A VOICE ACTIVITY DETECTOR BASED ON CEPSTRAL ANALYSIS

J.A. Haigh & J.S. Mason

Speech Research Group, Electrical Engineering Department,
University College Wales, SWANSEA, SA2 8PP, UK

ABSTRACT

This paper proposes a new approach to speech end-point detection based on cepstral analysis. The algorithm is based on explicit (static) modelling of speech and non-speech, and decisions are made on each incoming (overlapped) cepstral frame, according to model similarity scores.

The cepstral analysis provides excellent level-independence, meaning that parameter adjustment, decision thresholds etc, are unnecessary. A high degree of robustness to additive noise is demonstrated, even though the models are static. Accurate end-points are recovered with SNR levels of 0dB.

Keywords: **speech analysis, end-point detection, voice activity detection, robustness.**

1 INTRODUCTION

Many early voice-activity-detection (VAD) algorithms were based on a combination of short-term energy and zero-crossing-rate (ZCR) measurements, [1], [2] and [3]. Such algorithms detect speech by relying on the fact that an increase in energy is likely to occur somewhere between the two ends of a word or 'talkspurt'. Having identified this higher energy region, end-points can be found, essentially by moving backwards and then forwards along the time course, looking for falls in energy. Refinements of the end-points are performed using a combination of energy and ZCR measurements. Figure 1 shows an energy versus ZCR map fundamental to many such VAD algorithms.

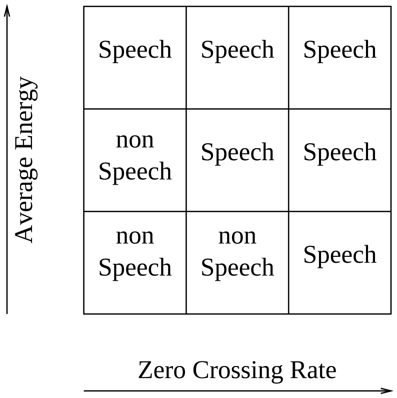

Figure 1: Energy versus ZCR map demonstrating the fundamental ideas behind many early VAD algorithms.

Perhaps the most important merit of these early algorithms is the relative ease with which they can be realised in real-time. Their biggest disadvantage proves to be sensitivity to changes in the statistics of the non-speech periods, and to a lesser extent, changes in levels. Even simple things, like changes in the signal offset, can have significant effects on short-term energy and ZCR measurements, and in turn lead to mal-operation.

McAulay [4] investigates VAD using a modified version of Robert's algorithm [5]. The implementation relies on energy measures during non-speech periods. Robert's [5] 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 difficulty in discriminating speech periods to noise/non-speech periods.

Le Floc'h [6] extends the work 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.

Recently, Tucker [7] reports on an algorithm, which is an extension of one first proposed by [8], and which exhibits a high degree of noise immunity. Detection of speech end-points is achieved by least squares periodicity measures, taken from a signal band-limited to the range 200-1000 Hz. This narrow band width helps reduce the probability of interference. Periodic noise though can still disrupt the decision process, and an automatic gain control is needed. Otherwise the algorithm is said to work successfully in high noise conditions with minimum threshold adaptation.

The proposed algorithm overcomes many of the difficulties mentioned above. The very fact that it is based on cepstral analysis means that it is essentially level-independent and therefore needs no AGC. Speech, non-speech discrimination proves to be good, even with high levels of background noise.

2 CEPSTRAL BASED VAD

Figure 2 contains a block diagram of the proposed VAD system. Separate models for speech and non-speech are created during training. These models need to be representative of their respective classes, although in practice we find that it is important to dominate the training data with speech evolutions and tails (for the speech model).

In testing, incoming cepstra are compared with the models and a frame-by-frame similarity score obtained from each. Assuming the two models to be pre-normalised, a simple ratio of the two distortions, passed through a fixed threshold, can be used for speech, non-speech decision.

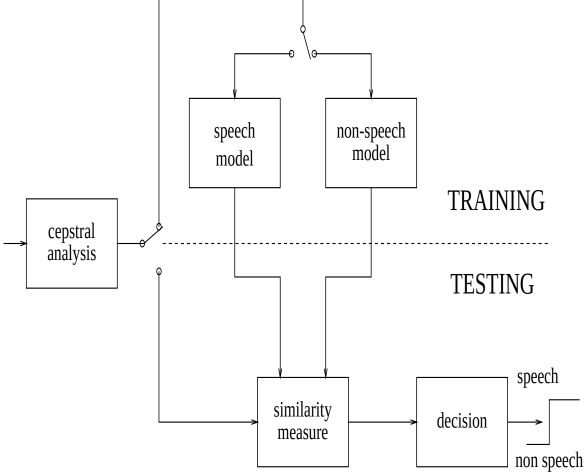

Figure 2: Block diagram of cepstral based VAD algorithm.

2.1 The models

The two models shown in Figure 2 do *not* have to encompass any time information, since the algorithm as described demands simple measures from each, on a frame-by-frame basis. Consequently a vector quantisation, codebook approach is perhaps an obvious first candidate, and the results below relate to this approach. However an alternative is being investigated based on multi-layer-perceptron models, the advantage of which lie in inherent discriminative training, and possibly computation.

2.2 Distance Measures

With a conventional model such as the codebook, an explicit similarity measure is required. The simplest we have examined, and one which proves to be successful, is the vector quantisation distortion from a standard Euclidean distance:

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

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 EXPERIMENTAL RESULTS

Two sets of data are chosen here for illustrative purposes. In the first case a recording of an utterance "hello" which is approximately 1.5 seconds is used to demonstrate the performance of the algorithm with clean and noisy test data.

Figure 3 illustrates the clean data and the resultant input to the decision function. Shown in the initial period of the test data is a period of non-speech which includes artificial impulse 'spikes' and noise from the ring of a telephone handset. These artifacts are deemed stern tests, and have been found to cause mal-function of other reported algorithms, including some of those discussed in the Introduction. The decision input for this period clearly shows accurate classification of the non-speech section is achieved.

The speech region of Figure 3 comprises the utterance "hello" immediately followed by a trailing breath sound. By including other examples of breath sound in the non-speech training data, the decision waveform shows that this breath example is successfully excluded from the speech class.

In the second example, Figure 4, the noisy case is shown for the same utterance. The time waveform shows a large amount of noise (SNR = 0dB across the whole of the recording). The decision input again clearly shows accurate end-pointing is obtainable. In such noisy cases energy and zero-cross based VAD algorithms have been shown to fail.

The second example is approximately 3 seconds long and another stern test for accurate end-pointing. Figure 5 shows significant variation in speech levels and characteristics, yet the decision input is relatively constant, enabling successful end-pointing with a fixed threshold.

4 CONCLUSIONS

The paper reports on the performance of a new cepstral based VAD algorithm, which explicitly models speech and non-speech in order to discriminate between the two.

The algorithm works with fixed thresholds since

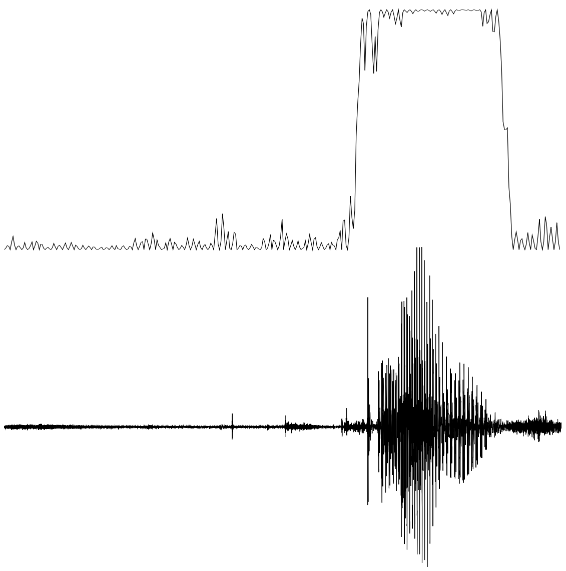

Figure 3: The time waveform (bottom) of the utterance "hello" for the clean case with corresponding speech, non-speech decision input.

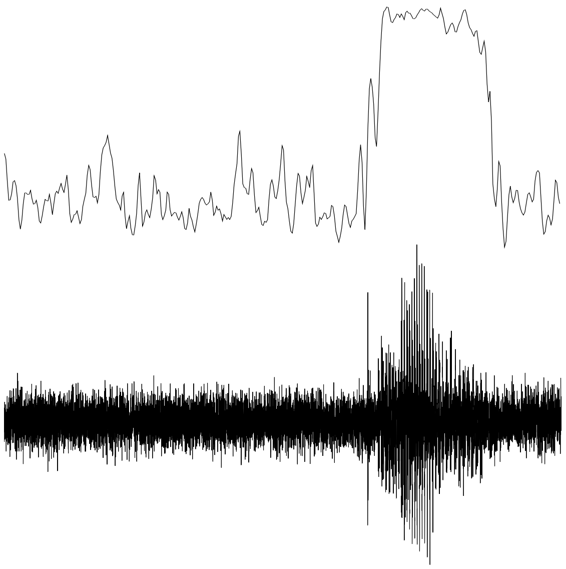

Figure 4: As for Figure 3, but with additive Gaussian noise (SNR = 0dB across the whole of recording).

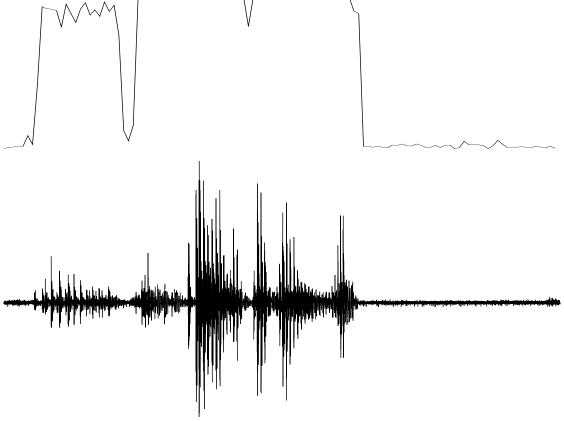

Figure 5: The time waveform of a 3-second recording with the corresponding decision input.

the cepstra exhibit a high degree of level-independence. Furthermore, by including within the models different *types* of non-speech background noise, eg breath noise, it is demonstrated that such (unseen) sections are correctly classified subsequently

Hence, good results are achieved without AGC, without monitoring the prevailing non-speech characteristics, and without threshold adjustments, even at SNR levels of 0dB.

ACKNOWLEDGMENTS

The Authors would like to acknowledge BT Labs who provided the stimulus for this work.

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-34, 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.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.

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

[6] 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.

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

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