# ON AMPLITUDE MODULATION FOR MONAURAL SPEECH SEGREGATION

Guoning Hu
Biophysics Program,
The Ohio State University
hu.117@osu.edu

DeLiang Wang
Department of Computer and Information Science & Center of
Cognitive Science
The Ohio State University
dwang@cis.ohio-state.edu

**Abstract – We propose a computational auditory scene analysis (CASA) model for monaural speech segregation. It deals with low-frequency and high-frequency signals differently. For high-frequency signals, it generates segments based on common amplitude modulation (AM) and groups them according to AM repetition rates. This model performs substantially better than previous CASA systems.**

## I. INTRODUCTION

In the real-world environment, target speech usually occurs simultaneously with acoustic interference. An effective speech segregation system will greatly facilitate many applications, including automatic speech recognition (ASR) and speaker identification. Many systems have been proposed to deal with speech segregation, primarily using blind source separation (BSS) [1] or speech enhancement techniques [2]. BSS performs well when there are enough sensors and the mixing signals satisfy some statistical independence. However, BSS techniques require at least two sensors, while many applications such as telecommunication and audio retrieval need a monaural (one sensor) solution. Speech enhancement techniques perform well in certain environments where some prior knowledge about target/interference is available. However, no system can efficiently separate speech from a variety of acoustic intrusions with one sensor.

While monaural segregation remains a difficult challenge for computational systems, the auditory system shows an impressive capacity for monaural segregation. ASA is the perceptual process in which an acoustic mixture is analyzed and separated into streams, corresponding to the acoustic sources [3]. Considerable research has been carried out to build monaural CASA systems [4-7]. Almost all existing systems rely on periodicity as a main grouping cue. However, the performance of these systems is limited and progress has stagnated in recent years. A main problem with the current systems is that they lack the ability to deal with high-frequency signals.

We study monaural speech segregation with particular emphasis on the high-frequency problem. For voiced signal, we note that the auditory system can resolve the first few harmonics in the low-frequency range but higher harmonics are unresolved unless they are much more intense than adjacent ones [8]. Psychoacoustic evidence suggests that different mechanisms are used to deal with resolved and unresolved harmonics [9]. Consequently, our model employs different methods to segregate target speech in the low-frequency range and in the high-frequency range.

According to Bregman [3], ASA takes place in two stages: segmentation (or analysis) and grouping. In segmentation, the acoustic input is decomposed into sensory segments, each of which would belong to one source. In grouping, those segments that are likely to respond to the same source are grouped together. Inspired by this suggestion, our model performs segregation in two corresponding stages across all frequency channels. More specifically, for low-frequency channels, our model generates segments based on temporal continuity and cross-channel correlation between responses from nearby channels. These segments are grouped by comparing periodicities of these responses with the estimated pitch of the target speech. On the other hand, high-frequency channels due to the wide bandwidths tend to respond to multiple harmonics of voiced speech, which are usually unresolved. These high-frequency responses are amplitude modulated, and their envelopes fluctuate at the frequency corresponding to the fundamental frequency (F0) [10]. Our model generates segments in the high-frequency range based on common AM and temporal continuity. These segments are grouped by comparing AM repetition rates with estimated F0 of the target speech.

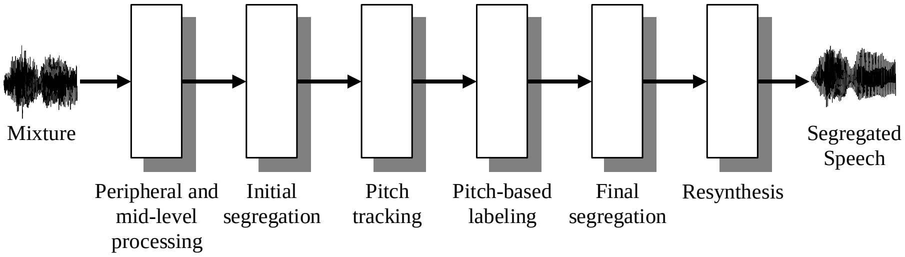

Fig. 1. The schematic diagram of the proposed system.

0-7803-7278-6/02/$10.00 ©2002 IEEE

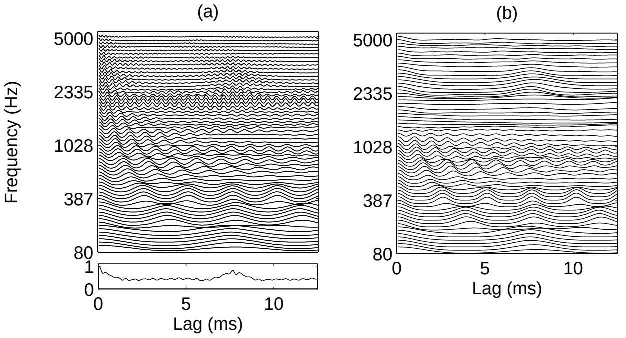

Fig. 2. a) The correlogram of individual responses for a mixture of a voiced utterance and the "cocktail party" noise at time frame 45 (i.e. 0.45 second after the start of the stimulus). For clarity, only half of all the channels are shown. The summary correlogram is shown in the bottom panel. b) The correlogram of response envelops at the same frame for the same input.

Section 2 describes the overall system. In section 3, systematic results and a comparison with an existing CASA model are given. Section 4 concludes the paper.

## II. MODEL DESCRIPTION

Our model is a multistage system, as shown in Fig. 1. Description for each stage is given below.

### A. Peripheral and Mid-level Processing

First, an acoustic input is analyzed by a peripheral model comprising cochlear filtering with a bank of 128 gammatone filters and subsequent hair cell transduction. This peripheral processing is done in time frames of $20\,ms$ long and $10\,ms$ overlap between consecutive ones. As a result, the input signal is decomposed into a group of cells. Each time-frequency cell contains the response of a certain channel in a certain frame. The envelope of the response is obtained by a lowpass filter with passband [0, 1 *k*Hz] and a Kaiser window of $18.25\,ms$. Mid-level processing is performed by computing a correlogram (autocorrelation function) of the individual responses and their envelopes. The global pitch contour is obtained from the summary correlogram. (See [7] for more details.)

As an example, Fig. 2 shows the correlogram of the responses and their envelopes at a particular frame and the corresponding summary correlogram. The input is a voiced utterance mixed with the "cocktail party" noise.

### B. Initial Segregation

Initial segregation takes place in two steps. First, segments are formed by grouping neighboring time-frequency cells based on temporal continuity and cross-channel correlation. In general, segments correspond to resolved components of the input signal, and most of them lie in the low-frequency range. Then, according to global pitch, segments are grouped into a foreground stream, which corresponds to the target speech, and a background stream, which corresponds to the intrusion. This process is same as the segregation process implemented in the Wang-Brown model through a two-layer oscillatory neural network [7], which provides a good basis for accurate pitch estimation.

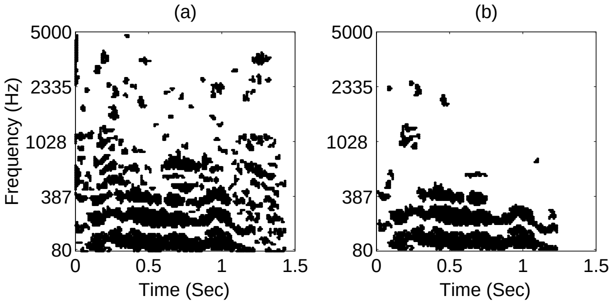

Fig. 3. (a) The segments generated in initial segregation. (b) The foreground stream formed in initial segregation. The input is the mixture of speech and "cocktail party" noise.

As an example, Fig. 3 shows the segments and the foreground stream generated in initial segregation. The input is the mixture of speech and "cocktail party" noise.

### C. Pitch Tracking

The target pitch contour is obtained using the same method we described previously in [11]. First, it is estimated from the foreground stream. Then the estimated pitch is checked according to two psychoacoustically-inspired constraints: 1) An accurate pitch period should be consistent with the periodicity of responses in the channels where the target speech dominates; 2) Pitch periods should vary smoothly in time. Unreliable pitch periods are replaced by new values obtained based on temporal continuity. (See [11] for more details.)

### D. Pitch-based Labeling

Each cell is labeled according to whether target speech dominates the corresponding response or not. To label cells, we can compare the periodicities of the corresponding responses with the estimated target pitch in time domain. More specifically, a cell of channel $i$ and frame $j$ is labeled as target speech dominant if

$$A(i, j, \tau(j)) / A(i, j, \tau_m) > \theta_d , \tag{1}$$

where $\tau(j)$ is the estimated pitch period in frame $j$, $A(i, j, \tau)$ is the autocorrelation function of the corresponding response, $\tau_m$ is the lag corresponding to the maximum of $A(i, j, \tau)$ for $\tau \in [2\,ms,\ 12.5\,ms]$, $\theta_d$ is the threshold. Here, we let $\theta_d = 0.85$.

This criterion, referred as the time criterion, works well in the low-frequency range where harmonics are resolved and therefore the pitch period of target speech corresponds to the global maximum of the autocorrelation function, as

0-7803-7278-6/02/$10.00 ©2002 IEEE

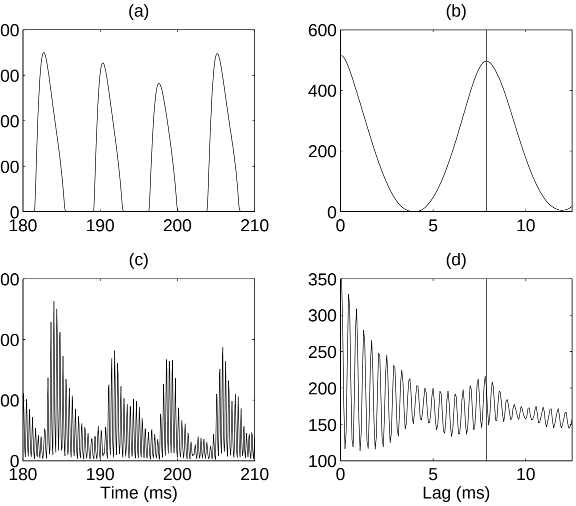

Fig. 4. (a) The response from the auditory filter whose center frequency is 140 Hz. (b) The corresponding autocorrelation function of the response in (a). (c) The response from the auditory filter whose center frequency is 2.6 *k*Hz. (d) The corresponding autocorrelation function of the response in (c). The input is the clean speech. The vertical lines in (b) and (d) mark the position of the lag corresponding to the pitch period.

shown in Fig. 4(a) and 4(b). However, it is not suitable for high-frequency channels because their responses are likely to contain multiple harmonics and therefore are amplitude modulated. As shown in Fig. 4(c) and 4(d), for a strongly amplitude-modulated response, the pitch period of target speech corresponds to a local maximum in the autocorrelation function instead of the global maximum. In addition, the peaks of the correlogram are steep, which makes this criterion less robust.

Here we propose a criterion for labeling cells with amplitude-modulated responses based on the following observation: for high-frequency responses where speech dominates, response envelopes fluctuate at the rate of F0 [10]. The new criterion compares AM repetition rate with estimated F0 at every sample, which is obtained by interpolating estimated pitch periods of target speech.

The AM repetition rate is obtained as follows. First, for each channel, the output from the corresponding gammatone filter is half-wave rectified and then bandpass filtered to remove DC component and other possible harmonics except the F0 component. Here we use a filter with passband [0.9 $\bar{f}$ , 1.2 $\bar{f}$ ] and a Kaiser window of $50\,ms \sim 100\,ms$ for response in every $100\,ms$ period. $\bar{f}$ is the average of estimated F0 in every $100\,ms$ period, and it determines the size of the corresponding Kaiser window. The instantaneous frequency (IF) of the rectified and filtered signal, obtained through a linear prediction algorithm in the spectral domain [12], indicates the AM repetition rate of the corresponding response.

As an example, Fig. 5(a) shows the output from a gammatone filter with center frequency 2.6 *k*Hz in several frames when the input is the clean speech. Fig. 5(b) shows the output from the same filter when the input is “cocktail party” noise. Fig. 5(c) shows the rectified and filtered output from the same filter when the input is the mixture of speech and “cocktail party” noise. Fig. 5(d) shows the obtained AM repetition rate. It matches the estimated F0 very well when the target speech signal dominates, and does not match the estimated F0 when the intrusion signal dominates.

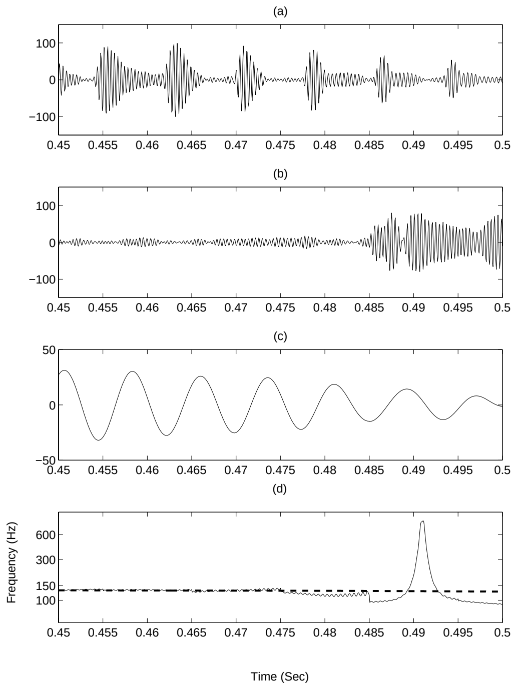

Fig. 5. (a) The output from a gammatone filter with center frequency 2.6 *k*Hz. The input is the clean speech. (b) The output from the same filter when the input is “cocktail party” noise. (c) The rectified and filtered output from the same filter. The input is the mixture of the speech and “cocktail party” noise. (d) Solid line: the AM repetition rate. Dash line: the estimated F0.

To measure the difference between the estimated F0 and the AM repetition rate for each cell, we calculate the root mean square of the difference between their logarithms. That is,

$$D(i,\,j)=\sqrt{\frac{1}{M}\sum_{k=0}^{M-1}[\log f_0(jT-k)-\log f_I(i,\,jT-k)]^2}\ , \tag{2}$$

where $f_0(t)$ is the estimated F0, $f_I(i,t)$ is the obtained AM repetition rate for channel $i$ . $M$ spans $20\,ms$, and $T=10\,ms$ . The smaller $D(i,\,j)$ is, the more likely it is for target speech to dominate the response of the corresponding cell. A cell of channel $i$ and frame $j$ is labeled as target speech dominant if:

$$D(i,\,j)<\theta_f\,, \tag{3}$$

where $\theta_f$ is the threshold. We refer this criterion as the frequency criterion.

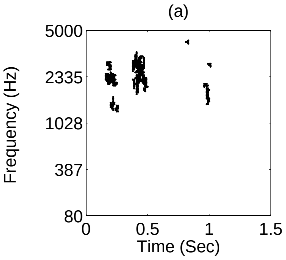

Fig. 6. The segments generated in final segregation. The input is the mixture of speech and "cocktail party" noise.

Psychoacoustic evidence shows that listeners can discriminate two simultaneous sounds with unresolved harmonics if the difference in F0 is more than 10% [9]. When there is a stable 10% difference between F0 and AM repetition rate, the corresponding measurement of their difference, $D(i, j)$, is about 0.1. However, consider that usually the difference between F0 and AM repetition rate are randomly distributed and F0 cannot be perfectly estimated, we choose $\theta_f$ to be 0.15, a looser threshold.

### E. Final Segregation

First, new segments are generated based on temporal continuity and common AM for cells that satisfy the frequency criterion. In this process, only the cells that are neither in the foreground stream nor in the background stream are included for the following considerations. There should be no conflict between this segmentation process and the one in initial segregation. Furthermore, the segments generated in initial segregation tend to reflect resolved components, and therefore shall be retained. The similarity of AM between the responses of nearby cells is measured by the cross-channel correlation of response envelopes. Segments are formed by grouping neighboring cells satisfying the above criteria. Most of them are in the high-frequency range. As an example, Fig. 6 shows the segments generated from the mixture of speech and "cocktail party" noise.

Then these segments are grouped into the foreground stream. Other segments in the foreground stream, which are grouped into this stream in initial segregation, are separated so that all the cells in one segment either satisfy or violate the time criterion. Some segments are removed from the foreground stream as a result, and they are put into the background stream if they contain cells violating the time criterion only.

Other cells that do not belong to either stream are grouped according to temporal and spectral continuity. More specifically, first, the background stream expands iteratively by grouping neighboring cells violating the time criterion or frequency criterion. It keeps on expanding until no more cells can be added. Then the foreground stream expands by grouping neighboring cells satisfying the time criterion or frequency criterion iteratively.

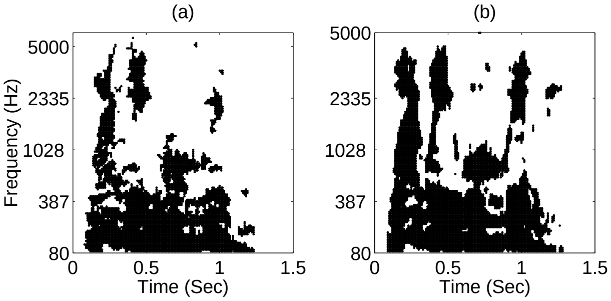

Fig. 7. (a) The final segregated target speech stream. (b) The ideal stream. The input signal is the mixture of speech and "cocktail party" noise.

### F. Resynthesis

Segregated target speech is resynthesized from the foreground stream. In resynthesis, the foreground stream works as a binary mask. It retains the signals corresponding to the cells in foreground stream, and removes other signals from the mixture[5].

## III. RESULTS AND COMPARISON

Our model is evaluated with a corpus of 100 mixtures composed of 10 voiced utterances mixed with 10 intrusions collected by Cooke [4]. The speech waveform resynthesized from the segregated speech stream is used for evaluation. For every mixture, the speech waveform resynthesized from the ideal stream composed of all the cells where target speech dominates, is used as the ground truth of target speech [11]. Theoretically speaking, the ideal stream gives the ceiling of performance for all binary masks. This evaluation methodology is supported by the following observations. In a critical band, a weak signal is masked by a stronger one [8]. In addition, the ideal stream is similar to the prior mask used in a recent study for ASR [13], which yields excellent recognition performance.

As an example, Fig. 7(a) shows the speech stream segregated from the mixture of speech and "cocktail party" noise. Fig. 7(b) shows the corresponding ideal stream of the mixture.

Let $S(t)$ be the resynthesized waveform by our model, $I(t)$ the waveform from the ideal stream, $e_1(t)$ the signal present in $I(t)$ but missing from $S(t)$, and $e_2(t)$ the signal present in $S(t)$ but missing from $I(t)$. We measure the ratio of energy loss, $R_{EL}$, and the ratio of noise residue $R_{NR}$ as follows:

$$R_{EL} = \sum_t e_1^2(t) \Big/ \sum_t I^2(t) . \qquad (4)$$

$$R_{NR} = \sum_t e_2^2(t) \Big/ \sum_t S^2(t) . \qquad (5)$$

0-7803-7278-6/02/$10.00 ©2002 IEEE

TABLE 1. THE RATIO OF ENERGY LOSS AND NOISE RESIDUE

| Intrusion | Proposed model | | Wang-Brown model | |
|---|---|---|---|---|
| | $R_{EL}$ (%) | $R_{NR}$ (%) | $R_{EL}$ (%) | $R_{NR}$ (%) |
| N0 | 2.52 | 0.01 | 6.99 | 0 |
| N1 | 7.76 | 1.08 | 28.96 | 1.61 |
| N2 | 1.96 | 0.11 | 5.77 | 0.71 |
| N3 | 5.27 | 1.64 | 21.92 | 1.92 |
| N4 | 5.45 | 0.91 | 10.22 | 1.41 |
| N5 | 3.38 | 0.02 | 7.47 | 0 |
| N6 | 2.20 | 0.09 | 5.99 | 0.48 |
| N7 | 4.14 | 1.71 | 8.61 | 4.23 |
| N8 | 2.56 | 1.34 | 7.27 | 0.48 |
| N9 | 10.00 | 19.08 | 15.81 | 33.03 |
| Average | 4.52 | 2.60 | 11.91 | 4.39 |

The results are shown in table 1. Each value is the average of each intrusion over 10 voiced utterances. Intrusions are: N0 — pure tone, N1 — white noise, N2 — noise bursts, N3 — “cocktail party” noise, N4 — rock music, N5 — siren, N6 — trill telephone, N7 — female speech, N8 — male speech, and N9 — female speech. The table also shows for comparison the results from the Wang-Brown model [7]. Each value is the average of a certain intrusion type. Compared with the Wang-Brown model, our model generates significantly smaller ratios of energy loss, especially for N1 and N3. Similar ratios of noise residue are obtained from both models except for N9 where our result is much better. We note that our overall improvement comes mainly from high-frequency channels.

To compare waveforms directly, we also measure a form of signal to noise ratio (SNR) in decibels using the resynthesized waveform from the ideal stream as ground truth:

$$SNR = 10\log_{10}[\sum_t I^2(t) \Big/ \sum_t (I(t)-S(t))^2\,]\,. \tag{6}$$

The average SNR for each intrusion is shown in Fig. 8. Compared with the Wang-Brown model, our model increases SNR for all the intrusions. The average improvement is about 4.5 dB. The performance of this model is also better than a preliminary version that does not carry out the segmentation process based on common AM and use the AM repetition rate for grouping cells [11].

## IV. CONCLUSION

Our monaural model contains two segregation processes: initial segregation and final segregation. Both processes can be implemented through the oscillatory neural network employed in the Wang-Brown model [7], and it is not further addressed here.

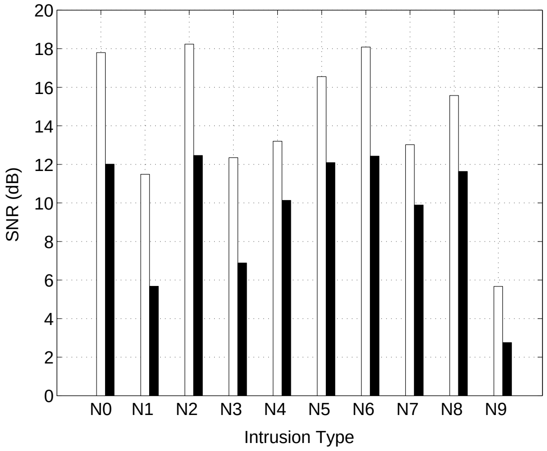

Fig. 8. SNR of segregated target speech. White bar: results from our model, and black bar: results from the Brown-Wang model. Different intrusion types are shown in the text.

Our model applies different mechanisms to segregate low-frequency speech and high-frequency speech. For high-frequency speech signals, it generates segments based on common AM and groups them by comparing AM repetition rates with estimated F0. Our model has been systematically evaluated on a mixture corpus, and it yields very good results. The performance of our model is substantially better than those pervious CASA systems evaluated on the same corpus [4] [5] [7] [11], especially in the high-frequency range. Our study demonstrates that computational investigation that incorporates ASA principles is a promising direction for monaural speech segregation, given the remarkable ability of the auditory system for the task.

## ACKNOWLEDGEMENTS

This research was supported in part by an NSF grant (IIS-0081058) and an AFOSR grant (F49620-01-1-0027).

## REFERENCES

[1] V. Zarzoso and A. K. Nandi, “Blind Source Separation,” *Blind Estimation Using Higher-order Statistics*, Boston: Kluwer Academic Publishers, 1999, pp. 167-252.

[2] D. O’Shaughnessy, *Speech Communications, Human and Machine*, 2nd Ed. New York: IEEE Press, 2000, pp. 323-336.

[3] S. Bregman, *Auditory Scene Analysis*, Cambridge, MA: MIT press, 1990.

[4] M. P. Cooke, *Modeling Auditory Processing and Organization*, U.K.: Cambridge University, 1993.

[5] G. J. Brown and M. P. Cooke, “Computational Auditory Scene Analysis,” *Computer Speech and Language*, Vol. 8, 1994, pp. 297-336.

[6] D. F. Rosenthal and H. G. Okuno, *Computational Auditory Scene Analysis*, Mahwah, NJ: Lawrence Erlbaum, 1998.

[7] D. L. Wang and G. J. Brown, “Separation of Speech from Interfering Sounds Based on Oscillatory Correlation,” *IEEE Trans. Neural Network*, Vol. 10, 1999, pp. 684-697.

[8] C. J. Moore, *An Introduction to the Psychology of Hearing*, 4th Ed. Academic Press, 1997.

[9] C. J. Darwin and R. P. Carlyon, “Auditory Grouping,” *Hearing*, 2nd Ed. Academic Press, 1995, pp. 387-424.

[10] H. Helmholtz, *On the Sensations of Tone*, Braunschweig: Vieweg & Son, 1863. (A.J. Ellis, English Trans., Dover, 1954).

[11] G. Hu and D. L. Wang, “An Improved Model for Speech Segregation,” *International Joint Conference of Neural Networks* 2001.

[12] R. Kumaresan and A. Rao, “Model-based Approach to Envelope and Positive Instantaneous Frequency Estimation of Signals with Speech Applications,” *J. Acoust. Soc. Am.*, Vol. 105, 1999, pp. 1912-1924.

[13] M. P. Cooke, P. D. Green, L. Josifovski, and A. Vizinho, “Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data,” *Speech Communication*, Vol. 34, 2001, pp. 267-285.

0-7803-7278-6/02/$10.00 ©2002 IEEE