# A SPEECH PRE-PROCESSING TECHNIQUE FOR END-POINT DETECTION IN HIGHLY NON-STATIONARY ENVIRONMENTS

R. Martínez, A. Álvarez, P. Gómez, M. Pérez, V. Nieto and V. Rodellar
Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n, 28660 Boadilla del Monte, Madrid, SPAIN
Tel.: +34.1.336.73.84, Fax: +34.1.336.74.12, e-mail: pedro@pino.datsi.fi.upm.es

## ABSTRACT

The determination of the precise moment in which speech begins or ends is an important problem in ASR. As showed in [1], small separations from the optimum beginning and ending point, imply a great decrease in the recognition accuracy. The presence of noise [2] [3], specially when its level is high (around 95 dB as in the case of this work), and its characteristics are highly non-stationary, is an added problem, since it can produce false shots (more probable when the noise includes speech sounds). That is the reason why in such conditions, it is important to have a pre-processing stage that removes as much noise as is possible, and that gives some clues that help to build an end-point detector for those environments.

The method here presented offers a pre-processing technique for highly noisy and non stationary environments, which at the same time that enhances the speech, gives an equalised version of the SNR improvement (Mean Spectral Energy Difference), whose main characteristic is that large differences in the level of noise are changed to a little ripple, while the presence of speech is distinguished by a large decrease in this Mean Spectral Energy Difference. Following this technique, any End-point Detection approach (explicit, implicit or hybrid [3]) may render acceptable results.

## 1 INTRODUCTION

This paper is intended to describe a practical application of certain Signal Processing Techniques used for End-Point Detection in Highly Non-Stationary Environments. These are characterised by high noise levels (of about 100 dB) and are specially harsh to deal with, as many different noise sources may be active at a time, varying dramatically with time, and are a paradigm for many similar cases (conference rooms, discotheques, automotive cabinets, industrial environments, etc.). One of the main problems posed by Isolated-Word Speech Recognition in these cases is found in the difficulty to establish reliable end points for the fragmentation of speech [2] [3]. The techniques used are based in two-microphone cancelling schemes with Adaptive Algorithms [4] [5] [6]. The General Framework is shown in Fig. 1. The recording scheme is based on a two-microphone structure, one for Noisy Speech (Primary), and the other for the Noise in itself (Reference). Assuming that the Speech Source is well separated from the Reference Microphone, the Noise is estimated by a *lattice filter*, and its *backward residuals* are used to adapt the weights of a *ladder filter*, in combination with the *Estimation of Noise* generated

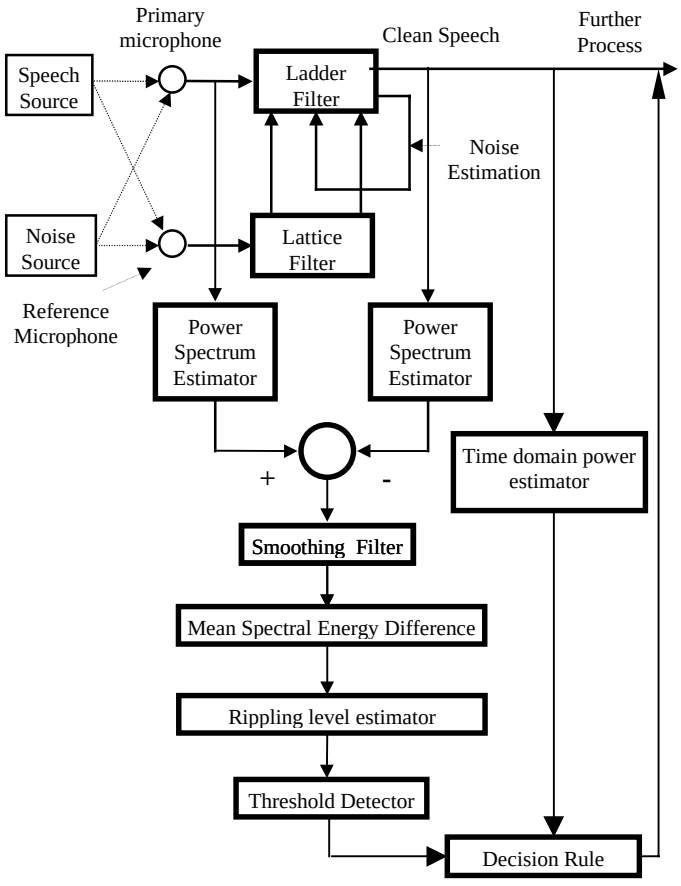



**Figure 1.** General Framework for the proposed methodology. A classical Adaptive Noise Cancellation Scheme (a combination of an *adaptive lattice and ladder filters*), Spectral Estimation of the Noisy and Clean Speech, and power estimation in the time domain for the detection of Speech and Speechless frames are being dynamically combined.

adaptively in this last filter, which generates Clean Speech as its output. The strategies classically used to combine the different estimators of Noise, and the *backward residuals*, different algorithms may be used to update the weights of the *ladder filter*. These give place to different implementations of the basic cancelling scheme. The algorithms studied have shown a good behaviour under hard work conditions: highly noisy and non stationary environments, no *a priori* knowledge of the characteristics of the signal or the noise, the need to preserve the quality of speech, and of immediate response, possible further processing of speech, etc.

## 2 METHODOLOGY OF THE PROVES

The speech traces used to produce the results shown were recorded with a two-microphone scheme as mentioned. A pair of high-quality cardioid microphones and a pre-amplifier were used to feed two channels to a *general purpose speech recording card.* Noise was reproduced at laboratory conditions from real recordings taken in a real scenario. The level of the noise was set to 95 dB, and the SNR about 0 dB. The resulting traces were stored and batch-processed in a *general purpose workstation*, using a C-C++ version of the above mentioned algorithms embedded in a User Interface to allow the audio and visual inspection of the recorded traces and the corresponding results. The microphones were placed at a distance of 20 cm., and the filter dimensions ranged correspondingly to 7+7 delay and processing stages for a sampling rate of 11025 Hz.

Once clean speech is obtained using adaptive cancellation techniques, an estimation of its Power Spectrum is generated and averaged, and then subtracted from the mean Power Spectrum of the Noisy Speech. The resulting values are smoothed in a special averaging filter, and an estimation of the *Mean Spectral Energy Difference* between the noisy and cleaned speech is obtained. This representation is specially useful to detect the presence of speech in noise for *End-Point Detection*. The *Mean Spectrum Energy Difference* between the Noisy Speech trace and the Reference Noise trace is evaluated using 256-sample FFT frames (with an overlap of 128 samples), averaging the energy over the whole spectrum for each frame, and smoothing the results in the time domain with an order-5 moving average smoothing filter over the frame set. The *Mean Spectrum Energy* for the Noisy Speech (upper trace) and the Cleaned Speech (lower trace) may be seen in Fig. 2.a.

In Fig. 2.b the *Mean Spectrum Energy Difference* is presented. In this trace the presence of speech is clearly enhanced, as the average energy level is around 10-12 dB (the amount of noise cancelled), except in the fragments where the most part of the contribution to the energy of the Noisy Speech is due to speech in itself. By direct inspection of figures 2.a and 2.b, we can notice that variations higher than 10 dB in the level of noise of the Mean Spectrum Energy of the speech produce a little ripple in the Mean Spectrum Energy Difference. The fragments of speech can be easily pointed out from the sudden and large decays in the Mean Spectrum Energy Difference (frames 163-217, 325-379, 487-541, 649-703, 685-919 and 973-1081), as seen in Fig. 2.b. Using the lower trace (cleaned speech) for the detection of beginnings and endings would require detection of the deviations in the energy from the baseline, which should require further processing (for example, an average filtering). If the trace in Fig. 2.b is used instead, this detection may be carried out using an absolute reference for thresholding from the more stable baseline (topline, to be more precise). This result is specially important in environments where the noise level is continuously changing, making it quite difficult to take an *a priori* decision of the location of speech fragments. This pre-processing technique can be used for helping End-point Detection, with independence of the final approach decided (explicit, implicit or hybrid [3]).

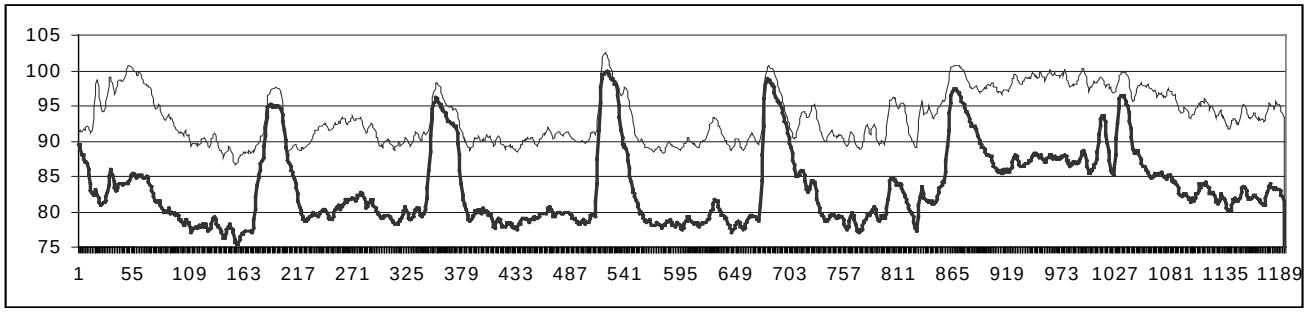



**Figure 2.a** Spectral Energy corresponding to Noisy Speech (upper trace) and Cleaned Speech (lower trace. The improvement does no sustain for energy peaks in which speech is not present (see specially those in the leftmost part of the figure).

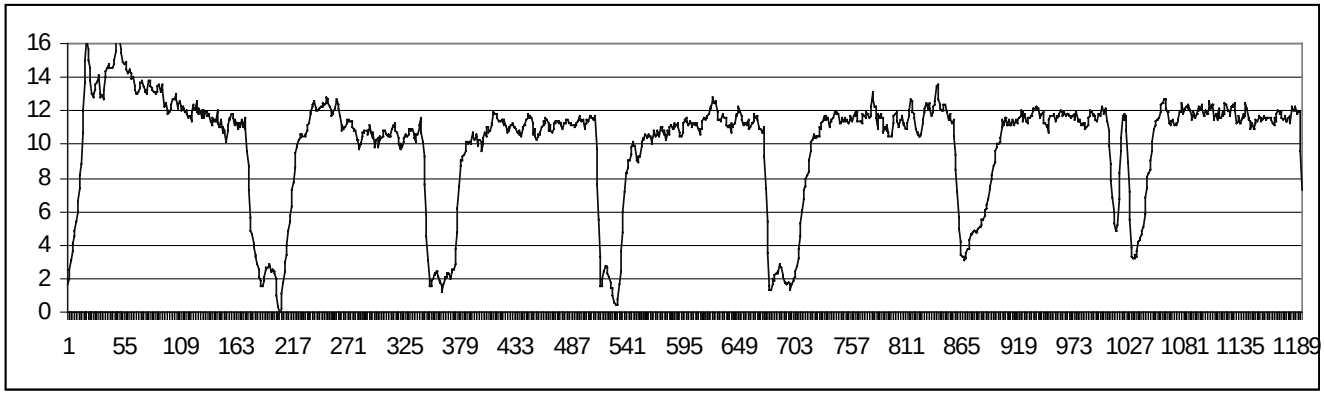



**Figure 2. b**. Mean Spectral Energy Difference between both traces in Fig 2.a. A Difference of 10-12 dB may be measured between both traces, which reduces when Speech is dominant over noise to almost zero. This trace may be used for End-Point Detection.

Once the Mean Spectral Energy is obtained, the mean level of the rippling (non-speech periods) is obtained to calculate a threshold. Decreases of the mean spectral Energy deep enough below that threshold, an with a sufficient duration are considered speech.

A time domain estimation of the energy of the signal is also evaluated, since some words ending with plosives are not well detected. The pressure wave generated can be easily detected in the time domain.

With both the spectral and the temporal information, the endpoints are determined.

## 3 RESULTS

Several utterances of the words “left”, “right”, “up”, “down”, “go”, “stop” were recorded with a two-microphone scheme. The endpoints of the filtered signal were first determined by inspection, and the results were compared against the endpoints detected using this scheme.

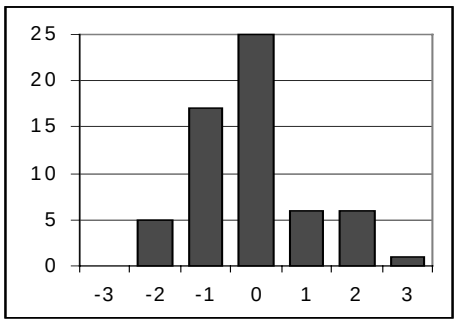
For such the smoothing filter was evaluated with 3 and 5 points windows.

The results are presented in frames. This frames have a length of 256 points (23.22 ms) and have an overlap of 128 points (11.61 ms). So an error of one frame in the determination of the beginning or ending point represents an error of 11.61 ms.

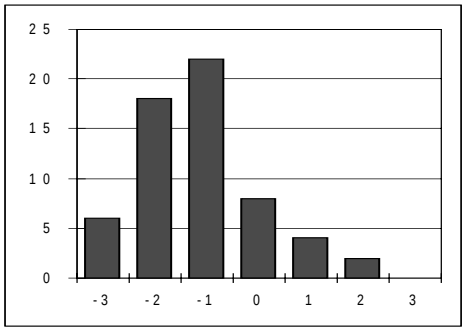
As the FFT frames are overlapped, the beginning and ending points of the words are present in two frames (frames 0 and –1 for the beginning, and 0 and +1 for the ending point), so both 0 and –1 frames for the beginning and 0 and +1 for the ending are correct.

It is better to estimate in advance the beginning of the real word inset rather than after it takes place (therefore, negative values for the frame are preferable than positive ones). For ending points the estimation is the opposite. It is better to produce late detects than anticipate ones, so the smoothing filter with a window of 5 points shows the better behaviour.

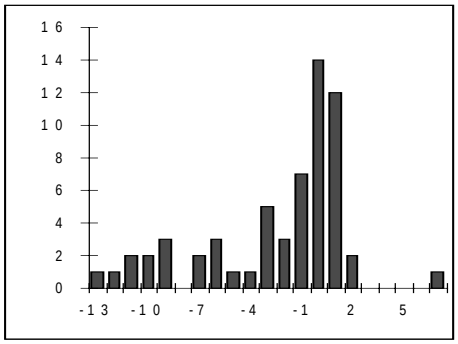
The importance of the time domain power estimator is also clear, as it improves the results. (This especially in the determination of the ending of /stop/, /left/ and /up/)



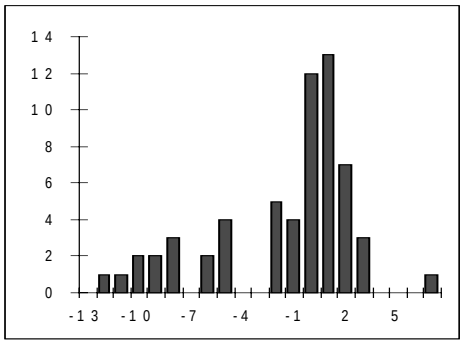

**3.a** Number of beginning-points detected as a function of the frame index for 10 utterances of the 6 mentioned words. A 3-point smoothing window was used



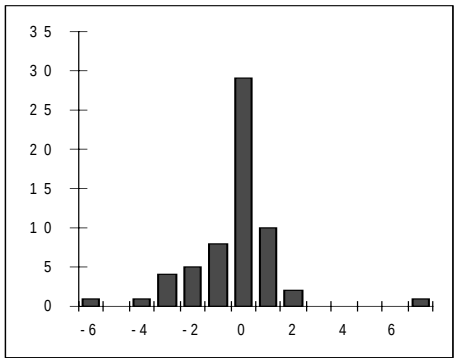

**3.b** Number of beginning-points detected as a function of the frame index for 10 utterances of the 6 mentioned words. A 5-point smoothing window was used



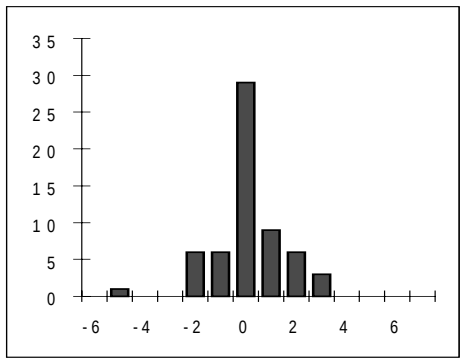

**3.c** Number of ending-points detected without time-domain power estimation. (3-point smoothing window).




**3.d** Number of ending-points detected without time-domain power estimation. (5-point smoothing window).




**3.e** Number of ending-points detected with time-domain power estimation. (3-point smoothing window).




**3.f** Number of ending-points detected with time-domain power estimation. (5-point smoothing window).

**Figure 3** Beginning and ending point detection deviations for 10 utterances of the words /left/, /right/, /up/, /down/, /go/, /stop/. Note that great errors in the detection of ending points (detections 7 to 13 frames before it really takes place -70 to 140 ms.-, corresponding to plosive endings) are solved if time domain power estimation is made.

## 4 ACKNOWLEDGEMENTS

This research is funded by ESPRIT Project *IVORY (Integrated VOice Recognition sYstem)*, no. 20277, and Grants TIC-95-1022 and TIC-96-1889-CE.

## 5 REFERENCES

[1] Wilpon, J. G., Rabiner, L. R. And Martin, T. B., "An improved word-detection algorithm for telephone-quality speech incorporating both sintactic and semantic constraints", *AT&T Tech. J.*, 63 (3): 479-498, March 1984.

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

[3] Rabiner, L., Juang, B.H., *Fundamentals of Speech Recognition,* Prentice-Hall, Englewood Cliffs, N.J., 1993.

[4] Widrow, B. et al. "Adaptive Noise Cancelling: Principles and Applications" *Proc IEEE,* vol. 63, no. 12, pp. 1692 - 1716, Dec. 1975.

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

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