Robust Voice Activity Detection using Cepstral Features

J.A. Haigh & J.S. Mason
Speech Research Group, Electrical Engineering Department
University College Swansea
SWANSEA, SA2 8PP, UK
Email: eehaigh@pyr.swan.ac.uk, eemasonj@pyr.swan.ac.uk
Phone: +44 792 294564
Fax: +44 792 295686

ABSTRACT

This paper reviews algorithms which rely on the analysis of time domain samples to provide energy and zero-crossing rates, together with more recent algorithms that use different methods for speech detection. We then examine a different approach using cepstral analysis, showing a high degree of amplitude and noise level independence.

We show that a cepstral based algorithm exhibits a high degree of independence to levels of background noise and successful speech endpointing can be achieved via thresholding cepstral distance measures. Through the use of a noise code-book we are able to provide a successful reference for Euclidean distance measures in the voice detection algorithm.

1 INTRODUCTION

Meaningful assessment of VAD algorithms must inevitably reflect the nature of the application. For example in recognition tasks, precise endpoints may be unnecessary, and the resultant recognition performance itself can be used as an indirect measure of VAD performance. However, assessment is more difficult in the case of 2-way voice transmission systems, where more accurate locations of 'talkspurts' are needed to assess quality of service and 'double-talk' occurrences for example.

Several long established VAD's have been used reliably in the detection of speech in a variety of applications, such as [1], [2] and [3]. In common with others, these are based on a combination of short-term energy and zero-crossing-rate (ZCR) measurements, and rely on the observation that for a 'talkspurt' to exist there is likely to be a burst of energy somewhere along its time course, energy being the main detection stimulus. Then, on the assumption that ZCR's for speech and background noise are generally different, the end-point decisions are refined using ZCR measures, this helping to define areas of low energy speech.

In the experience of the Authors such algorithms are inherently simple in terms of computation, and hence present no difficulties in real-time applications. Predictably we find they exhibit dependencies on estimates of the prevailing noise statistics, on dc offsets in the energy estimates, and a lack of robustness to changes in absolute levels. While adaptive algorithms, such as [3], show a greater degree of robustness due to their dynamic operation, nonetheless they do tend to depend on accurate measurements of background noise.

More recently [4] reports on a promising algorithm, first proposed by [5], with high noise immunity. Detection of speech periods is achieved by least squares periodicity measures, sampled data being in the frequency range of 200-1000 Hz. This narrow band width helps reduce the probability of interference. Periodic noise though can still disrupt the decision process. An automatic gain control is incorporated to help distinguish between low level speech and noise, with minimum threshold adaptation.

2 CEPSTRAL ANALYSIS FOR VAD

In an attempt to overcome some of the limitations mentioned above, particularly that of greater amplitude and noise independence, while retaining the characteristics of simplicity, we propose in this paper a new approach to VAD namely, one based on a form of cepstral analysis found in many recognition systems. We use a form here called PLP which has been shown to have some robustness to noise [6].

Such analysis can be viewed as an attempt to de-convolve the speech signal into its excitation component and a 'vocal apparatus' component, via homomorphic filtering. In normal speech analysis the cepstra vary as a function of time, reflecting the different speech sounds and, importantly in the context of speech recognition, provide an identifiable form of transcript for each particular utterance. These cepstral differences along the time course are key to the success of recognition systems. The basis of the proposed algorithm is the differences of cepstra when computed from speech and when computed from noise.

Recalling that fundamental to any VAD algorithm is the discrimination between background noise and speech, then the accurate modelling of the slowly varying parts of speech, as provided by the cepstra, is intuitively promising. Moreover, cepstra coming from noise are known to have a much lower variances than those from speech, especially in the case of lower 'quefrency'.

2.1 Normalisation and distance measures

Another distinct advantage of cepstral analysis is the almost complete lack of dependence on absolute signal levels. Of course, this must be the case if accurate modelling of the vocal tract (rather than the excitation) is to be achieved. The mechanics of this normalisation is attributable to a combination of the log function and cosine transform within the cepstral analysis.

The simplest discrimination measure we have examined, and one which proves to be successful, is the weighted Euclidean distance:

$$d = \frac{1}{p}\sum_{i=1}^{p}(c_i - c_i')^2$$

where p is the order of the cepstral analysis in this case 10, and c_i and c_i' are the i^{th} elements of two cepstral vectors.

3 CEPSTRAL VAD ALGORITHMS

McAulay [8] investigated VAD, using a modified version of Robert's algorithm [9]. The implementation relies on energy measures when background noise is slowly varying. Robert's [9] algorithm however, requires a relatively long training period to estimate noise and calculate the detection threshold. Also with high levels of narrow band noise the algorithm finds difficulties in discriminating speech periods to noise/non-speech periods.

Le Floc'h [10] reports improvements by the addition of spectral distance measures between the present frame and the average of the low energy spectra. The low energy frame spectra consists of noise and unvoiced speech periods and is expected to average towards noise. Distance measures are taken between this average and the present frame spectrum, decisions being based on a experimentally determined fixed threshold.

We have repeated work equivalent to that in [10], but have used initial frames of a recording to characterise non-speech. A better approach is to use a code-book designed to represent different noise classes. The difficulty remains to discriminate noisy speech, particularly in the case of low energy fricatives from background noise.

The spectral vectors for the test signal are compared against the code-book trained on noise vectors. Euclidean distance measures are used to compare the vector of the current frame to each vector of the code-book, the smallest Euclidean distance being adopted.

3.1 Experimental Results

Two sets of data are chosen here for illustrative purposes. In the first case a recording of approxi-

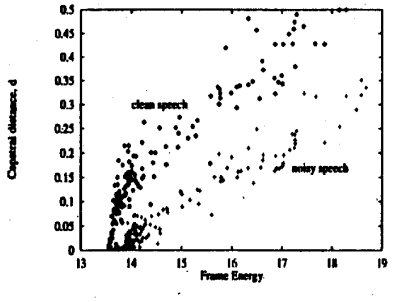

Figure 1: A scatter plot of Euclidean distance against log energy for speech, where ◇ indicates clean speech and + indicates noisy speech (SNR = 15dB)

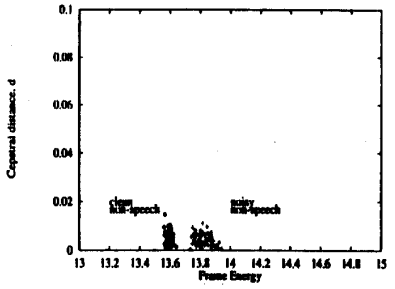

Figure 2: A scatter plot of Euclidean distance against log energy for non-speech periods. Note expanded scale for d by a factor of 5

mately 20 seconds has been used to generate scatter plots of distance measures, d, plotted against the frame energy, the energy being defined as:

$$energy = log_e \sum_{i=1}^{n} x_i^2$$

where n is the number of samples, in this case $n = 256$ corresponding to 25.6ms of speech and x_i is a 12 bit sample.

Figure 1 shows a scatter plot for clean and noisy speech periods from hand labelled data. As would be expected there is an increase in energy, hence the shift to the right in the case of noisy speech. However on the distortion axis while there is some decrease in d, overall this is small, showing itself in a small decrease in slope of the scatter plot trend. In the case of non-speech, Figure 2, again as would be predicted there is a shift to the right for non-speech plus noise, but it is very clear that there is no overall increase in d, implying a threshold of about 0.01.

The second case is an utterance of "sheep". The recording is approximately 2 seconds long, the make-up of which provides a stern test for accurate end-pointing. We then add a very high level of Gaussian noise (SNR -2dB across the recording) and some large noise 'spikes' and repeat the VAD experiment. Figures 3 and 4 show the clean and noisy cases respectively.. In the noisy case VAD's based on energy and zero-cross measurements fail in these conditions, especially with impulse noise. However it can be seen that, the new cepstral VAD offers a high degree of noise immunity and the recording can be successfully end pointed.

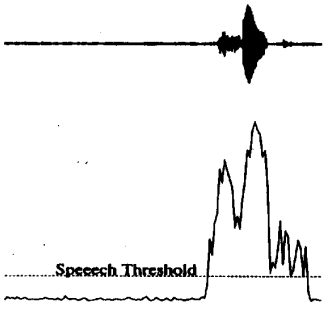

Figure 3: The time waveform (top) and the corresponding cepstral distance d for an utterance of sheep.

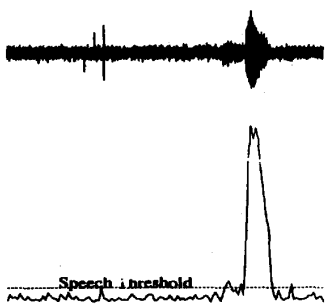

Figure 4: As for Figure 3 but with white Gaussian and impulse noise added (SNR = -2dB for the whole recording).

4 CONCLUSIONS

The paper reports on the performance of a new cepstral based VAD algorithm. Existing algorithms explicitly attempt estimates of background noise, and require some form of adaptation in order to accommodate normal variations levels and adverse conditions.

Cepstral analysis has been optimised over a number of years in the context of speech recognition. It thus provides an excellent model of the slow but meaningful variations in speech. Such variations tend not to be exhibited by non-speech, and this contrast is shown to provide the basis of a robust VAD system. Good results are achieved without any explicit measurement of, or adjustments for either absolute signal levels or background noise levels, desirable characteristics which are not to be found in any other VAD algorithms known to the Authors.

REFERENCES

[1] L.R. Rabiner and M.R Sambur, *"An algorithm for determining the endpoints of isolated utternaces"*, The Bell System Technical Journal, Vol. 54, No. 2, pp. 297, February 1975.

[2] H.H. Lee and C.K. Un, *"A study of On-Off characteristics of conversational Speech"*, IEEE Transactions on Communications, Vol. COM-84, No. 6, pp. 630, June 1986.

[3] J.A. Jankowski, *"A new digital voice-activated switch"*, Comstat Technical Review, Vol. 6, No. 1, pp. 159, Spring 1976.

[4] R. Tucker, *"Voice activity detection using a periodicity measure"*, IEE Proceedings, Vol. 139, No. 4, August 1992.

[5] M. J. Irwin, *"Periodicity estimation in the presence of noise"*, Inst. Acoust. Conf. 1979, Windemere, UK, and JSRU Report 1009, 1980.

[6] J. Junqua, H. Wakita, *"A comparative study of cepstral lifters and distance measures for all pole models of speech in noise"*, ICASSP-89, Vol. 1, pp. 476-479, 1989.

[7] D.K. Freeman, G. Cosier, C.B. Southcott and I. Boyd, *"The voice activity detector for the pan-european digital cellular mobile telephone service"*, ICASSP-89, Vol. 1, pp. 369-372, 1989.

[8] R.J. McAulay and M.L. Malpass, *"Speech enhancement using a soft-decision noise suppression filter"*, IEEE Trans, ASSP, Vol. 28, No. 2, pp. 137-147,, April 1980.

[9] J. Roberts, *"Modification of piecewise LPC "*, MITRE Working Paper WP-21752, May 1978.

[10] A.Le Floc'h, R. Salami, B. Mouy and J-P. Adoul, *"Evaluation of linear and non-linear spectral subtraction methods for enhancing noisy speech"*, ESCA, pp. 131-134, November 1992.

ACKNOWLEDGMENTS

The Authors would like to acknowledge British Telecom who provided the stimulus for this work.