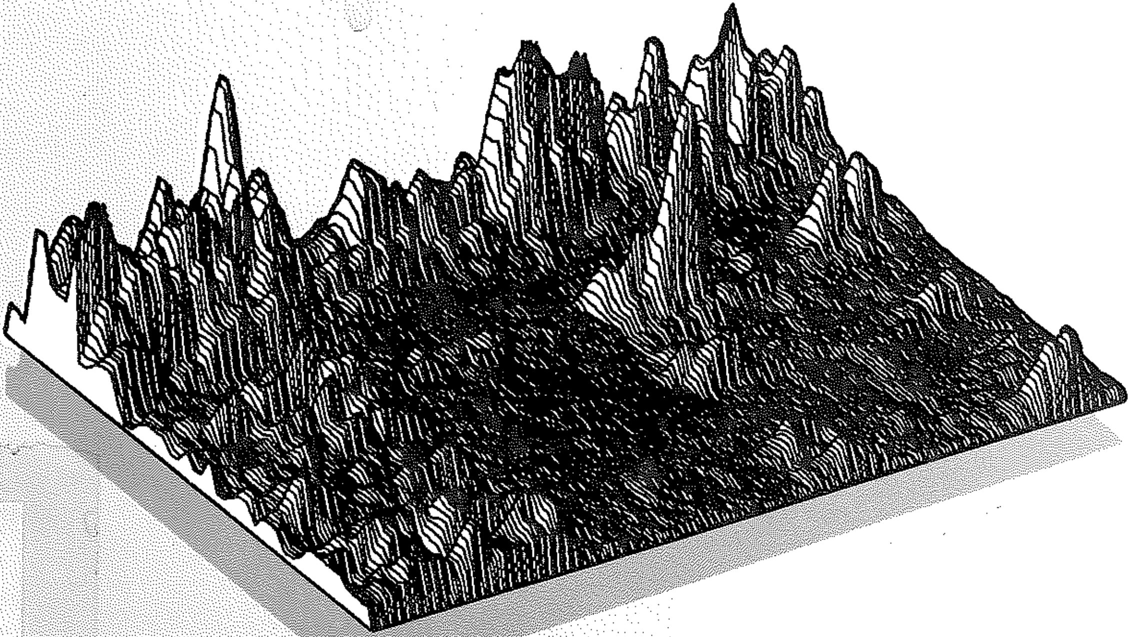
WILEY   WILEY PROFESSIONAL COMPUTING

VISUAL REPRESENTATIONS OF SPEECH SIGNALS



Edited by Martin Cooke, Steve Beet and Malcolm Crawford

AUDITORY MODELLING FOR SPEECH ANALYSIS AND RECOGNITION

1

Piero Cosi
Centro di Studio per le Ricerche di Fonetica
C.N.R. P.zza Salvemini, 13 - 35131 Padova, ITALY

1 INTRODUCTION

Cochlear transformations of speech signals result in an auditory neural firing pattern significantly different from the spectrogram, a popular time-frequency-energy representation of speech. Phonetic features may correspond in a rather straightforward manner to the neural discharge pattern with which speech is coded by the auditory nerve. For these reasons, even an ear model that is just an approximation of physical reality appears to be a suitable system for identifying those aspects of the speech signal that are relevant for recognition.

A recently developed joint Synchrony/Mean-Rate (S/M-R) Auditory Speech Processing (ASP) scheme [8] was successfully applied in speech recognition tasks, where promising results were obtained for speech segmentation and labelling [9]. Moreover, results reported elsewhere in the literature show that a combination of the same ASP scheme with multi-layer artificial neural networks produced an effective generalisation amongst speakers in classifying vowels both for English [1] and Italian [2].

The joint S/M-R ASP scheme will be very briefly described and its application to the problem of speech segmentation and labelling, both for clean and noisy speech, will be introduced and analysed.

2 AUDITORY SPEECH PROCESSING

The computational scheme proposed in this paper for modelling the human auditory system is derived from a joint Synchrony/Mean-Rate model proposed by Seneff [8]. The overall system includes three blocks: the first two of them deal with peripheral transformations occurring in the early stages of the hearing process while the third one attempts to extract information relevant to perception.

In fig. 1 a block diagram of the joint S/M-R ASP scheme is displayed together with its mathematical counterpart (for a complete description of the model refer to [8]). The first two blocks represent the auditory periphery. They are designed using knowledge of the well-known responses of the corresponding human auditory stages. The third unit attempts to apply a useful processing strategy for the extraction of important speech properties like spectral lines related to formants and also to show enhanced sharpness of onset and offset of different speech segments. 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 analysed by the first block, a 40 channel critical-band linear filter-bank whose single channels were designed in order to optimally fit physiological data.

The second block of the model is called the hair cell synapse model. 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 fibres. The outputs of this stage, in accordance with Seneff [8], represent the probability of firing as a function of time for a set of similar fibres acting as a group.

Visual Representations of Speech Signals: Martin Cooke & Steve Beet (eds.)

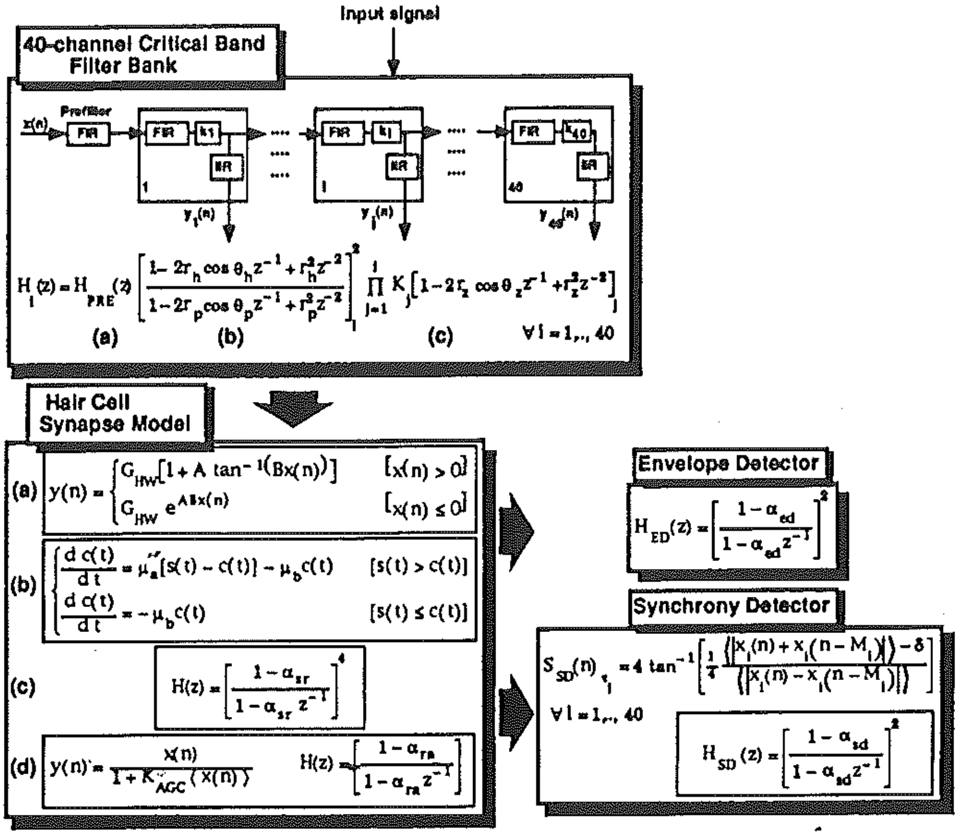

Input signal
40-channel Critical Band Filter Bank
Prefilter
x(n)
FIR
k1
IIR
1
$y_1(n)$
k_i
i
$y_i(n)$
k_{40}
40
$y_{40}(n)$
$H_i(z) = H_{PRE}(z)\left[\frac{1-2r_h\cos\theta_h z^{-1}+r_h^2z^{-2}}{1-2r_p\cos\theta_p z^{-1}+r_p^2z^{-2}}\right]_i^2 \prod_{j=1}^{i} K_j\left[1-2r_z\cos\theta_z z^{-1}+r_z^2z^{-2}\right]_j$
(a)
(b)
(c)
$\forall i = 1,.., 40$
Hair Cell Synapse Model
(a) $y(n) = \begin{cases} G_{HW}[1+A\tan^{-1}(Bx(n))] & [x(n)>0] \\ G_{HW}\, e^{ABx(n)} & [x(n)\le 0] \end{cases}$
(b) $\begin{cases} \frac{d\,c(t)}{d\,t} = \mu_a[s(t)-c(t)]-\mu_b c(t) & [s(t)>c(t)] \\ \frac{d\,c(t)}{d\,t} = -\mu_b c(t) & [s(t)\le c(t)] \end{cases}$
(c) $H(z) = \left[\frac{1-\alpha_{sr}}{1-\alpha_{sr}z^{-1}}\right]^4$
(d) $y(n) = \frac{x(n)}{1+K_{AGC}\langle x(n)\rangle}$ $H(z) = \left[\frac{1-\alpha_{ra}}{1-\alpha_{ra}z^{-1}}\right]$
Envelope Detector
$H_{ED}(z) = \left[\frac{1-\alpha_{ed}}{1-\alpha_{ed}z^{-1}}\right]^2$
Synchrony Detector
$S_{SD}(n)_i = 4\tan^{-1}\left[\frac{1}{4}\frac{\langle|x_i(n)+x_i(n-M_i)|\rangle-\delta}{\langle|x_i(n)-x_i(n-M_i)|\rangle}\right]$
$\forall i = 1,.., 40$
$H_{SD}(z) = \left[\frac{1-\alpha_{sd}}{1-\alpha_{sd}z^{-1}}\right]^2$

Fig. 1 Block diagram of the Synchrony/Mean rate Auditory Speech Processing scheme

The third and last block of the ear model is a double-unit block with two parallel outputs. The Generalized Synchrony Detector (GSD), which implements the known "phase-locking" property of nerve fibres, represents the first unit and is designed with the aim of enhancing spectral peaks due to vocal tract resonances. The second unit, called the Envelope Detector (ED) computes the envelope of signals at the output of the previous stage of the model and seems more important for capturing the very rapidly changing dynamic nature of speech. The outputs of this unit should be most important in characterising transient sounds.

The computation time of the joint S/M-R ASP system proposed in this paper is about 150 times real-time on a SUN 4/280. The system structure is suitable for parallelisation 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 computation time is about 10 times real-time [3].

In fig. 2 the output of the model applied to 'clean.syl' (a) from the Sheffield data set is illustrated for the envelope (b) and the synchrony detector (c) modules respectively. In (a), manual segmentation made by an Italian mother-tongue phonetician is superimposed on the speech waveform. The multi-line structure drawn in (b) refers to a particular output of the segmentation procedure which finally produces the target segmentation shown in (c). The use of the GSD parameters allows the production of 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. Figure 3 shows the same output resulting from the application of the model to the 'dirty.syl'. It is evident from a comparison of figs. 2c and 3c that the formant structure is well preserved by the S/M-R ASP, even if the speech is corrupted by noise.

3 SPEECH RECOGNITION

Various studies [6,9] suggest the effectiveness of ASP techniques for speech analysis and recognition, especially in adverse speech conditions [6]. Results of the application of this model in previous recognition experiments [1] were also compared with those obtained by using a classical FFT-based front-end. In that particular vowel recognition task the use of ear model coefficients showed better recognition performance than the use of classical FFT-based coefficients. Furthermore, other results on Italian phoneme recognition experiments [2] provided other evidences in favour of 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 analysis and recognition tasks.

4 SPEECH SEGMENTATION AND LABELLING

Following visual inspection of ASP parameters produced in clean and noisy speech analyses, as those previously described in figs. 2b-c and 3b-c, the use of ASP techniques was considered and tested for speech segmentation purposes. We compared results obtained segmenting both 'clean' and 'dirty' sentences of "Fred can go, Susan can't go, and Linda is uncertain", using a semi- automatic segmentation tool called "SONOGRAFIA" [7] which is entirely based on Multi-Level (ML) segmentation theory [5]. ASP and FFT parameters were used as input to the segmentation system in order to evaluate and compare their performance aligning speech in clean and adverse conditions.

As previously underlined, figs. 2 and 3 show the ML segmentation tree (the "Dendrogram" [5]) automatically built by the system analysing the 'clean' (fig. 2b) and the 'dirty' sentence (fig. 3b), using ASP parameters as input. In figs. 2 and 3, the ML structure is superimposed to the envelope output only to have a reference, but it is built considering both envelope and synchrony parameters. The same ML structure, but produced using FFT parameters instead of ASP ones, is shown for the 'clean' (a) and 'dirty' (b) case in fig. 4. The final target segmentation is found with minimal human intervention, which is limited exclusively to fixing the vertical point determining the final target segmentation (corresponding to that found on the horizontal line built at this point), and eventually deleting over-segmentation landmarks forced by this choice. Segmentation marks were always automatically positioned by the system and never adjusted by hand. Inspecting figs. 2 and 4, it is evident that a segmentation vertical point is more easily found in fig. 2b, by reference to ASP parameters, than in fig. 4a, using FFT parameters. Moreover no over-segmentation marks were produced when using ASP parameters, while some of them were forced by the use of FFT parameters without regarding the vertical segmentation choice.

Much clearer evidence in favour of the ASP parameters results by inspecting figs. 3b and 4b referring to the segmentation of the 'dirty' sentence. Even if speech is clearly degraded by quite a relevant noise, ASP parameters lead SONOGRAFIA 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 ML segmentation structure. In other words, throughout the examples we examined, over-segmentation marks (gross errors), always produced by the use of FFT parameters, were 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 [9]. 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 complicated ones. At present, no attempts have been made to build such an automatic system; instead, SONOGRAFIA was used, as a very useful

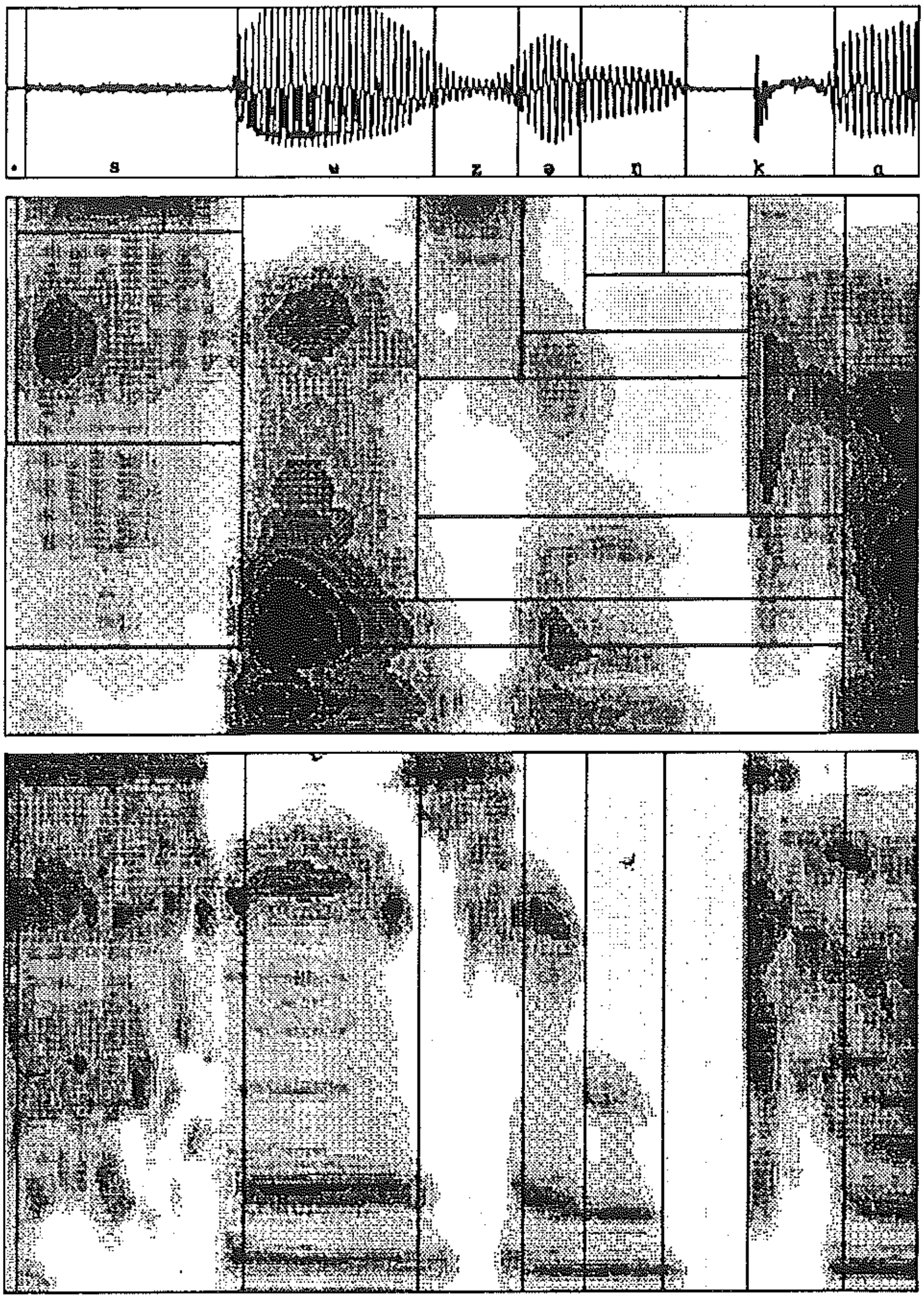


Fig.2 Output of the ASP as applied to 'clean.syl' (a), for the envelope (b) and the synchrony detector (c) modules respectively.
In a) manual segmentation made by an Italian mother-tongue phonetician is superimposed on the waveform. Multiline structure superimposed on the envelope output in b) refers to the "dendrogram", a particular output of the segmentation procedure used (see text): the resulting segmentation is shown in c).

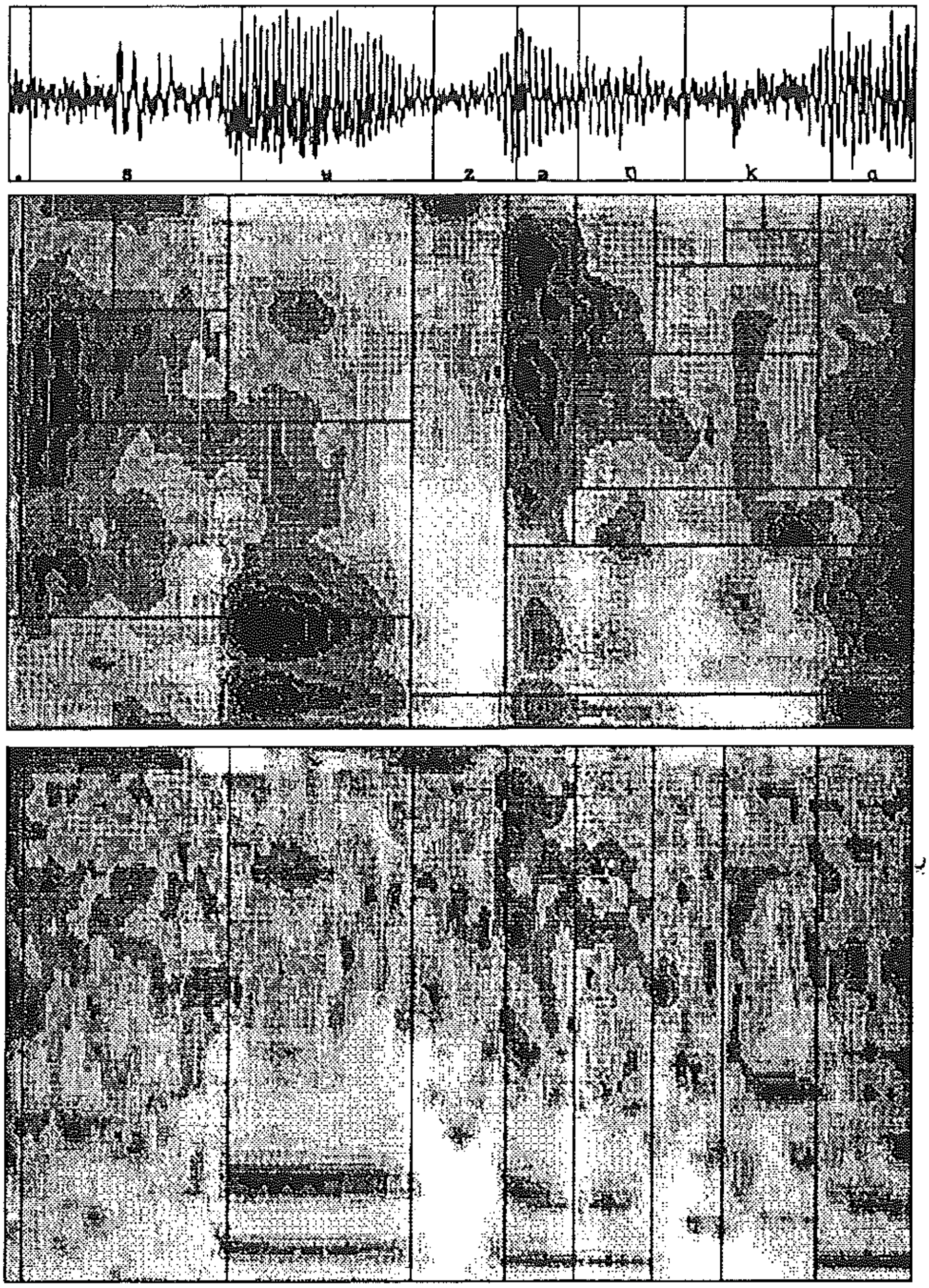


Fig.3 Output of the ASP as applied to 'dirty.syl' (a), for the envelope (b) and the synchrony detector (c) modules respectively.
In a) manual segmentation made by an Italian mother-tongue phonetician is superimposed on the waveform. Multiline structure superimposed on the envelope output in b) refers to the "dendrogram", a particular output of the segmentation procedure used (see text): the resulting segmentation is shown in c).

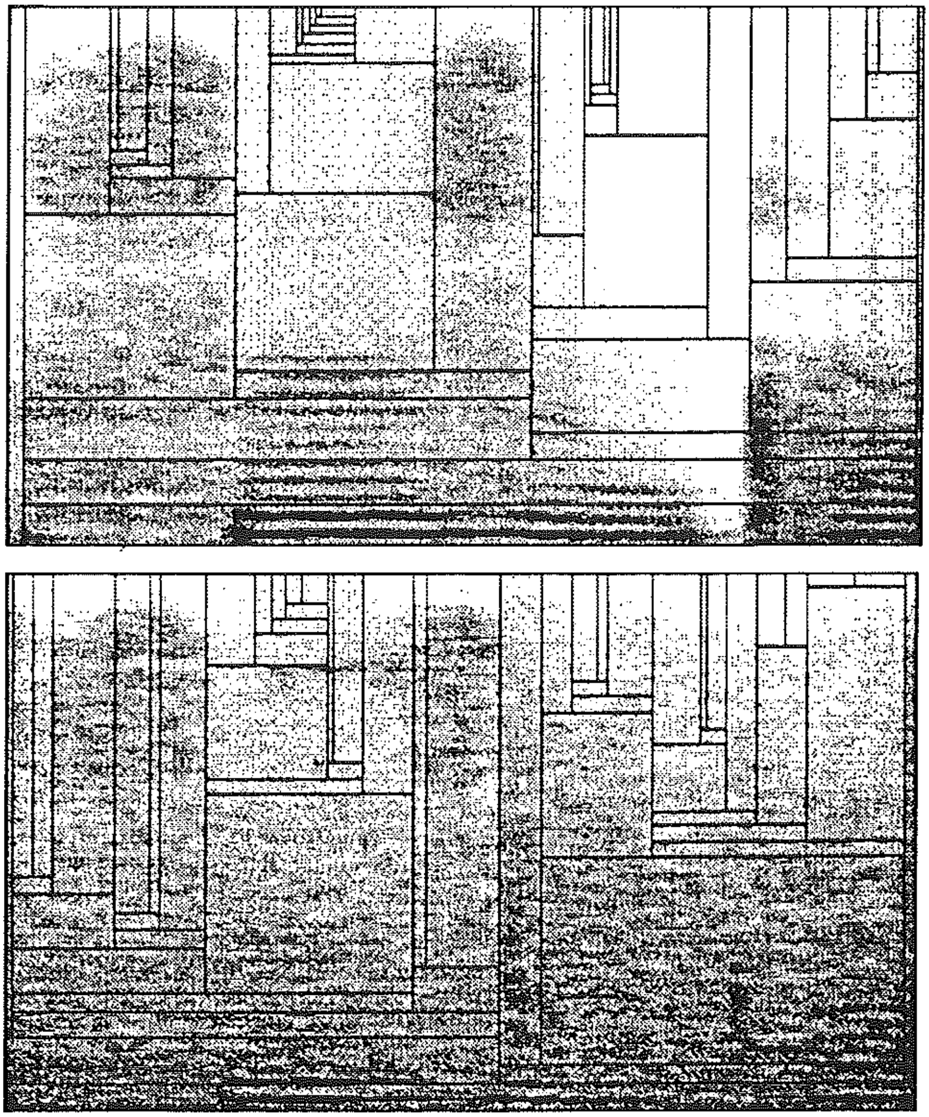


Fig.4 ML structure produced by the segmentation system using FFT instead of ASP parameters; upper: 'clean.syl'; lower: 'dirty.syl'

semi-automatic tool, in order to speed up segmentation procedure and to limit human intervention in fixing segmentation marks.

Finally, speech segmentation discrepancies (fine errors) were computed for both 'clean' and 'dirty' sentences, comparing SONOGRAFIA semi-automatically produced landmarks (test segmentation) with those produced by a manual segmentation (reference correct segmentation) made by a phonetician by using audio and visual facilities (see Appendix A). Figure 5 illustrates segmentation histograms referring to the application of SONOGRAFIA with ASP parameters to both the 'clean' (a) and the 'dirty' (b) sentence . Considering a 20 ms error criterion [4] (i.e. considering an error to be the positioning of a segmentation mark outside a 40 ms interval centred on the correct reference mark) 87% and 90.3% correct segmentation was achieved in the 'clean' and 'dirty' case respectively.

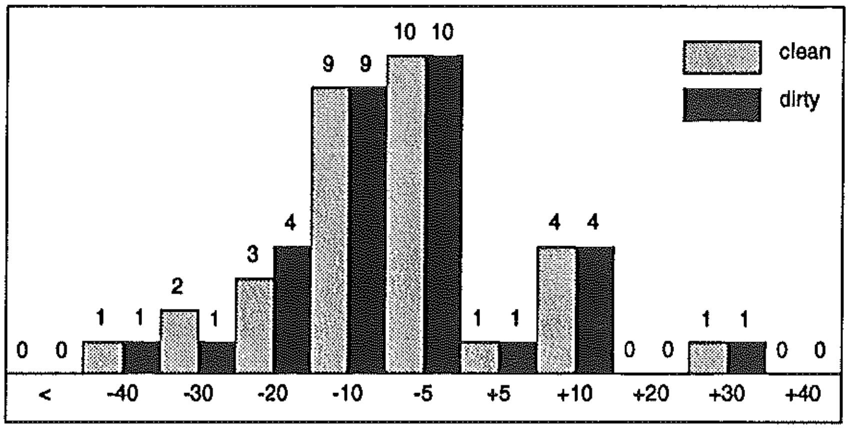


Fig. 5 Segmentation histogram referring to the application of SONOGRAFIA, with ASP parameters, to 'clean' and 'dirty'

5 CONCLUSIONS

Various results suggest the effectiveness of ASP techniques for speech analysis and recognition. As for segmentation, considering both gross-errors (over-segmentation marks) and fine-errors (segmentation discrepancies) ASP parameters seem to constitute a very effective, hopefully better, alternative to classical speech parameters. In order to verify and make reliable the results presented in this paper (surprisingly higher performance in the noisy speech conditions than in the clean ones) much more experiments need to be performed, but in the mean time these preliminary results could be considered as a very promising starting point for further research.

REFERENCES

[1] P. Cosi, Y. Bengio & R. De Mori (1990), 'Phonetically-based multi- layered neural networks for vowel classification', *Speech Comm.*, 9(1), 15-29.

[2] P. Cosi, R. De Mori & K. Vagges (1990), 'A neural network architecture for italian vowel recognition', *Proc. VERBA '90*, Rome, 221-228.

[3] P. Cosi, L. Dellana, G.A. Mian & M. Omologo (1991), 'Auditory model implementation on a DSP32C board', *Proc. GRETSI '91*, Juan-les-Pins.

[4] P. Cosi, D. Falavigna & M. Omologo (1991), 'A preliminary statistical evaluation of manual and automatic segmentation discrepancies', *Proc. EUROSPEECH '91*, Genova, 693-696.

[5] J.R. Glass & V.W. Zue (1988), 'Multi-level acoustic segmentation of continuous speech', *Proc. ICASSP '88*, 429-432.

[6] M.J. Hunt & C. Lefebvre (1988), 'Speaker dependent and independent speech recognition experiments with an auditory model', *Proc. ICASSP '88*, 215-218.

[7] A. Marzal & J. Puchol (1991), 'Sonografia: an interactive segmentation system of acoustic signals based on multilevel segmentation for a personal computer', *ESPRIT-II BRA-ACCOR Progress Reports*, 2.

[8] S. Seneff (1988), 'A Joint synchrony/mean-rate model of auditory speech processing', J. Phonetics, 16, 55-76.

[9] V.W. Zue, J. Glass, M. Philips & S. Seneff (1989), 'Acoustic segmentation and phonetic classification in the SUMMIT system', Proc. ICASSP '89, S8.1, 389-392.

APPENDIX A

Manual segmentation produced by an Italian mother-tongue phonetician of 'clean.wav': "Fred can go, Susan can't, and Linda is uncertain" (SAMPA alphabet):

0, ...	17300, s	29689, 6 (/6U/)	41576, s
4296, f (+/u/?)	19356, }	32225, 1	43271, @ (/@:/)
5448, @	22411, z	33639, I (/I~/)	46033, ? k
7246, (/@~/)	22053, @	34868, n	47978, n
9596, e (/e~/)	22763, N	36051, d	50621, ...
10538, N	23814, k	36878, @	57280 <END>
11791, g	25207, A (/A~/)	38030, s (/zx/)	
12612, A	27841,	N 39237, V (/V~/)	
15034, e(/Ae/)	28934, k	40423, n	