A ROBUST BEGIN-END POINT DETECTOR FOR HIGHLY NOISY CONDITIONS

R. Martínez, A. Alvarez, P. Gómez, V. Nieto, V. Rodellar, M. Rubio and M. Pérez
Dept. Arquitectura y Tecnol. de Sist. Inf., Facultad de Informática
Universidad Politécnica de Madrid, Campus de Montegancedo, s/n
Boadilla del Monte, 28660, Madrid, SPAIN
e-mail: pedro@pino.datsi.fi.upm.es

ABSTRACT

Most recognition methods, which have shown to be highly efficient under noise-free conditions fail dramatically with S/N ratios around or below 10 dB. One of the consequences of these high noise levels is that most Begin-End Point Detectors fail to separate properly the speech segments of the noise ones. Therefore, the speech recognition mechanisms will not have a clear boundary to start the processing of the signal, and as a consequence, speech segments will be lost, and noisy segments will be used in recognition. The overall reliability of the Speech Recognition System will dramatically experience the consequences of this impairment in its results. What is being proposed in this paper is to use the side information provided by an Adaptive Noise Canceller for the dynamic detection of word boundaries.

1 INTRODUCTION

Speech Recognition in Noisy Environments is of vital importance for the success of certain applications in the domain of Communications, Automotion, Avionics, and other harsh situations where common recognizers fail to meet acceptable standards due to the negative influence of environmental noise collected during the recording of the Speech Trace [1]. Speech Recognition under high noisy conditions is a difficult task. Fig. 1 shows an example of a set of isolated One-Command Words, corresponding to the sequence */left/, /right/, /up/, /down/, /go/,* and */stop/*, recorded by a primary microphone within a field of about 90-95 dB SPL. In this example the environmental conditions are quite harsh. Two of the words are quite difficult to distinguish, as the energy of the noise is at the same level of the energy of the speech signal, and only the central two words are about 8 dB over the level of noise. The non-stationary characteristic of the noise (in level and in spectral distribution) makes it more difficult to find a rule for the detection of the presence of voice (sudden bursts of noise could shoot the speech detection mechanism). An added difficulty is the nature of the noise, which consists of a mixture of music, car racing motor sounds, and speech. This is a very important point, as the presence of speech interference could be considered a valid word.

In such situation a noise canceling device should help the speech recognition process offering a cleaned version of the speech trace, which should render better results in recognition rates. We can also expect that the detection of the word boundaries in this circumstances should be much easier. So in a first approach we can consider the possibility to directly use the cleaned speech trace to detect of the voice periods. Nevertheless, we can take advantage of the information of the original trace.

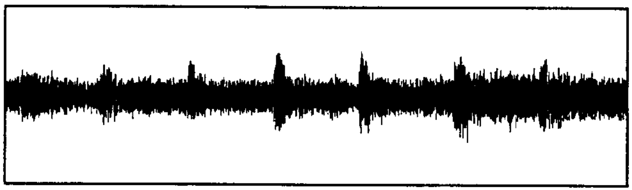

Figure 1.a. Noisy Speech trace (Primary Microphone)

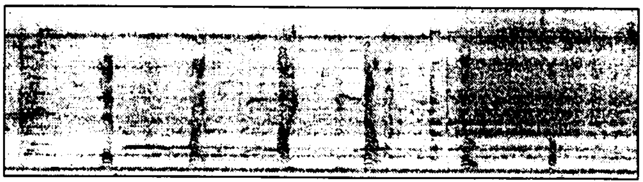

Figure 1. b. Power spectrum of the Noisy Speech Trace

2 METHODOLOGY

The preprocessing stage proposed, which may be seen in Fig. 2, is based in a two-Microphone array (Speech Source or Primary, and Noise Source or Reference).which give two signal channels that feed an Adaptive Lattice-Ladder Filter [2-4]. According with this scheme several algorithms may be devised to update the weights of the filter. The one finally selected in this research (although other ones were initially checked [5]) is RLS Lattice Ladder algorithm (Recursive LSL algorithm using *a posteriori* estimation errors) [3,4]. The adaptation step used for this filter was α = 0.9999, and the factor of lattice initial residual error energy was taken as $\varepsilon = 5.10^8$, this value ensuring a high lock-up performance and an acceptable degree of stability during lock up in non-stationary conditions.

For practical considerations, the best microphone separation is in the range from 15-30 cm. Shorter distances should imply a high cross-talk, and as a consequence, the voice registered by the reference microphone should be considered noise and cancelled. On the other hand, farther separations produce a loss of correlation between the noise received by the Reference and the Primary microphone. Another important reason to keep the microphone separation as close as possible is the length of the filters. With a sampling rate of 11025 Hz, the filter dimensions corresponding to a separation of 15-30 cm. are 10 to 20 stages (double the separation in samples, as the reference signal has to be delayed to prevent noise arriving first to the Primary Microphone). That number of taps may overwhelm the processor power. In our case the limit imposed by the processor (DSP Card using TMS320C31-50), is 14 taps (20 cm.) [6].

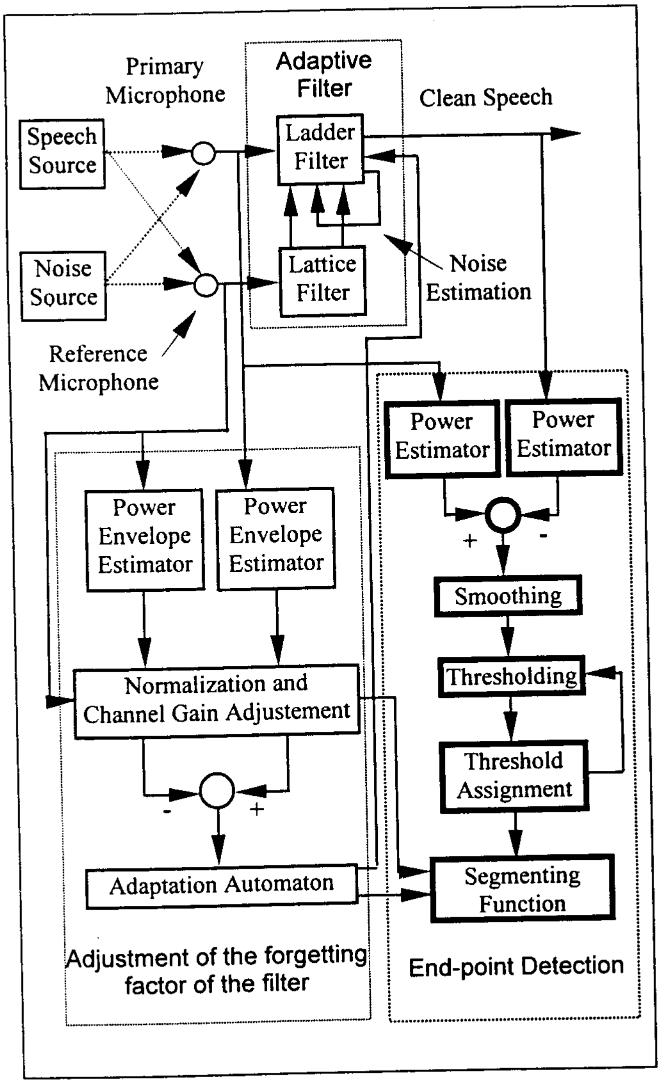

Figure 2. General framework for the proposed end-point detection methodology.

2.1 Forgetting factor Adjustment Block

There are three great blocks in the general framework:

- Adaptive filter (which has been described previously).
- End-Point detector.
- Forgetting factor adjustment block.

It was observed that the optimum values for the adapting factor α were not independent from the S/N ratio. The system, which had a good behavior in low S/N ratios, did not work properly when the level of speech was neatly superior to the level of noise. In such cases, an unlocking period of the algorithm was observed. In general, a value for α close to the unity produces a slower adaptation rate, but keeps the cancellation ability within reasonable limits. On the other hand, smaller values for α would result in a faster re-locking process at the cost of decreasing the overall accuracy of the noise canceller. Fortunately, these instabilities take place after the ending of high energy words, so these cases can be easily detected by comparing the power envelope of the Reference and the Primary Channel. A channel gain adjustment mechanism ensures that the comparison between channels is done in the same power conditions. This power information feeds a state automaton which decides the best forgetting factor for the filter. A low factor is used when low power envelope difference follows a great difference (high S/N ratio). This information, as well as the state of the automaton help the detection of word boundaries.

2.2 End-Point detector

In a first approach, we could attempt to use the information of the difference of the power envelope between the Primary and the Reference Microphones. This method should overcome the problem of the continuously changing level of noise, as these changes should be registered in the same way by both microphones. Nonetheless an End-Point Detector like that should only be capable to detect speech in high S/N ratio conditions. Otherwise the presence of speech would completely be hidden by the ambient noise.

A second idea could be to directly use the cleaned speech at the output of the adaptive filter (Fig. 3). But as we can see, although the presence of speech is now more clear, there are still zones that are at a similar level as noise, making it very difficult to precisely decide the boundaries of the words.

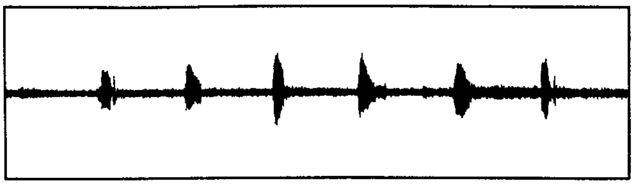

Figure 3.a. Cleaned Speech Trace

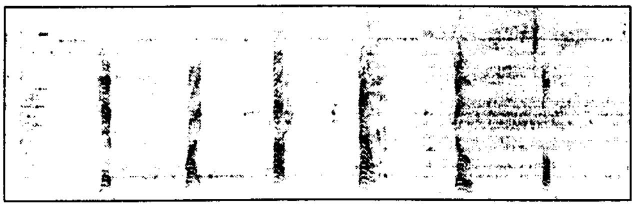

Figure 3.b. Power Spectrum of the Cleaned Speech Trace

The same effect observed in the waveform and in the spectrum, can be noticed in Fig 4.a. The upper trace represent the *Average Power* of the noisy speech (Primary Microphone Trace). The lower trace is the *Average Power* of the cleaned speech (output of the adaptive filter). As we can see, the noise level is not constant either in the noisy trace or in the cleaned trace. In the middle of the recording the level of noise for the cleaned trace is about 80 dB, but in the first part and at the end, the level of noise is more than 85 dB. However, the difference between those two traces (improvement in the S/N ratio) is almost constant when speech is not present, decreasing with the presence of speech (which is not cancelled).

The method proposed here relies on the continuous tracking and comparison of the instantaneous power of both the primary *s(n)* and the clean speech *c(n)* signals, given as follows:

$$S(m) = 10\ log10 \sum_{k=-N}^{N-1} s^2(m\ N+k) \qquad (1)$$

$$C(m)= 10\ log10 \sum_{k=-N}^{N-1} c^2(m\ N+k) \qquad (1')$$

$$D(m) = S(m) - C(m) \qquad (2)$$

$$\bar{D}(m)=1/3 \sum_{k=-2}^{0} D(m+k) \qquad (3)$$

$$\vartheta_m =1/15 \sum_{k=-14}^{0} \bar{D}(m+k)\ -3 \qquad (4)$$

With N = 128 being half the size of a 256-sample sliding frame, m being the frame index. $\bar{D}(m)$, the smoothed difference between both energy traces *S(m)* and *C(m)* will remark the segments where speech is explicitly present above the noise. This difference is compared with an adaptive threshold ϑ_m, which is used to assign values to a segmenting function σ_m. That threshold is not updated when speech is detected.

The conditions of the segmentation function are:

1) Begin-points are always detected by the adaptive threshold ϑ_m.

2) If $\vartheta_m < \bar{D}(m)$ end-point is detected.

3) When a difference lower than 2 dB is detected between ϑ_m and $\bar{D}(m)$ during more than 7 frames, the value of the threshold is decreased in 1.5 dB.

4) When more than 20 frames in which S/N is high enough to cause the forgetting factor adjustment block modify α, then the end-point condition is changed to one based on the difference between the power envelope estimations of the reference and the primary channels. (In such conditions, the S/N ratio is high enough to make it possible to detect the speech periods).

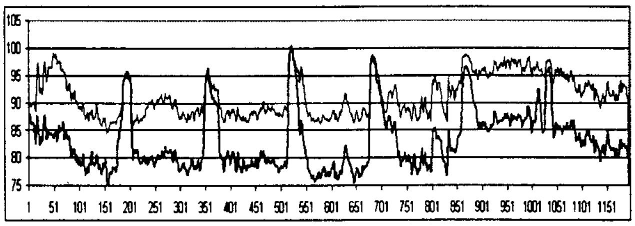

Figure 4.a. Average power S(m) of the Primary Microphone (upper trace), and of the filtered signal C(m) (lower trace).

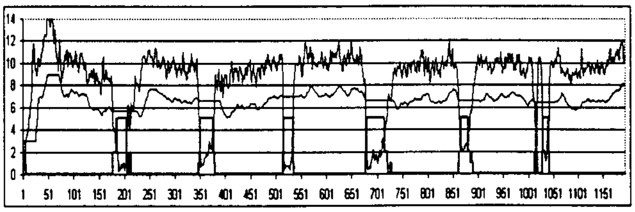

Figure 4.b. Average power difference as evaluated in (3) (upper trace), adaptive threshold ϑ_m, (middle trace) and segmenting function σ_m. (lower trace)

3 RESULTS

To test the performance of the proposed thresholding method for the dynamic segmentation of the speech trace a set of experiments was conducted which is summarized in what follows. An LSL (direct update) Lattice-Ladder Filter was used with the conditions exposed in [5]. The output signal, and its corresponding spectrogram are given in Figs. 3.a and b. Figure 4.a shows the average powers in Equations (1-1'). Figure 4.b gives the average power difference as evaluated from (3) (upper trace), the threshold function ϑ_m (middle trace) and the segmenting function σ_m (lower trace). It may be seen that the segmentation produces a set of blocks which remark noticeably the structure of the speech trace (see the detection of the fricative in */stop/* precisely detached from the background noise).

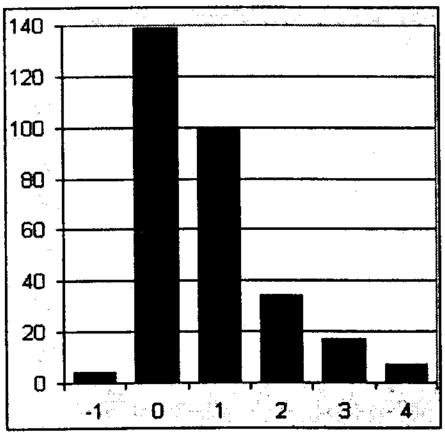

Figure 5.a. Number of begin points for 50 utterances of the 6 words relative to frame index.

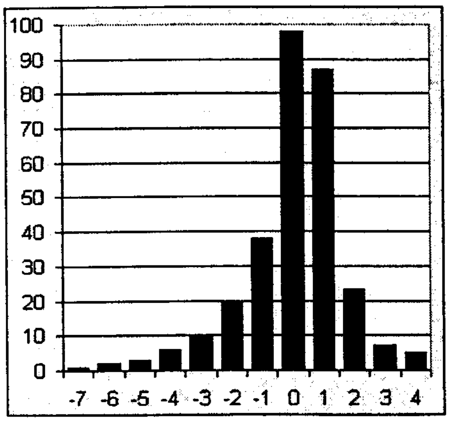

Figure 5.b. Number of end points for 50 utterances of the 6 words relative to frame index

Finally, the segmenting ability of the method was measured using a set of 50 utterances of the 6-word traces, produced by 10 different speakers with S/N ratios in the range from 0-5 dB . The results of the segmenting method were checked against a visual inspection method and are shown in Figures 5.a and b. In Fig. a the histogram plotting the number of begin point detections is plotted as a function of the relative number of frames shifted. Thence, there was coincidence (0-frame shift) for 139 begin points. Seven begin points were detected 4 frames in delay. Figure 5.b, on its turn, gives the same figures for end-point detection. In this case there was also good agreement between both methods for 98 cases. The deviation for the worst case was of -7 frames in one case. As a conclusion, it must be pointed out that the segmentation methods proposed produce a cleaner speech trace, keep the average cancellation behavior as constant as possible, and help in deciding on the begin-end point detection process. Applications can be found in Robust Isolated-Word Speech Recognition, Clean Speech Communications, and others [7].

4 ACKNOWLEDGMENTS

The present Research is being carried out under grants TIC95-0122, TIC96-1889-CE, ESPRIT IVORY n° 20277 and TIC97-1011.

5 REFERENCES

[1] Furui, "Recent Advances in Robust Speech Recognition", *Proc. of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-à-Mousson, France, 17-18 April 1997, pp. 11-20.

[2] Deller, J. R., Proakis, J. G. and Hansen, J. H. L., *Discrete Time Processing of Speech Signals*, MacMillan, 1993.

[3] Haykin, S., *Adaptive Filter Theory*, 3rd Ed., Prentice-Hall, Englewood Cliffs, N.J., 1996.

[4] Proakis, J. G., *Digital Communications*, 2nd. Ed, McGraw Hill, 1989.

[5] Martínez, A. Álvarez, V. Nieto, V. Rodellar and P. Gómez, "ASR in Highly Non-Stationary Environments using Adaptive Noise Canceling Techniques", *Proc. of the ESCA-NATO Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-à-Mousson, France, 17-18 April, 1997, pp. 181-184.

[6] Martínez, R., A. Álvarez, V. Nieto, V.Rodellar, and P.Gómez, "Implementation of an Addaptive Noise Cancelleron the TMS320C31-50 for Non-Stationary Environments", *Proc. of the 13th International Conference on Digital Signal Processing*, Santorini, Greece, 4-4 July, 1997, pp. 181-184.

[7] IVORY - ESPRIT project n° 20277: http://moral. datsi.fi.upm.es/projects/IVORY/IVORY.html.