Knowing the Wheat from the Weeds in Noisy Speech

H. Agaiby, T. J. Moir
Department of Electronic Engineering and Physics
University of Paisley, Paisley, PA1 2BE, UK
Tel. +44 141 848 3409, FAX: +44 141 848 3404, E-mail: hany@diana22.paisley.ac.uk

ABSTRACT

This paper introduces a word boundary detection algorithm that works in a variety of noise conditions including what is commonly called the 'cocktail party' situation. The algorithm uses the direction of the signal as the main criterion for differentiating between desired-speech and background noise. To determine the signal direction the algorithm calculates estimates of the time delay between signals received at two microphones. These time delay estimates together with estimates of the coherence function and signal energy are used to locate word boundaries. The algorithm was tested using speech embedded in different types and levels of noise including car noise, factory noise, babble noise, and competing talkers. The test results showed that the algorithm performs very well under adverse conditions and with SNR down to -14.5dB.

1. INTRODUCTION

Word boundary detection has been the theme of a significant number of research activities during the last two decades. The increasing interest in this theme is driven by the search for excellence in many speech applications such as speech recognition and speech enhancement, as well as discontinuous transmission (DTX) in cellular mobile telephone systems. Word boundary detection contributes to the performance of systems working in the above areas in various degrees. For example, an evaluation of a discourse system using an isolated-word recogniser [1] showed that more than half the recognition errors were due to the word boundary detector. In the area of mobile phones, it has been estimated that the use of DTX could approximately double the capacity of the radio system. A variety of algorithms have been proposed for speech detection, each of which is based on some criteria to differentiate between speech and silence or noise. The algorithms also vary in terms of the level of noise levels they can cope with and the detection quality emphasised. For example, while algorithms targeting speech recognition applications may put high emphasis on accuracy, those targeting speech logging or speech enhancement may compromise accuracy for robustness under severe noise conditions.

2. A REVIEW OF DETECTION CRITERIA

Word boundary detection algorithms presented in the literature rely on a variety of criteria to differentiate between speech and silence or background noise. In the following some commonly used criteria are reviewed with emphasis on their underlying strengths as well as restrictions.

2.1 Signal Energy

A number of algorithms use the signal energy as the main criterion for detecting word boundaries. A commonly used algorithm for isolated word recognition systems [2] uses signal energy to produce one or more sets of word endpoint sets. The algorithm is in a hybrid form that depends on feedback from the recognition scores to select the most likely endpoint sets for a particular utterance. It can however be used in an explicit form with little degradation to its performance. Another algorithm that was described by Savoji [3] uses energy metrics based on some statistical knowledge of speech for broadly detecting word endpoints. The endpoints are then adjusted using measures of signal energy and zero crossing. The main strength of the above algorithms is their ability to cope with speech artefacts and multi-syllabic words. The later algorithm has an added advantage that it can be easily implemented in real-time. The underlying assumption in the above algorithms, however, is that the noise level is fairly constant and significantly lower than the speech such that an increase in the signal energy indicates speech. An energy based algorithm that is more robust to noise was described by Junqua et al. [4]. The algorithm was tested on a HMM recogniser using noisy speech of SNR down to 5dB and with different types of noise. The obtained recognition accuracy when using the algorithm was almost as good as when using manual labelling. However, the algorithm was not tested with SNRs below 5dB which is typical in many speech enhancement applications. Moreover, none of the above algorithms has the capability to differentiate between one speech signal and another.

2.2 Periodicity

An algorithm that is based on a periodicity measure of the input signal was described by Tucker [5]. The algorithm relies on the periodic nature of the voiced parts of the speech to detect speech embedded in background noise. A least-squares periodicity estimator (LSPE) is calculated and speech is detected whenever this estimator exceeds a fixed threshold. The algorithm was shown to operate reliably in SNRs down to 0dB and detects most speech at -5dB. Nevertheless, the algorithm is intended to be used mainly in speech-logging applications and does not aim to find the exact word boundaries. The restriction on using this algorithm for applications where accuracy is of a prime importance stems from the fact that the algorithm does not detect unvoiced parts of the speech and that it takes 6 frames of 25ms each before a decision can be made. To differentiate between speech and other periodic interference a pre-processor is used to detect, and if possible remove, expected types of interference. The pre-processor, however, has to be tuned to a specific type of interference.

2.3 Coherence

Le Bouquin and Faucon [6] describe a voice activity detector that is based on Magnitude Squared Coherence (MSC). The algorithm calculates the MSC for signals at two microphones separated by 73cm then compare it with a predefined threshold, and whenever the MSC is higher than the threshold speech is detected. The performance of the algorithm was reported using speech recorded in a car with SNR down to 0dB. The performance of the algorithm was not reported in other environments where other types of noise exist or for SNRs below 0dB. The algorithm also assumes that the disturbing noises are spatially decorrelated while speech is highly correlated, conditions that are not valid in many situations. Moreover, in a cocktail party situation, the coherence may not differentiate between one speaker and the other.

2.4 Linear Prediction Parameters and Cepstrum

An LPC based algorithm [7] has been used for the more complex problem of voiced-unvoiced-silence detection. A spectral characterisation of each of the three classes is obtained during a training session. An LPC distance measure is then used together with an energy distance for discriminating between the classes. In another algorithm cepstral analysis was also used for voice activity detection [8]. In this algorithm models for speech and non-speech are created during training. The signal cepstra are then compared with these models and a speech/ non-speech decision is taken based on the similarity between the models and the calculated cepstra. The problem with the above two algorithms is that they require training under conditions similar to those encountered by the algorithm during its operation. This requirement becomes impractical if the algorithm is to work in a variety of noisy conditions. Another LPC based algorithm that does not require training is described in [9]. Word boundaries are estimated by comparing the squared residual prediction error of the speech signal with a threshold. The use of this criterion is based on empirical observations that showed a strong correlation between the squared residual prediction error and the nominal endpoints. The algorithm, however, can only cope with short transient pulses and low level noises such as those generated by breathing and small microphone movements, and not with higher noise levels.

In addition to the above criteria, zero crossing is used extensively to detect weak fricatives at the boundaries of speech utterances usually at a refinement stage. In general, speech detection algorithms use more than one measure for more accurate detection of speech endpoints. For example an algorithm described by B.S. Atal, and L.R. Rabiner [10] uses five measurements for voiced-unvoiced-silence classification of speech. For each speech segment, the algorithm computes a non-Euclidean distance measure from the set of measurements and the segment is assigned to the class with the minimum distance. Nevertheless, even when a number of parameters are used, word boundary detection at low SNR and in the presence of competing talkers is still far from being a solved problem. This was the motivation for developing the algorithm described below.

3. AN ALTERNATIVE SOLUTION

An alternative criterion to differentiate between desired-speech and undesired-speech or noise is the physical position of the signal sources. Although the talker position has been used previously in multi-mrophone based techniques to enhance noisy speech e.g. [11], its use in endpoint detection has not been thoroughly investigated before. The motivation of this criterion stems from simple observation of users of speech processing applications such as speech recognition systems, hands free phones, and hearing aids. In the above applications the position of the wanted speech source can, in most of the cases, be estimated to be within a predefined area, usually facing the microphones. An algorithm that is based mainly on the above criterion is here described. The algorithm assumes a viewing zone within which a speech source is considered to be wanted-speech and signals coming from outside this zone are considered noise. This approach has the advantage of being able to differentiate between wanted-speech and other unwanted-speech signals. The algorithm uses the time delay between signals received at two microphones to estimate the direction of the dominant signal source. This estimate together with an estimate of the coherence function between the two signals are used to determine initial values for word boundaries. The use of the coherence function serves to isolated spatially un-correlated noise and reduce the effect of reverberation. The initial boundary positions are then refined using measures of the signal energy. With the use of a third microphone, the algorithm can be easily modified to detect the source position rather than direction. The time delay between the signals received at the two microphones; x1, and x2 is estimated using the maximum likelihood (ML) estimator method described in [12]. In the ML method the estimated time delay is defined as the time τ at which the generalised cross correlation function $R^{(g)}_{y1y2}(\tau)$ is maximum, where:

$$R^{(g)}_{y1y2}(\tau) = \int_{-\infty}^{\infty} \psi_g(f) G_{x1x2}(f) e^{j2\pi f\tau} df \qquad (1)$$

where

$$\psi_g(f) = \frac{|\gamma_{x1x2}(f)|^2}{|G_{x1x2}(f)|\left[1-|\gamma_{x1x2}(f)|^2\right]} \qquad (2)$$

Where γ_{x1x2} is the coherence function between $x_1(t)$, and $x_2(t)$ is defined as:

$$\gamma_{x1x2}(f) = \frac{G_{x1x2}(f)}{\sqrt{G_{x1x1}(f)G_{x2x2}(f)}} \qquad (3)$$

Where G_{x1x2} is the cross spectral density between the two input signals, while G_{x1x1}, and G_{x2x2} are the autospectral densities. All spectral densities were estimated using time averaging by a simple recursive formula [13].

Valid speech was assumed when:

Estimated time delay $\leq T_{max}$ (5)

and Estimated coherence $\geq C_{min}$ (6)

Where T_{max} is fixed and is calculated based on the desired viewing zone, and C_{min} is a function of an average of the signal energy E_{av} calculated as:

$$E_{av} = \lambda * E_{av} + (1-\alpha) * E(m). \qquad (7)$$

where m is the frame index, and α is a forgetting factor ($0 \leq \alpha \leq 1$). The speech endpoints are then refined by adding a ‘Head’ and ‘Tail’ frames at the beginning and the end of the estimated wanted-speech. The number of these frames is in turn functions of E_{av}.

4. EXPERIMENTS AND RESULTS

A number of experiments were conducted to evaluate the performance of the proposed algorithm. Speech and noise material from Noisex-92 corpus were used. The speech utterances were a sequence of digits that differs from one speaker to the other. Two sets of tests were conducted. In the first set, the noisy speech was recorded in a medium size room (6*8*5m) using two omni-directional microphones 20 cm apart. Three recordings were made: the first one represents a typical factory unit with factory noise sources distributed in the room, an unwanted-speech source and a wanted-speech source. In the second recording, car noise was played in the background with one wanted-speech source, the SNR was measured for this recording and was found to be -14.5dB (segmental with all silent periods excluded from the calculations). The third recording represents a typical office environment with one wanted-speech and two unwanted-speech sources in addition to noises from other office equipment.

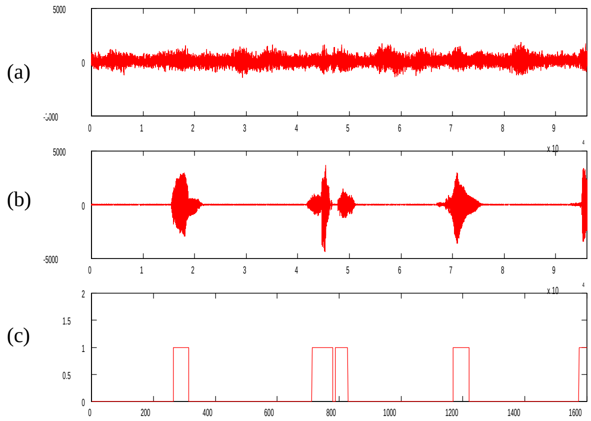

Fig 1 Factory background noise experiment. (a) noisy speech. (b) original wanted-speech. (c) output of the speech detector

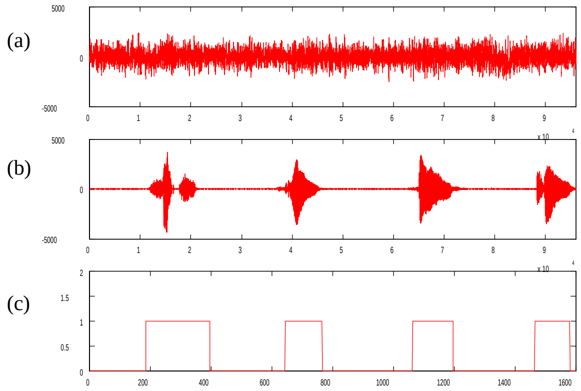

Fig 2 Car background noise experiment. (a) noisy speech. (b) original wanted-speech. (c) output of the speech detector

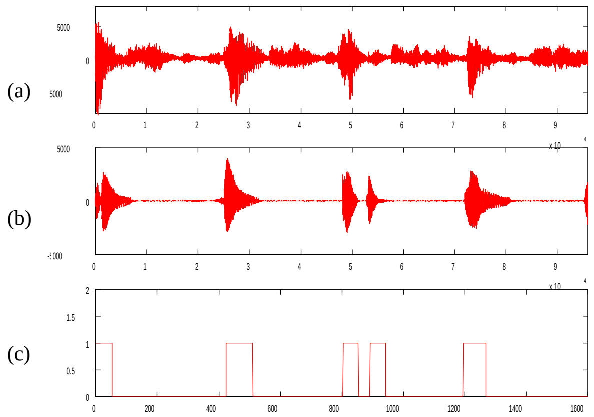

Fig 3 Office environment experiment. (a) noisy speech.
b) original wanted-speech. (c) output of the speech detector

In the second set of tests, speech and noise signals were mixed using simulation software that takes into account the position of the signal sources, as well as, reverberation. Three noise sources were used: 2 speech babble noise, and one factory noise distributed in a 6x8x6 m room. Noises were added at 6 levels of SNR ranging from 18dB to -12dB, in 6dB steps. A total of 24 utterances were used with each utterance comprising 4 words spoken discretely. In all experiments, both the time delay and the coherence function between signals received at the two microphones were estimated with 10 msec overlapping frames (62.5% overlap). Figures 1-3 show the results of the first set of experiments. Alternatively, Table 1 presents a summary of the results of the second set. For each word, the average and standard deviation of the endpoint estimates at the various SNR's are reported alongside with manual labelling. Endpoints are estimated in 'no. of frames' with each frame corresponding to 3.75msec.

	word no.1		word no.2		word no.3		word no.4	
	Start	End	Start	End	Start	End	Start	End
Speaker 1								
Manual	91	197	530	682	952	1102	1393	1546
Auto Av.	85	190	533	697	963	1094	1392	1562
stdev	7.9	12	12	19	9.8	7.3	7.2	11
Speaker 2								
Manual	81	192	515	633	928	1076	1362	1520
Auto Av.	76	179	504	617	955	1065	1376	1510
stdev	12	8.7	12	6.6	11	10	8.8	7.9
Speaker 3								
Manual	100	222	495	645	893	1031	1288	1441
Auto Av.	103	196	490	663	899	1018	1289	1423
stdev	2.4	15	6.1	9	10	18	4.4	30
Speaker 4								
Manual	111	210	494	587	895	1020	1298	1417
Auto Av.	104	199	489	583	902	1020	1299	1405
stdev	1.1	14	5.5	9.3	9.3	16	8.9	7.1

Table 1 Average and standard deviation for endpoint estimates for the second set of experiments.

The results of the first set of experiments show that all the speech parts were detected and no false detection occurred even under the most severe noise conditions. Gaps in multi-syllable words were detected as non-speech which is suitable for speech enhancement applications. If these gaps are to be considered as speech, e.g. for speech recognition applications, some statistics about the on-off pattern of speech [14] can be used for that purpose as in [2] and [3]. The results of the second set of experiments show that automatically estimated endpoints are very close to those obtained manually, and in general less than 70ms different from the manual labels in the various noise conditions.

5. CONCLUSION

A word boundary detection algorithm is presented that is based on the position of signal source. The algorithm was tested in a variety of noise conditions including competing talkers in both real and simulated environments with SNRs down to -14.5dB and proved to be both robust and accurate. Moreover the algorithm can be implemented in real-time applications. These properties make it suitable for the majority of word boundary detection applications in areas such as speech enhancement, and speech recognition.

6. REFERENCES

1. J-C Junqua, *"Robustness and cooperative multimodel man-machine communication applications"*, Proc. Second Venaco Workshop and ESCA ETRW, Sept. 1991.
2. L.F. Lamel et al., *"An Improved Endpoint Detector for Isolated Word Recognition"*, IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-29, No. 4, pp 777-785, 1981.
3. M.H. Savoji, *"A Robust Algorithm for Accurate Endpointing of Speech Signals"*, Speech Commun., Vol. 8, pp 45-60, 1989.
4. J-C Janqua et al., *"A Robust Algorithm for Word Boundary Detection in the Presence of Noise"*, IEEE Trans. Speech, Audio Proc., Vol.2, No.3, pp 406-412, July 1994.
5. R. Tucker, *"Voice Activity Detection Using a Periodicity Measure"*, IEE Proc.-I, Vol.139, No. 4, pp 377-380, August 1992.
6. R. Le Bouquin and G. Faucon, *"Study of a voice activity detector and its influence on a noise reduction system"*, Speech Commun. Vol. 16, pp 245-254, 1995.
7. L.R. Rabiner and M. R. Sambur, *"Application of an LPC Distance Measure to the Voiced-Unvoiced-Silence Detection Problem"*, IEEE Trans. Acoust, Speech, Signal Proc., Vol. ASSP-25, No. 4, pp 338-343, 1977.
8. J. A. Haigh and J. S. Mason, *"A Voice Activity Detector Based on Cepstral Analysis"*, Proc. EUROSPEECH '93, pp 1103-1106, Berlin, Germany, 1993.
9. C. Tsao and R. M. Gray, *"An Endpoint Detector for LPC Speech Using Residual Error Ahead for Vector Quantization Applications"*, Proc. ICASSP-84, pp. 18B.7.1-18B7.4, San Diago, CA, USA, 1984.
10. B.S. Atal and L.R. Rabiner, *"A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition"*, IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-24, No. 3, pp 338-343, June 1976.
11. J. Kates, and M.R. Weiss**,** *"A comparison of Hearing -aid Array-Processing Techniques"*, J. Acoust. Soc. Am., Vol. 99, No. 5, pp. 3138-3148, 1996.
12. C.H. Knapp, and G.C. Carter *"The Generalized Correlation Method for Estimation of Time Delay"*, IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-24, No. 4, pp 320-327, Aug 1976.
13. J.B. Allen et al., *"Multimicrophone Signal Processing Technique to Remove Room Reverberation From Speech Signals,"* J. Acoustic Soc. Amer., Vol. 62, No. 4, pp 912-915, 1977.
14. P.T. Brady, *"A Technique for Investigating On-Off Patterns of Speech,"* Bell Syst. Tech. J., Vol. 44, No.1, pp. 1-22, Jan. 1965.