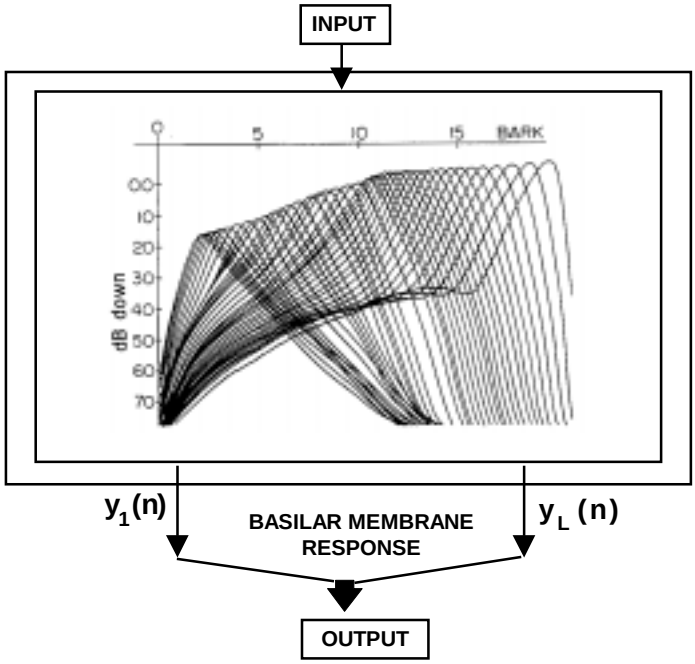

Fig. 3. Transfer Functions of the 40-channel critical-band linear filter bank.

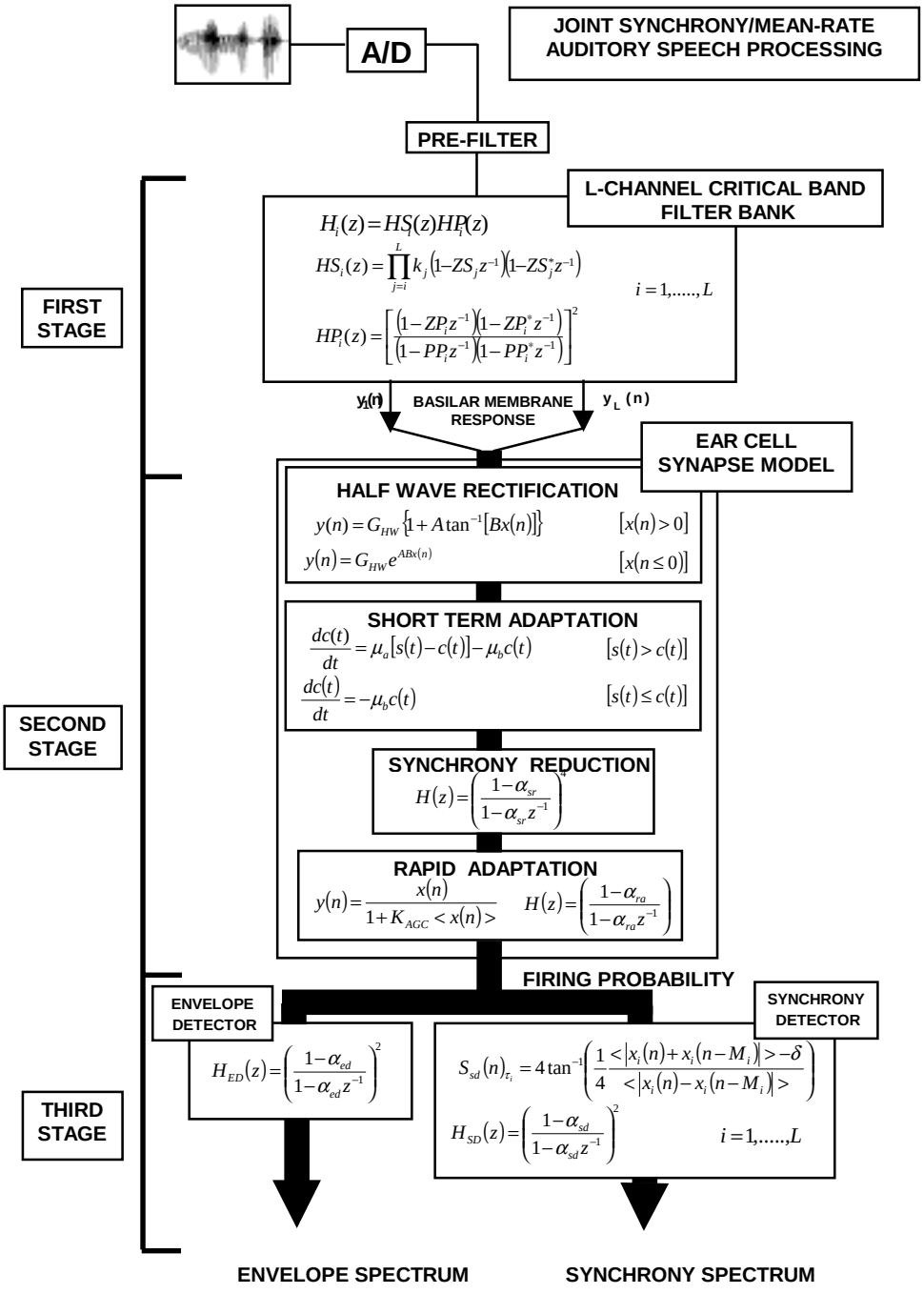

Fig. 4. Mathematical framework of the joint Synchrony/Mean-Rate model of Auditory Speech Processing.

The second stage of the model is called the *hair cell synapse model* (see Fig. 1). It is nonlinear and is intended to capture prominent features of the transformation from basilar membrane vibration, represented by the outputs of the filter bank, to probabilistic response properties of auditory nerve fibers. The outputs of this stage, in accordance with S. Seneff [23], represent the probability of firing as a function of time for a set of similar fibers acting as a group. Four different neural mechanisms are modeled in this nonlinear stage. A *half-wave rectifier* is applied to the signal in order to simulate the high level distinct directional sensitivity present in the inner hair cell current response. This rectifier is the first component of this stage and is implemented by the use of a saturating non linearity. The istantaneous discharge rate of auditory-nerve fibers is often significantly highest during the first part of acoustic stimulation and decreases thereafter, until it reaches a steady-state level. The *short-term adaptation* module, which controls the dynamics of this response to non steady-state signals which is due to the neurotransmitter release in the synaptic region between the inner hair cell and its connected nerve fibers, is simulated by the so called "membrane model", which was conceived following the work by R.S. Goldor [33]. This model influences the evolution of the neurotransmitter concentration inside the cell membrane. The third unit implements the observed *gradual loss of synchrony* in nerve fiber behaviour as stimulus frequency is increased, and it is implemented by a simple low-pass filter. The last unit is called *Rapid Adaptation* and implements the very rapid initial decay in discharge rate of auditory nerve-fibers occurring immediately after acoustic stimulation onset, followed by the slower decay, due to short-term adaptation, to a steady state level. This module performs "Automatic Gain Control" and is essentialy inspired by the refractory property of auditory nerve fibers [34]. The final output of this stage is affected by the ordering of the four different components due to their nonlinear behaviour. Consequently, as underlined by S. Seneff [23], each module is positioned by considering its hypothesized corresponding auditory apparatus (see Fig. 1). As for the mathematical implementation of the four modules of the hair-cell synapse model, this is illustrated in the central block of Fig. 4. Fig. 5 describes the result of the application of the model to a simple 1000Hz sinusoid. Left and right plots refer respectively to the global 60ms stimulus and to its corrsponding first 10ms window in different positions along the model.

The third and last stage of the model, mathematically described on the bottom of Fig. 4, is formed by the union of two parallel blocks: the *Envelope Detector* (ED), implemented by a simple low-pass filter, which, in accordance with S. Seneff [23], by smoothing and downsampling the second stage outputs, appears to be an excellent representation for locating transition between phonemes, thus providing an adequate basis for phonetic segmentation, and the *Synchrony Detector* (SD), whose block diagram as applied to each channel is shown in Figure 6, which implements the known "phase locking" property of the nerve fibers. This block enhances spectral peaks due to vocal tract resonances. In fact, auditory nerve fibers tend to fire in a "phase-locked" way responding to low frequency periodic stimuli, which means that the intervals between nerve fibers tend to be integral multiples of the stimulus period. Consequently, if there is a "dominant periodicity" (a prominent peak in the frequency domain) in the signal, with the so called *Generalized Synchrony Detector* (GSD) processing technique [29-30], only those channels whose central frequencies are

closest to that periodicity will have a more prominent response. The use of GSD parameters allow to produce spectra with a limited number of well defined spectral lines and this represents a good use of speech knowledge according to which formants are voiced sound parameters with low variance.

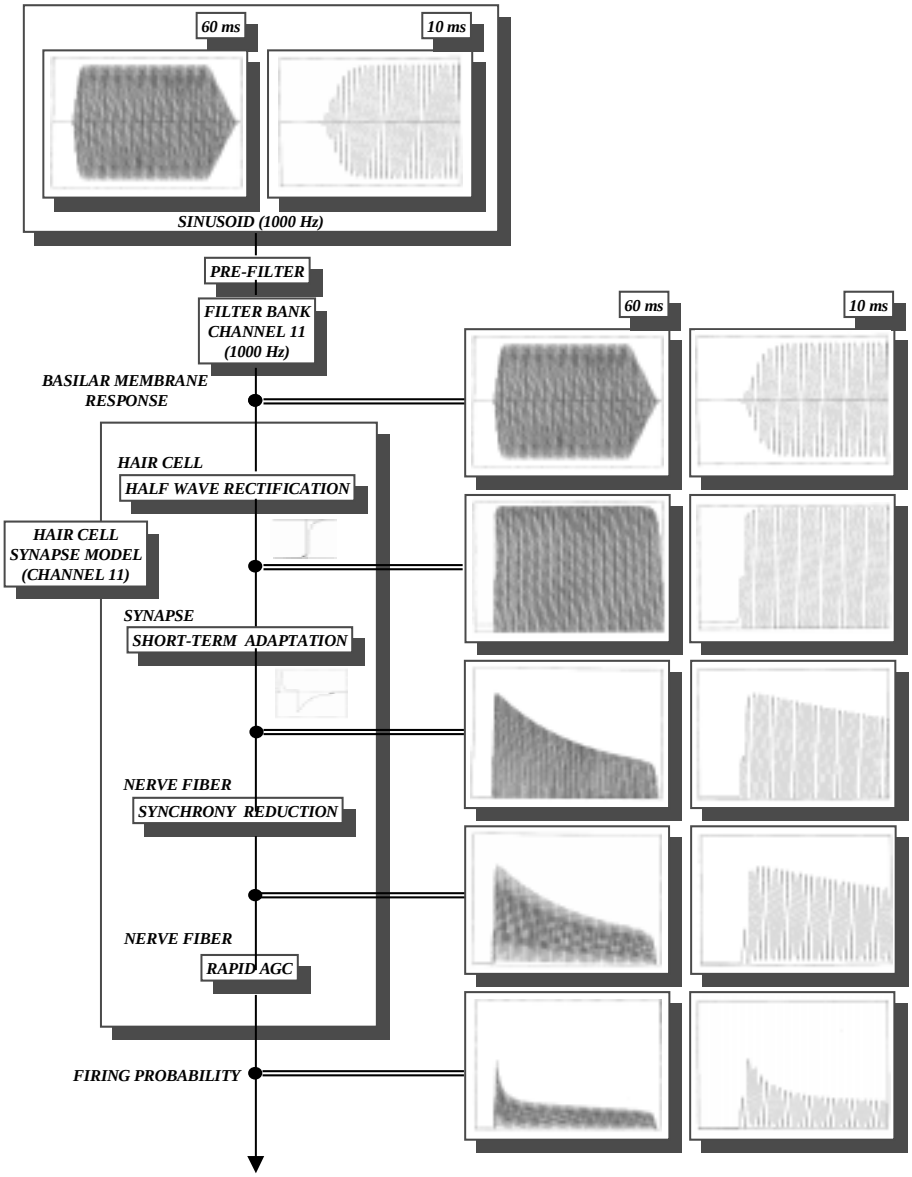

Fig. 5. Result of the application of the four modules implementing the hair-cell synapse model to a simple 1000Hz sinusoid. Left and right plots refer to the global 60ms stimulus and to its corrsponding first 10ms window, in different positions along the model.

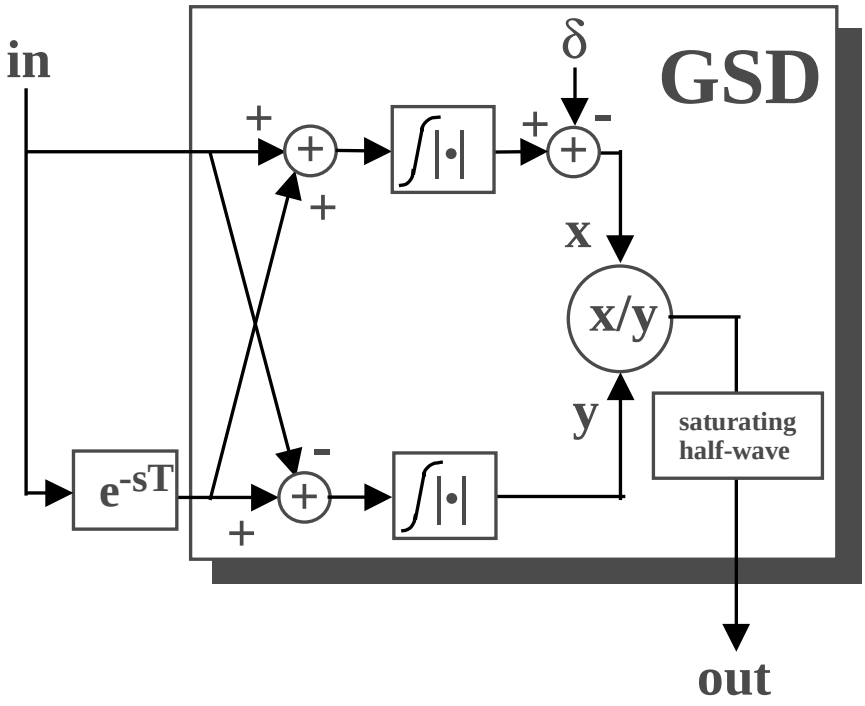

Fig. 6. Block diagram of the Generalized Synchrony Detector (GSD) module.

In Fig. 7, an example of the output of the model, as applied to the English sentence 'Susan ca(n't)' uttered in a *clean* (on the left) and in a *noisy* (on the right) environment by a female speaker [27] (last two consonants are omitted), is illustrated for the envelope (b) and the synchrony (c) detector module respectively. In (a), manual segmentation made by an Italian mother-tongue phonetician is superimposed to the speech waveform. Plots on the right of the figure refer to the noisy case when the sentence is highly corrupted by a superimposed natural noise [27] The effectiveness of using this model is quite evident from a comparison between the two sonogram-like plots produced by the GSD (c). Observing the low frequency components it is evident that the formant structure is well preserved even if speech is greatly corrupted by quite a relevant noise.

The computation time of the joint S/M-R model of ASP is about 150 times real-time on a SUN 4/280. The system structure is suitable for parallelization with special purpose architectures and accelerator chips. At the present time the model has been also implemented on a floating-point Digital Signal Processor and the obtained computation time is about 10 times real-time [35].

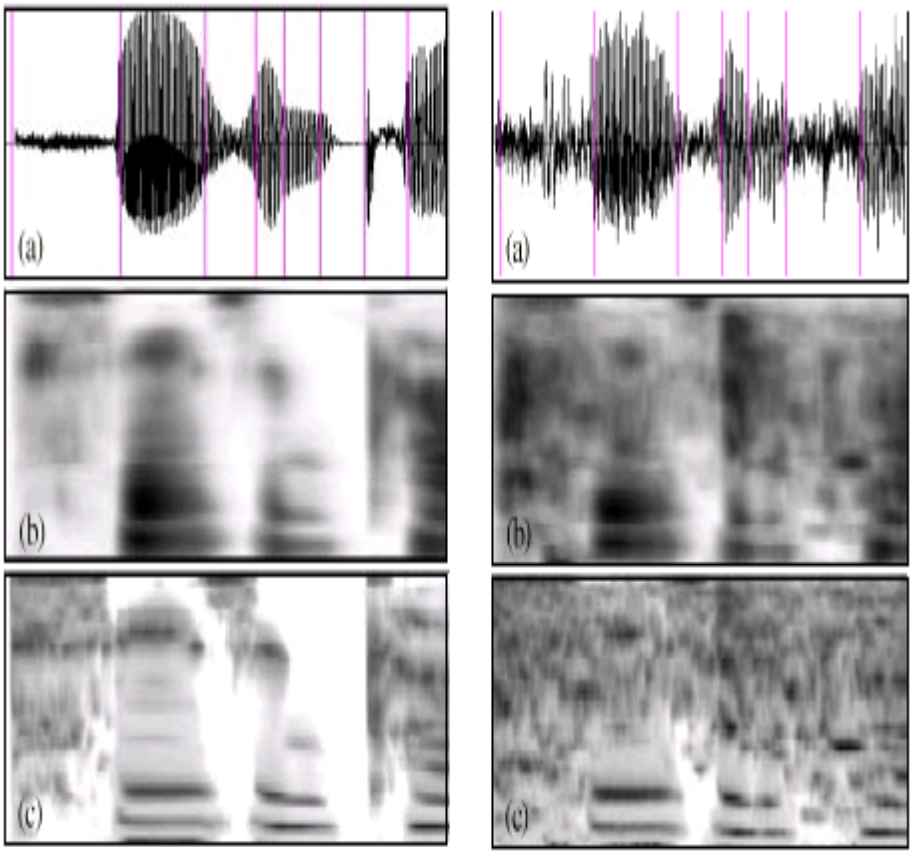

Fig. 7. Output of the model, as applied to the English sentence 'Susan ca(n't)' uttered in a *clean* (on the left) and in a *noisy* (on the right) environment by a female speaker (last two consonants are omitted). In (a), manual segmentation made by an Italian mother-tongue phonetician is superimposed to the speech waveform. (b) and (c) represent the ED and SD spectrogram-like plot respectively. Noisy sentence has been created starting from the clean one by superimposing natural speech noise.

3. Auditory Modelling and Speech Segmentation.

The joint S/M-R model of Auditory Speech Processing provides an adequate basis both for phonetic segmentation and for phonetic labelling or identification. In fact, the envelope detector module, which smooths second stage outputs, well preserves dynamic features of input speech thus allowing to discover transitions between phonemes very precisely and efficiently, while the synchrony detector, producing spectra with a limited number of well defined spectral lines, provides a useful representation for identifing different sounds.

In order to determine the location of onset and offset events corresponding to different phonemes of input speech, two viewpoints are essentially considered. The first one is that these events represent a local maximum or minimum in some parameter representing the speech signal, since at these points the signal is undergoing significantly more change than in the neighboring environment. As for the second viewpoint, to which the segmentation algorithm considered in this work belongs, speech is considered as a temporal sequence of quasi-stationary acoustic segments, and the points within such segments are more similar to each other than to the points in adjacent segments. The segmentation problem can thus be simply reduced to a local clustering problem where the decision to be taken regards the similarity of any particular frame with the sound immediatly preceeding or following it. Furthermore, using only relative measures of acoustic similarity, this technique should be quite independent of the speaker, vocabulary, and background noise.

The semi-automatic segmentation algorithm briefly summarised in the following has been developed by J.R. Glass and V.W. Zue [36-38] and is called *Multi Level Segmentation* (MLS) algorithm. The present implementation is based entirely on both ED and SD parameters while, in its original formulation, only ED parameters are used.

For each target frame, within its left and right window of Δ frames lenght (Δ can be set to different values), an average value for each analysis vector component is computed. Depending on an euclidean-based similarity measure, forward and backward distances between the current frame and the right and left window are calculated and a decision is taken in associating the current frame to its immediate past or to its immediate future. Various strategies can be adopted in defining forward and backward distances allowing the possibility of adapting the sensitivity of the association to the local environment [38]. After all frames have been analyzed various adjacent regions are created. These initial 'seed regions' constitute the basis for the following 'hierarchical structuring' segmentation procedure suggested by the fact that the speech signal is characterized by short events that are often quite distinct from their local environment. This hierarchical technique, incorporating some kind of temporal constraint, is quite useful in order to appropriately rank the significance of acoustic events. The clustering scheme utilized to produce a multi-level description of the speech signal is based essentially on the same framework used for locating 'seed acoustic events'. In fact, starting from previously calculated initial 'seed regions', each region is associated with either its left or right neighbor using an euclidean-based similarity measure, where the similarity measure is computed with a distance measure applied to the average spectral analysis vector of each region. Two regions are merged together to form a single region when they associate with each other and this new created region subsequently associates itself with one of its neighbors. The process is repeated until the whole utterance is analysed and described by a single acoustic event. By keeping track of the distance at which two regions merge into one, a multi-level description usually called *dendrogram,* like that described in Fig. 8 referring to the Italian sentence 'Che senso ha.....?' (What does it mean.....?) can be constructed.

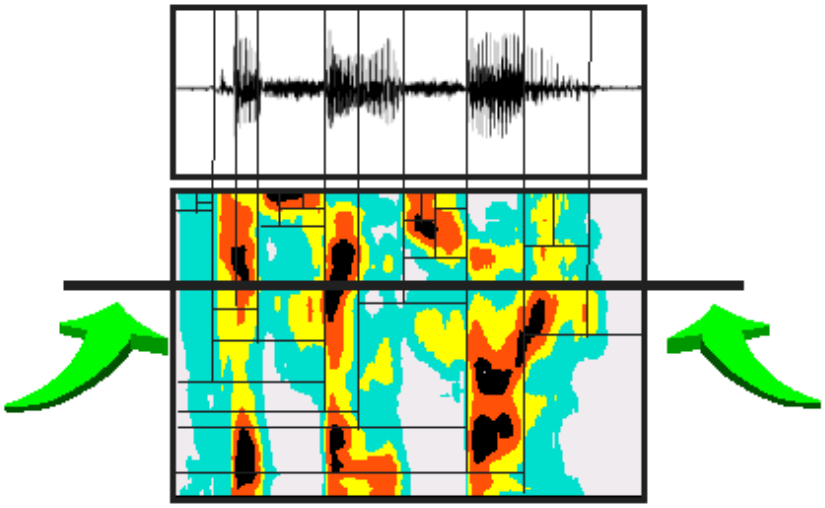

Fig. 8. Multi-level segmentation tree ('Dendrogram') built on auditory speech representation for the Italian sentence. 'Che senso ha.....?' (What does it mean.....?)

The final target segmentation can be extracted automatically [24] by appropriate pattern recognition techniques whose aim is to find the optimal segmentation path given the dendrogram structure and the target phonemic transcription of the input sentence, but also with minimal human intervention, which is limited exclusively on fixing the vertical point determining the final target segmentation (corresponding to that found on the horizontal line built on this point), and eventually deleting over-segmentation landmarks forced by this choice. Even using the above described manual intervention, segmentation marks are always automatically positioned by the system and never adjusted by hand.

Advantages of using auditory models vs classical "short-term" analysis approaches for automatic speech segmentation have been shown in litterature [36-38], especially in adverse conditions [27]. A graphic example of the output produced by the application of this algorithm with two different input signal representations to the same noisy English sentence considered in Figures 7 is illustrated in Fig. 9. The same algorithm, applied to a "FFT-based spectrogram" (Fig. 9b) and to the "AM-based" one (Fig 9a), produces a more confusable segmentation "dendrogram" in the first case, from which the final target segmentation is much more difficult to extract.

Even if speech is clearly degraded by quite a relevant noise, ASP parameters lead MLS algorithm to compute very clear and reliable segmentation landmarks, while, on the contrary, FFT parameters cause serious problems in finding a possible segmentation line throughout the dendrogram structure. In other words, over-segmentation marks (gross errors), always produced by the use of FFT parameters, are totally or heavily reduced by the use of ASP parameters. This result leads obviously to a better starting point for building a real automatic segmentation system [24]. In fact,

walking through the "dendrogram" from left to right, in order to automatically find the optimal segmentation path, clean multi-level structures would surely be more useful than very complicate ones.

The MLS system has been incorporated in CASPAR, an automatic transcription and alignment system developed at MIT [39-40] which has been used also for the transcription alignment of the TIMIT database produced by the DARPA consortium. In a formal evaluation of TIMIT automatic alignment, the boundary locations produced by the system agree well with those produced by human transcribers. For example, over 75% of the automatically generated boundaries are within 10 ms of a boundary entered by trained phoneticians [41]. Other speech semi-automatic segmentation experiments, both in clean and noisy conditions, [27] have shown more than 90% agreement between semi-automatic and human positioned landmarks.

As for the computation complexity of the MLS algorithm, considering the fact that it does not make use of the entire utterance for emitting segmentation hypothesis but it shows a local behaviour, it is capable of analysing speech signal virtually istantaneously.

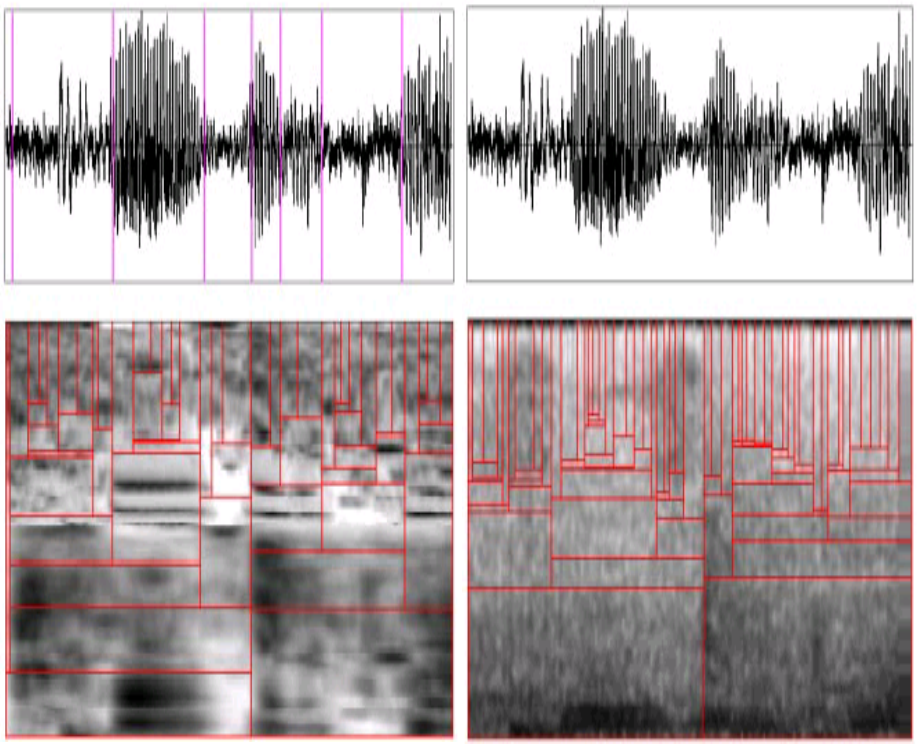

Fig. 9. Comparison of "dendrograms" produced by the application of the MLS algorithm to the English noisy sentence "Susan ca(n't)" (last two consonants are omitted), using AM parameters (a) and FFT parameters (b).

4. Auditory Modelling and Speech Recognition.

As already underlined, various evidences [24-28], suggest the effectiveness of ASP techniques for speech analysis and recognition, especially in speech adverse conditions [27-28]. Results of the application of this model in previous recognition experiments [25] have been compared with those obtained by using a classical FFT-based front-end. In that particular case, a vowel identification task, the use of AM parameters has shown better recognition performance than the use of classical FFT-based coefficients. Moreover, in that experiment a combination of Seneff's Auditory Modelling technique and multi-layer neural networks gives rise to an effective generalization among speakers in coding vowels.

Furthermore, considering an extremely difficult Italian phonetic recognition problem [26], the automatic discrimination of the so called Italian i-set (/bi/, /tSi/, /di/, /dZi/, /i/, /pi/, /ti/, /vi/), plus other two i-like stimuli /Li/, /si/ (see SAMPA Phonetic Alphabet [42]), the achieved speaker independent mean recognition rate is around 65% which, given the effective difficulty of the task, can be considered quite acceptable and promising. In this experiment using Recurrent Neural Networks (RNN) as the global recognition framework, input speech signal is sampled at 16kHz, in a quiet office room, it is analyzed with the joint S/M-R model of ASP in order to produce adequate speech signal representation, and successively segmented using the MLS algorithm in order to locate onset and offset of target stimuli. In Fig. 10 a block-diagram of the whole system used in [26] is described.

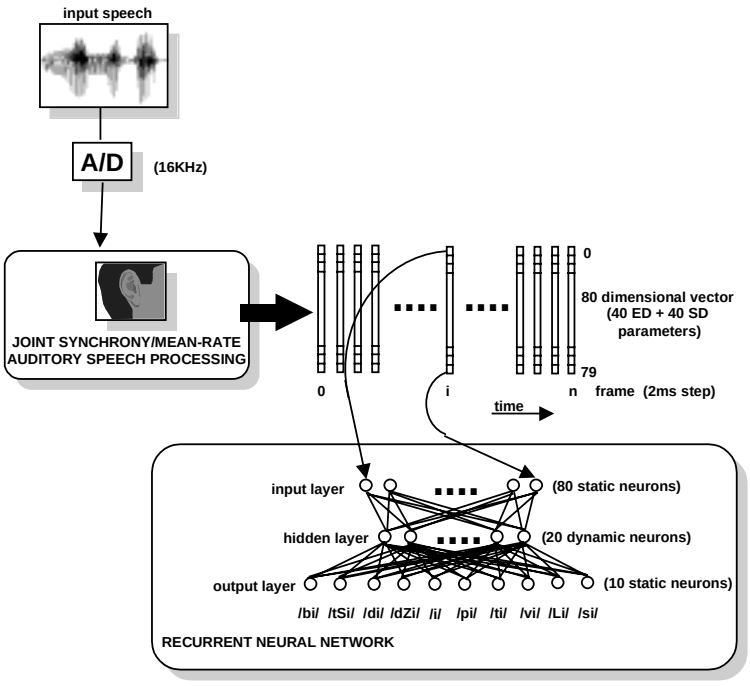

Fig. 10. Block diagram of the speech recognition system described in [26] (see text).

More studies are needed to reinforce the conclusion that the proposed perception-based auditory analysis could perform better than other acoustic production-based front-end (LPC, MEL-scale cepstrum, etc. ...) in speech recognition tasks.

5. Conclusions

In this work a few evidences in favour of using Auditory Modelling tecniques instead of more classical production-based ones (FFT, LPC, Cepstrum, ets...) in Speech Processing technology are summarized

Auditory Modelling is still a young research field. Further knowledge on human auditory functioning has been acquired during last years and surely more discovering will be made in the future. When it will be possible to easily incorporate all this knowledge into effective real-time Digital Speech Processing algorithms perhaps speaker independent speech recognition could become a reality.

Aknowledgements

This work has been made possible by the kind help of Stephanie Seneff of MIT who was even too patient in clarifying her Joint Synchrony/Mean-Rate Model of Auditory Speech Processing thus speeding up implementation time.

References

[1] L.R. Rabiner and R.W. Shafer (1978), "Digital Processing of Speech Signals", Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1978.
[2] J.D. Markel and A.H. Gray (1976), Jr., "Linear Prediction of Speech", Springer-Verlag, Berlin, Heidelberg, New York, 1976.
[3] H.W. Strube (1976), "Linear prediction on a warped frequency scale", Journal of Acoustical Society of America, JASA Vol. 68(4), Oct. 1980, pp. 1071-1076.
[4] M. Blomberg, R. Carlson,K. Elenius and Bjorn Granstrom (1983), "Auditory Models and Isolated Word Recognition", STL-QPSR Vol. 4, 1983, pp.1-15.
[5] H. Hermansky, B.A. Hanson and H. Wakita, "Perceptually Based Linear Predictive Analysis of Speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-84), paper 13.10, pp.509-512.
[6] H. Hermansky, J.C. Junqua, "Optimization of Perceptually Based ASR Front-Ends", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-88), New York, N.Y. April 11-14, 1988, paper S5.10, pp.219-222.

[7] S.B. Davis and P. Mermelstein (1980), "Comparison of Parametric Representation of Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP Vol. 28(4), pp. 357-366.
[8] S. Furui (1986), "Speaker Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum", IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP Vol. 34(1), pp. 52-59.
[9] L.R. Rabiner, J.G. Wilpon and F.K. Soong (1988), "High Performance Connected Digit Recognition Using Hidden Markov Models", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-88), New York, N.Y., April 11-14, 1988, paper S3.6, pp.119-122.,
[10] K.F. Lee (1989), "Automatic Speech Recognition; The Development of the SPHINX System", Kluwer Academic Publisher, Boston, 1989.
[11] O. Rioul and M. Vetterli, "Wavelets and Signal Processing", IEEE Signal Processing Magazine, October 1991, pp.14-38.
[12] B. Delgutte (1980), "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers" , Journal of the Acoustical Society of America, JASA Vol. 68, 1980, pp. 843-857.
[13] B. Delgutte and N.Y.S. Kiang (1984), "Speech coding in the auditory nerve: I. Vowel-like sounds", Journal of Acoustical Society of America, JASA Vol. 75, 1984, pp. 866-878.
[14] B. Delgutte and N.Y.S. Kiang (1984), "Speech coding in the auditory nerve: II. Processing Schemes for Vowel-like sounds", Journal of Acoustical Society America, JASA Vol. 75, 1984, pp. 897-907.
[15] B. Delgutte and N.Y.S. Kiang (1984), "Speech coding in the auditory nerve: III. Voiceless fricative consonants", Journal of Acoustical Society of America, JASA Vol. 75, 1984, pp. 887-896.
[16] B. Delgutte and N.Y.S. Kiang (1984), "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics", Journal of Acoustical Society America, JASA Vol. 75, 1984, pp. 897-907.
[17] E. D. Young and M.B. Sachs (1979),"Representation of steady-state vowels in the temporal aspects of the discharge pattern of populations of auditory nerve fibers", Journal of Acoustical Society of America, JASA Vol. 66, 1979, pp. 1381-1403.
[18] M. B. Sachs and E. D. Young (1980), "Effects of nonlinearities on speech encoding in the auditory nerve", Journal of Acoustical Society of America, JASA Vol. 68, 1980, pp. 858-875.
[19] M. I. Miller and M. B. Sachs (1983), "Representation of stop consonants in the discharge patterns of auditory-nerve fibers", Journal of Acoustical Society of America, JASA Vol. 74, 1983, pp. 502-517.
[20] D. G. Sinex and C. D. Geisler (1983), "Responses of auditory-nerve fibers to consonant-vowel syllables", Journal of Acoustical Society of America, JASA Vol. 73, 1983, pp. 602-615.
[21] N.Y.S. Kiang, T. Watanabe, E. C. Thomas and L. F. Clark (1965), Discharge patterns of single fibers in the cat's auditory-nerve fibers, Cambridge, MA, MIT press, 1965.
[22] S. Greenberg eds. (1988), "Representation of Speech in the Auditory Periphery", Journal of Phonetics, Special Issue, Vo. 16(1), January 1988.

[23] S. Seneff (1988), "A joint synchrony/mean-rate model of auditory speech processing", Journal of Phonetics, Special Issue, Vol. 16(1), January 1988, pp. 55-76.
[24] V.W. Zue, J. Glass, M. Philips and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT System",Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-89), paper S8.1, pp. 389-392.
[25] P. Cosi, Y. Bengio and R. De Mori,(1990), "Phonetically-Based Multi-Layered Neural Networks for Vowel Classification", Speech Comm., Vol. 9, N. 1, Feb 1990, pp. 15-29.
[26] P. Cosi, P. Frasconi, M. Gori and N. Griggio, "Phonetic Recognition Experiments with Recurrent Neural Networks", Proc. International Conference on Spoken Language Processing (ICSLP-92), Banff, Alberta, Canada, October 12-16, 1992, pp. 1335-1338
[27] P. Cosi, "Ear Modelling for Speech Analysis and Recognition" (1992), Proceedings of "Comparing Speech Signal Representations", ESCA Tutorial and Research Workshop, Sheffield, England, 8-9 April 1992; paper ISSN 1018-4554 (to be published in J. Wiley & sons L.t.D. book).
[28] M.J. Hunt and C. Lefebvre,"Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-88), New York, N.Y., April 11-14, 1988, paper S5.9, pp. 215-218.
[29] S. Seneff (1984) , "Pitch and spectral estimation of speech based on an auditory synchrony model", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-84), San Diego, CA, March 19-21, 1984, pp. 36.2.1-36.2.4.
[30] S. Seneff (1985), "Pitch and spectral analysis of speech based on an auditory synchrony model", RLE Technical Report, No. 504 , Mass. Inst. of Techn., 1985.
[31] S. Seneff (1986), "A computational model for the peripheral auditory system: application to speech recognition research", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-86), Tokyo, April 7-11,1986, pp. 37.8.1-37.8.4.
[32] E. Zwicker and E. Terhardt, "Analytical expression for critical-band rate and critical bandwidth ad a function of frequency", Journal of Acoustical Society of America, JASA Vol. 68(5), 1980, pp. 1523-1525.
[33] R.S. Goldhor (1985), "Representation of Consonants in the Peripheral Auditory System: A Modeling Study of the Correspondence between Response Properties and Phonetic Features", RLE Technical Report , N. 505, MIT press, 1985.
[34] D.H. and A. Swami (1983), "The transmission of signals by auditory-nerve fiber discharge patterns", Journal of Acoustical Society of America, JASA Vol. 74, pp. 493-501.
[35] P. Cosi, L. Dellana, G.A. Mian and M. Omologo (1991), "Auditory Model Implementation on a DSP32C-Board", Proc. GRETSI-91, Juan Les Pins, 16-20 Sep 1991.
[36] J.R. Glass and V.W. Zue (1986), "Signal Representation for Acoustic Segmentation", Proc. First Australian Conference on Speech Science and Technology, November 1986, pp.124-129.

[37] J.R. Glass and V.W. Zue (1988), "Multi-Level Acoustic Segmentation of Continuous Speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-88), New York, N.Y., April 11-14, 1988, paper S10.6, pp. 429-432.
[38] J.R. Glass (1988), "Finding Acoustic Regularities in Speech: Application to Phonetic Recognition", Ph. D Thesis, May 1988, MIT press.
[39] H.C. Leung and V. W. Zue (1984), "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-84), San Diego, CA, March 19-21, 1984, pp. 2.7.1-2.7.4.
[40] H.C. Leung (1984), "A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech", S.M. Thesis, Department of Electrical Engineering and Computer Science Massachussets Institute of Technology, January 1985.
[41] S. Seneff and V.W Zue (1988), "Transcription and Alignment of the TIMIT Database", Unpublished manuscript to be distributed with the TIMIT database by NBS, 1988.
[42] A.J. Fourcin, G. Harland, W. Barry and W. Hazan eds. (1989), "Speech Input and Output Assessment, Multilingual Methods and Standards, Ellis Horwood Books in Information Technology, 1989.