D, DD, DDD, DDDD……
EVIDENCE AGAINST FRAME-BASED ANALYSIS TECHNIQUES

Piero Cosi

Istituto di Fonetica e Dialettologia – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY),
e-mail: cosi@csrf.pd.cnr.it www: http://www.csrf.pd.cnr.it

ABSTRACT

The need of Δ, ΔΔ, ΔΔΔ, ΔΔΔΔ.... measures is a clear sign of the loss in the representation capability of classical frame-based analysis techniques. Mainly coarticulation effects in fluent speech are hidden and obscured by the classical short-time analysis technique. In fact, almost every acceptable ASR system is forced to introduce this kind of post-processing technique, in order to obviate to that loss.

Following previous work on Auditory Modeling (AM) techniques for speech analysis front-end for automatic speech segmentation (ASS) and automatic speech recognition (ASR), evidences against frame-based analysis techniques, thus against the need of Δ, ΔΔ, will be given and exploited in this paper.

Various examples, mostly on plosives or other non-stationary consonants, will be illustrated, with the aim of underlying the superiority of "*sampling after processing*" relatively to "*framing before processing*" in speech analysis tasks.

1. INTRODUCTION

Most speech processing schemes relay on the underlying assumption that the properties of the speech signal change relatively slowly with time. From this assumption, various *short-time* processing methods were suggested [1], all working on short segments of speech, often overlapped, isolated and processed as if they were derived from a sustained sound with fixed properties.

Unfortunately, continuous speech is characterized by sounds described by time-varying spectral patterns and, a part from few rare cases, most of which are slowly uttered vowels, no steady-state period is present is fluent speech.

The introduction of Δ parameters, and for automatic speech recognition (ASR) applications the cepstral domain is often considered [2], was exactly intended to provide information characterizing dynamic changes within a speech sound and transitions from one speech sound to another. This particular post-processing technique was, in fact, designed with the aim of overcoming the loss in the representation capability of classical frame-based analysis techniques and was incredibly effective in boosting performance of ASR systems comparing to that previously obtained [3-4].

In the real world, as a consequence of coarticulation, the spectral pattern of each phoneme is highly modified and profoundly transformed by its adjacent phonemes. What is called neutralization or reduction in phonetics, in other words the fact that the ideal or target spectrum of a phoneme is almost never realized in continuous speech, is essentially due to the coarticulation mechanism. Due to all these facts, it is emerging in the speech community that even the Δ-strategy is insufficient to solve the problem of a true description of the continuous speech signal.

Evidences in favor of dynamic spectral features as well as instantaneous spectral features were found and nowadays it is commonly believed that these features play an important role in human speech perception [5]. Furui demonstrates, by using isolated syllables truncated at the initial or final end, that the portion of the utterance where spectral variation was locally maximum contained the most important phonetic information in the syllable [6].

Despite the physical modification of phonemes, depending on the context in which they occur, our perception mechanism looks at continuous speech as a process of concatenation of "constant" phonemes, regardless of their evident modification. In other words, a so-called categorical perception mechanism should exist in order to overcome the target-undershooting problem. Various human hearing mechanisms, such as the context-independent spectral compensation or prediction in which, by overshooting the spectral dynamics [7-8], the target spectrum is perceived, or the context-dependent hearing mechanism of the contrast effect, were hypothesized, in the past, in order to explain this perceptual behavior.

A computational model for this perception mechanism originally proposed by Kuwabara [9], was followed by the important work of Furui [10], which inspired various researchers in developing new ASR systems [11-13] adopting new signal processing schemes which directly incorporate these new perceptual findings based on emphasizing the spectral dynamics.

The main idea inspiring this work was that of adopting an auditory-based model of speech processing which does not make use of the *short-term* assumption, and transforms the speech signal in a more "smooth" process. The final goal will be that of justifying the hypothesized superiority of "*sampling after processing*" relatively to "*framing before processing*" in speech analysis tasks. In other words, we asked ourselves the following question:

> "*why framing speech and trying to reconstruct the information we loose doing that with perceptual based strategy, instead of develop and use auditory based front-ends that analyze speech without the need of framing it and smooth the parameter vector in such a way that is much more effective to sample it at the end of the whole process*?".

Looking at our previous works on Auditory Modeling (AM) techniques for speech analysis front-end for automatic speech segmentation (ASS) [14-15] and ASR [16-18], some answers to the above question will be suggested. Some evidences against frame-based analysis techniques, thus against the need of Δ-strategy or of the more refined spectral-dynamics-strategy, will be given.

2. AUDITORY MODEL

Our work is focused on the Seneff's model of auditory speech processing (ASP) [19], but other auditory models could be considered if they satisfy the two requests underlined in the introduction, that is: *no-framing analysis* and *smoothing output*.

The reason why this model was adopted in this work both, in segmentation and in recognition tasks, is well underlined by Seneff in [19]:

> "*The parameters of the model were adjusted to match existing experimental results of the physiology of the auditory periphery. The output of this model is delivered to two parallel channels, each of which produces spectral representations appropriate for distinct subtasks of a speech recognition system. One path yields an overall energy measure for each channel that can be identified with the average rate on neural discharge. The outputs of this path appear to be useful for locating acoustic events and assigning segments to broad phonetic categories. In the other path, the extent of dominance of periodicities at each channel's center frequency is captured by a synchrony measure, which yields a spectral representation with enhanced spectral contrast, relative to the mean-rate spectrogram. The outputs of this stage show distinct formant peaks during sonorant regions, with smooth transitions over time, as well as preserving spectral prominences in the high-frequency region for fricatives and stops*."

As you can see in Figure 1, referring to the Italian word /'panka/ (bench), uttered in isolation by an Italian male talker, the output of the Seneff's model is quite effective in smoothly tracking the dynamic modifications of speech. Transitions from one phonetic segment to the next are clearly delineated by onsets and offsets in the output representation given by the auditory spectrogram, and this is probably due to the forward masking mechanism which is directly included in the model.

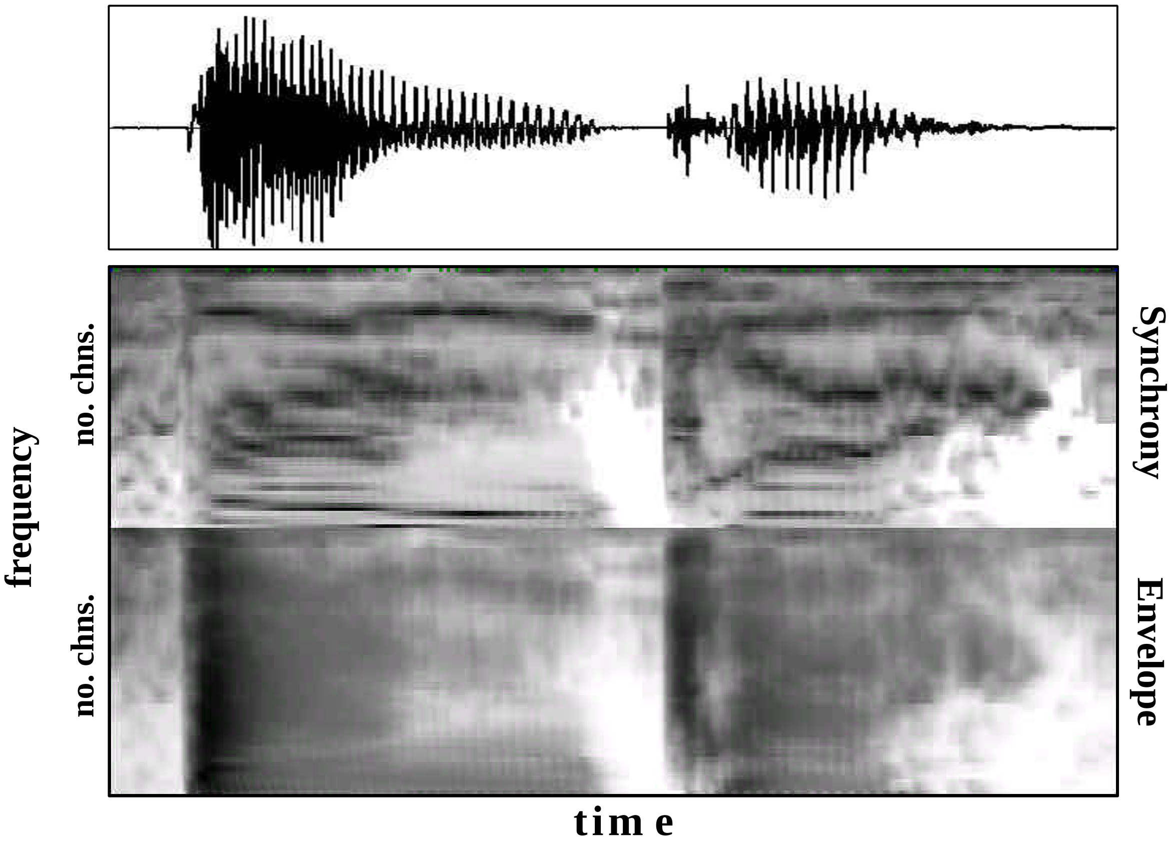

Figure 1. Output of the Seneff's auditory model speech processing applied to the Italian words /'panka/ (bench). Both envelope and synchrony parameters [19] are considered in the same spectrogram plot.

3. SEGMENTATION

As previously underlined, the Seneff's ASP appears to be more suitable than other classical frame-based analysis techniques for locating acoustic events, thus it constitutes a good starting point to build an effective segmentation system. Following the work of Glass [20] on the so-called Multi-Level Segmentation[1] (MLS) theory [21] a PC-based segmentation system[2] called SLAM (Semi Automatic Segmentation Module) was designed and implemented [14-15].

SLAM makes use of the MLS hierarchical technique, that, incorporating some kind of temporal constraints is quite useful to appropriately rank the significance of acoustic events. By a recursive technique, involving the computation of Euclidean-based similarity measure for each target frame, some initial adjacent "seed regions", which constitute the basis for the MLS "hierarchical structuring" segmentation procedure, are created. These regions, using the same similarity measure, are themselves merged together, and, by keeping track of the distance at which two regions merge into one, a multi-level structure describing the hypothesized segmentation landmarks, usually called *dendogram* [21], is built up.

The effectiveness of the combination of the MLS strategy with the auditory modeling versus other frame-based analysis techniques such as FFT and LPC could be verified in Figure 1 where two segmentation examples referring to two Italian syllables /'ba/ and /'ka/ are illustrated. Looking at the dendograms with the hypothesized segmentation landmarks, superimposed to every spectrogram plots in the figure, it is quite evident that with the Seneff's ASP the final correct segmentation is much more easily identified.

Along the line followed for segmenting and labeling American English speech material by the SUMMIT system [22], various Italian speech data were semi-automatically segmented and labeled with SLAM obtaining similar accuracy than that obtained by manual labeling by expert phoneticians [23].

4. RECOGNITION

Various studies suggest the effectiveness of auditory-based speech processing techniques for speech recognition [24], especially in adverse conditions [25].

For example, the results of the application of Seneff's ASP in vowel classification experiments were compared with those obtained with a classical front-end built with an FFT-based filter-bank. A subset of the American-English vowels built up with the 10 vowels /i, ɪ, ɛ, æ, ʌ, ə, ɑ, ɔ, ʊ, u/ were extracted from the words BEEP, PIT, BED, BAT, BUT, FUR, FAR, SAW, PUT, BOOT.

With a neural network-based system, a 96% correct classification performance was obtained with the auditory-based front-end while with a classical FFT-based front-end an 87% correct classification performance was achieved [16]. Similar results were obtained for the complete set of Italian vowels /i, e, ɛ, a, ɔ, o, u/ which were extracted from the words PIPA, PEPE, PEPPA, PAPA, POPE, POPPA, PUPA.

Furthermore, other results on Italian phoneme recognition experiments [17-18] provided other evidences in favor of the idea that the proposed ASP could perform better than other acoustic production-based front-end in speech classification tasks.

5. SUMMARY

Various evidences suggest the effectiveness of the application of auditory speech processing techniques for speech analysis.

As for segmentation, considering both gross-errors (over-segmentation) and fine-errors (segmentation discrepancies) ASP parameters seem to constitute a very effective tool and a better alternative to other classical frame-based analysis parameters.

As for recognition, much more experiments need to be performed to convince ASR people to adopt this kind of processing, but an increased interest on this matter seems to be already activated by preliminary results on comparing these new techniques against classical ones especially in noisy conditions.

Moreover, the well-known objection to auditory-based front-ends, that is the too high computational cost of such processing techniques, will be easily overcome in the future, due to the tremendous increase in speed and capacity characteristics, of new designed computer processing chips.

6. ACKNOWLEDGMENTS

This work has been made possible exclusively thanks to S. Seneff and J.R. Glass who gave me important suggestions for implementing the joint Synchrony/Mean-Rate (S/M-R) model of Auditory Speech Processing [19], and for developing the MLS segmentation strategy [20-21].

[1] Within the framework of MLS theory [21], 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. Following this viewpoint, the segmentation problem can be simply reduced to a local clustering problem where the decision to be taken regards the similarity of any particular frame with the signal immediately preceding or following it. Using only relative measures of acoustic similarity, this technique should be quite independent of the speaker, vocabulary, and background noise.

[2] SLAM version 1.0 works on Windows 3.1, 3.11, 95 and NT and is available at the following ftp site: www.csrf.pd.cnr.it

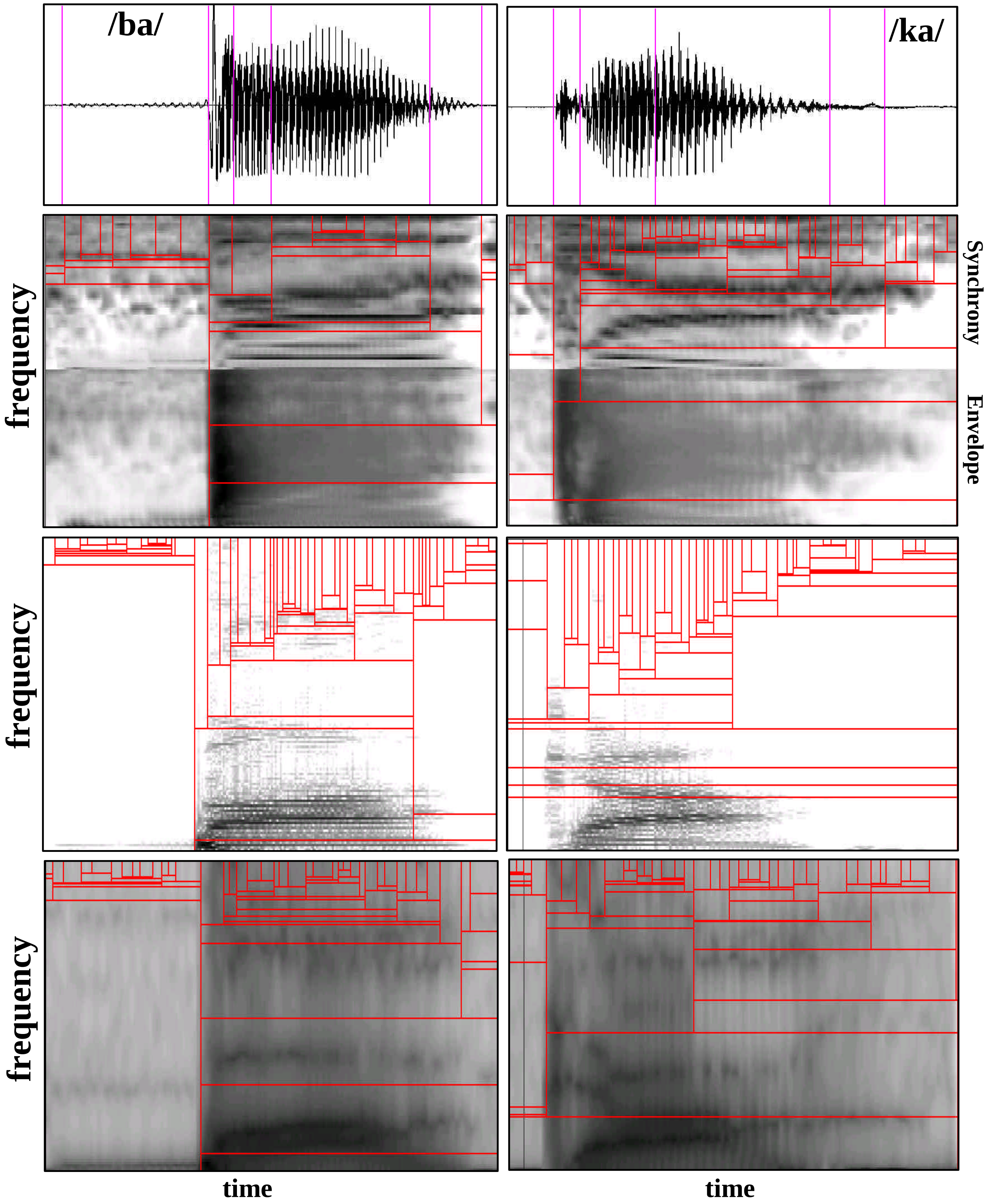

Figure 2. Example of segmentation of the Italian syllables /'ba/ and /'ka/ with different auditory and spectral representations. For each syllable, the Seneff's AM, the narrow-band FFT, and the LPC-derived spectrograms are illustrated below the waveform from top to bottom. It is quite evident that in the first case (AM) the segmentation landmarks are much more easily identified from the dendogram of hyphoteses produced by SLAM (see text).

7. REFERENCES

[1] Rabiner L.R. and Shafer R.W. (1978), *Digital Processing of Speech Signal*, Prentice Hall,, Alan V. Oppenheim Series Editor, Englewood Cliffs, New Jersey.

[2] Davis S.B. and Mermelstein P., "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Volume ASSP-28, no. 4, 1980, pp. 357-366.

[3] Rabiner L.R., Wilpon J.G. and Soong F.K., "High Performance Connected Digit Recognition, Using Hidden Markov Models", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-88*, New York, Apr 11-14, 1988, pp. 119-122.

[4] Lee K.F., *Automatic Speech Recognition: The Development of the SPHINX System*, Kuwler Academic Publisher Boston, 1989.

[5] Ruske G., "Auditory Perception and its Application to Computer Analysis of Speech", in *Computer Analysis and Perception, Volume II, Auditory Signals*, Suen C.Y. and De Mori R. Editors, CRC Press, 1982, pp.2-42.

[6] Furui S., "On the Role of Dynamic Characteristics of Speech Spectra for Syllable Perception", *Proc. of Fall Meeting of Acoustical Society of Japan*, 1-1-2, Oct 1984.

[7] Brady P.T., House A.S. and Stevens K.N., "Perception of Sounds Characterized by a Rapid Changing resonant Frequency", *Journal of Acoustical Society of America, JASA*, vol. 33, 1961, pp. 1357-1362.

[8] Lindblom B.E.F and Studdert-Kennedy M, "On the role of Formant Transition in Vowel Recognition", *Journal of Acoustical Society of America, JASA*, vol. 42, 1967, pp. 830-843.

[9] Kuwabara H., "An Approach to Normalization of coarticulation effects for vowels in connected speech", *Journal of Acoustical Society of America, JASA*, vol. 77, 1985, pp. 686-694.

[10] Furui S., "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum, *IEEE Trans. on Acoustics, Speech and Signal Processing*, Volume ASSP-34, no. 1, 1986, pp. 52-59.

[11] Furui S., "Speaker Independent Isolated Word Recognition Based on Emphasized Spectral Dynamics", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-86*, Tokyo, Japan, pp. 1991-1994.

[12] Doddington G.R., "Phonetically Sensitive Discriminants for Improved Speech Recognition", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-89*, Glasgow, Scotland, May 23-26, 1989, pp. 556-559.

[13] Wilpon J.G., Lee C.H. and Rabiner L.R., "Improvements in Connected Digit Recognition Using Higher Order Spectral and Energy Features", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-91*, Toronto, Ontario (canada) 1991, pp. 349-352.

[14] Cosi P., "SLAM: Segmentation and Labeling Automatic Module", *in Proc. of 3rd European Conference on Speech Technology, EUROSPEECH-93*, Berlin, Germany, 21-23 September, Vol. 1, 1993, pp. 88-91.

[15] Cosi P., "SLAM v1.0 for Windows: a Simple PC-Based Tool for Segmentation and Labeling", *Proc. of International Conference on Signal Processing Applications & Technology, ICSPAT-97*, San Diego, CA, USA, September 14-17, 1997, pp. 1714-1718.

[16] Cosi P., Bengio Y. and De Mori R., "Phonetically-Based Multi-Layered Neural Networks for Vowel Classification", *Speech Communication*, North Holland, Vol. 9, No. 2, 1990, pp. 15-29.

[17] Cosi P., Magno Caldognetto E., Ferrero F.E., Dugatto M. and Vagges K., "Speaker Independent Bimodal Phonetic Recognition Experiments", *Proc. of ICSLP-1996*, Philadelphia, PA USA, October 3-6, 1996, Vol. 1, pp. 54-57.

[18] Cosi P., Dugatto M., Ferrero F.E., Magno Caldognetto E., and Vagges K., "Phonetic Recognition by Recurrent Neural Networks Working on Audio and Visual Information", *Speech Communication*, North Holland, Vol. 19, No. 3, 1996, pp. 245-252.

[19] Seneff S., "A joint synchrony/mean-rate model of auditory speech processing", *Journal of Phonetics*, Vol. 16(1), Jan 1988, pp. 55-76.

[20] Glass J.R., "Finding Acoustic Regularities in Speech: Application to Phonetic Recognition", *Ph.D Thesis*, May 1988, MIT press.

[21] Glass J.R. and Zue V.W, "Multi-Level Acoustic Segmentation of Continuous Speech", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-88*, New York, Apr 11-14, 1988, pp. 429-432.

[22] V.W. Zue, J. Glass, M. Philips and S. Seneff, "Acoustic Segmentation and Phonetic Classification in the SUMMIT System", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-89*, Glasgow, Scotland, May 23-26, 1989, pp. 389-392.

[23] Cosi P., Falavigna D. and Omologo M., "A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies", *Proc. of 2ndEuropean Conference on Speech Technology, EUROSPEECH-91*, Genova, Sep 24-26, 1991, pp. 693-696.

[24] Hunt M.J. and Lefebvre C., "Speaker Dependent and Independent Speech Recognition Experiments with an Auditory Model", *Proc. of IEEE International Conference on Acoustic Speech and Signal Processing, ICASSP-88*, New York, Apr 11-14, 1988, pp. 215-218.

[25] Jankowski C.R. Jr., Vo H-D. H.& Lippmann R.P., "A Comparison of Signal Processing Front Ends for Automatic Word Recognition", *IEEE Trans .on Speech and Audio Processing*, Volume SAP-3, N. 4, Jul, 1995, pp. 286-293.