A ROBUST WORD BOUNDARY DETECTION ALGORITHM WITH APPLICATION TO SPEECH RECOGNITION

H. Agaiby, T. J. Moir

Department of Electronic Engineering and Physics, University of Paisley, High Street,
Paisley, PA1 2BE, Scotland, UK
e-mail:hany@diana22.paisley.ac.uk

ABSTRACT

A new robust word boundary detection algorithm is described in this paper that performs well under a variety of noise conditions including competing talkers. The algorithm uses the direction of the signal as the main criterion to differentiate between wanted-speech and background noise. A 'viewing zone' is assumed within which a speech source is considered desired-speech and signals coming from outside this zone are considered noise. The algorithm uses the time delay between signals received at two microphones to estimate the direction of the dominant signal. This estimate together with an estimate of the coherence function between the two signals as well as measures of the signal energy are used to determine word boundaries. Two state-of-the-art speech recognisers were used to evaluated the performance of the algorithm. For each recogniser, the recognition accuracy is measured with manually labelled noisy speech and compared when speech is automatically processed using the proposed algorithm. The results showed that the algorithm performs as well as manual labelling under signal-to-noise ratios as low as 0dB.

1. INTRODUCTION

The accurate detection of word boundaries is crucial to the performance of speech recognisers. 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. However, most of the word boundary detection algorithms proposed in the literature impose a number of restrictions on the type and level of background noise that are impractical in many applications. Algorithms based on signal energy, for example [2], and [3] assume an almost noise free environment. On the other hand algorithms targeting discontinuous transmission applications (DTX), e.g. [4], and [5], while more robust in the presence of noise, do not exhibit the accuracy required for speech recognition. None of the above algorithms or other word boundary detection algorithms such as [6], and [7] can differentiate between wanted-speech and competing talkers. Nevertheless the variety of environments in which speech recognition systems work requires a word boundary detector that is both flexible and robust enough to work with various background noise types and levels while maintaining the high level of accuracy required for speech recognition. The objective of the algorithm presented in this paper was to fulfil as much as possible the above requirements while maintaining the computational simplicity needed for real-time operation. The algorithm uses the direction of the signal as the main criterion to differentiate between wanted-speech and background noise. Two state-of-art speech recognisers were used to evaluate the performance of the proposed algorithm. The recognisers accuracy was first measured using manually labelled noisy speech and the results compared against those obtained when the noisy speech is automatically processed using the word boundary detection algorithm.

2. ALGORITHM DESCRIPTION

The main criterion used to differentiate between wanted-speech and background noise in the proposed algorithm is the direction of the signal. The selection of this criterion rather than criteria used by other algorithms such as signal energy, periodicity, zero crossing, coherence function, stems from the simple observation that in most of speech recognition applications the position of the user can be considered to be within a predefined area facing the system sensors. Even if the position of the user is initially outside the predefined area it is, in most of the cases, easier for the user to adjust his/her position than to control the types or levels of background noise. Moreover, the use of this criterion enables the algorithm to differentiate between wanted-speech and other unwanted speech signals in what is commonly called 'cocktail party' situation. The algorithm is based on the use of only two microphones and is therefore only able to estimate the direction of the signal source and not its position. Nevertheless, experiments showed that using the direction of the signal as a main criterion together with estimates of the coherence

0-7803-4137-6/97/$10.00©1977 IEEE

function and signal energy were quite adequate in many speech recognition applications. The algorithm uses the time delay between signals received at two microphones 20cm apart 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. These initial values are then refined using measures of the signal energy.

The algorithm processes the speech in overlapping frames of 10msec with 62.5% overlapping. The relatively short frame length gives the algorithm the potential of real time implementation. The frames are first windowed using a hamming window, then an estimate of the time delay T_d between the two input signals $x_1(t)$, and $x_2(t)$ is calculated using the maximum likelihood (ML) estimator method [8]. 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

$$\psi_g(f) = \frac{|\gamma_{12}(f)|^2}{|G_{x1x2}(f)|\left[1-|\gamma_{12}(f)|^2\right]} \qquad (2)$$

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 [9]:

$$G_{xixj}(f,m) = \lambda\gamma_{xixj}(f,m-1) + (1-\lambda)X_i(f,m)\,X_j^{*}(f,m)$$

$$i, j = 1, 2 \qquad (4)$$

where: m is the frame index
λ is a forgetting factor ($0 \le \lambda \le 1$)

The algorithm decides on a valid speech when both the estimated time delay and the coherence function are above predefined thresholds:

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

and Estimated coherence $\ge \gamma_{min}$ (6)

The time delay T_{max} is fixed and calculated based on the estimated position of the speaker, and γ_{min} is a function of a measure of the average signal energy E_{av} calculated from:

$$\gamma_{min} = \gamma_0 + (c_1 - E_{av})/c_2 \qquad (7)$$

with:

$$E_{av} = \lambda * E_{av} + (1-\lambda) * E(m) \qquad (8)$$

where m is the frame index, c1, c2 are constants and λ is a forgetting factor ($0 \le \lambda \le 1$). The time delay condition in (5), creates a viewing zone opposite to the system sensors while the coherence function condition (6) reduces the effect of reverberating signals and uncorrelated noise within the viewing zone. The word boundaries are then refined by adding 'Head' and 'Tail' frames at the beginning and the end of the estimated wanted-speech respectively. The 'Head' and 'Tail' frames are to account for the weak beginnings and ends of the words dominated by high level of noise. These are in turn linear functions of E_{av} such that at low levels of noise few or no frames are added and the number of frames increases with the level of noise.

3. EVALUATION

A number of experiments were conducted to evaluate the performance of the proposed algorithm. In these experiments noisy speech was applied to two state-of-the-art speech recognisers [10] and the recognition accuracy measured. The noisy speech was then processed first using manual labelling then automatically using the proposed algorithm. In both cases the word boundaries are labelled and noise alone periods replaced by silence. The processed speech in each case was then applied to the two recognisers and the recognition accuracy measured.

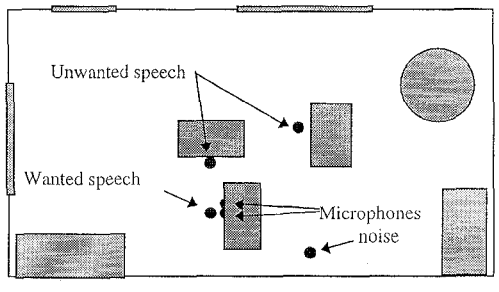
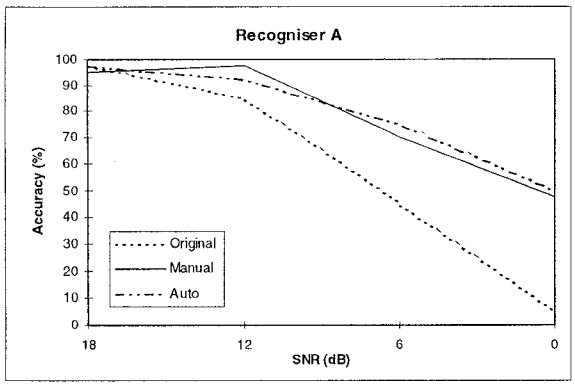


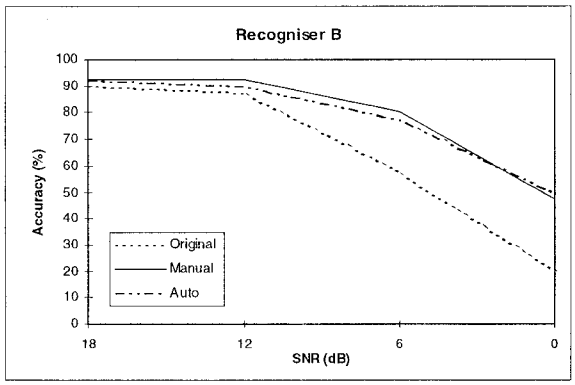
Figure 1 Arrangement for office environment experiment

For this purpose speech and noise material from Noisex-92 corpus were mixed to produce a number of noisy speech utterances. The speech utterances were the sequence of digits (0-9) while noises types were factory noise, and speech babble noise. Speech and noise signals were mixed using simulation software that takes into account the position of the signal sources as well as reverberation. The experiments can be divided in two sets: the first set of experiments was designed to

simulate an office of 6x8x6m (WxDxH) (figure 1) with one desired-speech source, two undesired speech sources and a factory noise source. The noises were added to the speech in four SNR levels of 18dB, 12dB, 6dB, and 0dB (the SNR calculation was segmental with all silent periods excluded). In the second set of experiments a factory unit of 10x15x6m (WxDxH) was simulated with one desired speech source opposite to the microphones and four noise sources distributed in the room namely: two undesired speech sources and two factory noise sources. Again, noises were added in four levels of segmental SNR levels of 18dB, 12dB, 6dB, and 0dB . In total, 160 words were used in the two sets of experiments; 40 words at each noise level.



(a)



(b)

Figure 2 Speech Recognition Performance for the original, manually processed, and automatically processed noisy speech for (a) recogniser A (b) recogniser B.

The unprocessed and processed noisy speech utterances were applied to the two recognisers and the resulting recognition accuracy calculated. Figure 2 depicts the recognition accuracy against the SNR of the noisy speech for the two recognisers in the three cases namely; using unprocessed speech, using manually labelled speech and using the proposed algorithm. The results show that for both recognisers the recognition accuracy using automatic word boundary detection is almost the same as that obtained with a manually labelled speech for both recognisers and with SNR down to 0dB. A significant improvement in the recognition accuracy was achieved at almost every SNR level with up to 45% for recogniser A, and 30% for recogniser B at 0dB.

4. CONCLUISON

An automatic word boundary detection algorithm was presented and shown to perform as well as manual labelling in noisy environments with various types of noises at SNR down to 0dB. The algorithm is capable of differentiating between desired-speech and other types of noise including competing talkers, and can be implemented in real time.

5. 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. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-29, No. 4, 1981, pp 777-785.
3. M. H. Savoji, "A Robust Algorithm for Accurate Endpointing of Speech Signals," *Speech Communication*, Vol. 8, 1989, pp 45-60.
4. D. K. Freeman et al, "The Voice Activity Detector for the Pan-European Digital Cellular Mobile Telephone Service," *ICASSP 1989*, Vol. I, pp. 369-372.
5. R. Trucker, "Voice Activity Detection Using a Periodicity Measure," *IEE Proc.-I*, Vol. 139, No. 4, August 1992, pp 377-380.
6. R. Le Bouquin and G. Faucon, "Study of a Voice Activity Detector and its Influence on a Noise Reduction System," *Speech Communication,* Vol. 16, 1995, pp 245-254,.
7. J-C Janqua et al., "A Robust Algorithm for Word Boundary Detection in the Presence of Noise," *IEEE Trans. on Speech and Audio Processing,* Vol.2, No.3, July 1994, pp 406-412.
8. C. H. Knapp, and G. C. Carter "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, No. 4, Aug 1976, pp 320-327.
9. J.B. Allen et al., "Multimicrophone Signal Processing Technique to Remove Room Reverberation From Speech Signals," *J. Acoustic Soc. Amer.*, Vol. 62, No. 4, 1977, pp 912-915.
10. H. Agaiby et al., "Commercial Speech Recognisers Performance under Adverse Conditions, A Survey," *Proc. ESCA-NATO workshop on Robust Speech Recognition For Unknown Communication Channels*, Pont-A_Mouson, France April 1997.