An Entropy based Robust Speech Boundary Detection Algorithm for Realistic Noisy Environments

Kim Weaver, Khurram Waheed and Fathi M. Salem
Circuits, Systems and Neural Networks Laboratory
Michigan State University
East Lansing, MI 48824-1226

Abstract-**This paper addresses the issue of automatic word/sentence boundary detection in both noiseless and noisy backgrounds. We present our proposed speech boundary detection algorithm using a time-domain entropic contrast function. The entropic contrast exhibits well-behaved characteristics as compared to energy-based methods resulting in immunity to endpoint cut-off issues for the latter. This algorithm is capable of estimating the speech boundaries both in noiseless and distinctive noise backgrounds such as a fan, car engine, radio etc. For the case of wide-spectrum colored background noise such as jazz, opera, songs, rock music etc., we further propose a modification in the preprocessing stage by incorporating a frequency-weighting scheme to emphasize the speech contents. This improved scheme provides proper speech segmentation even in the presence of wide-spectral background noise with no change in the computational cost versus our earlier proposed algorithm. A complete time-domain implementation is sought due to its lower computational burden and its suitability for real-time implementations using DSPs, FPGAs, ASICs etc. The algorithm improves the accuracy of word boundary estimates by a factor of at least 25% for the case of isolated (and 16% for connected) speech. For continuous speech, the algorithm can determine sentence boundaries thus allowing for power efficient implementation of speech recognition engines by rejecting extended periods of silence.**

I. INTRODUCTION

Automatic segmentation of speech in real-world noisy surroundings is a very challenging problem. For isolated word recognition for a limited vocabulary, this implies determination of correct isolated word boundary and rejection of extra-aural artifacts such as breath, mouth and lip clicks etc. For the connected speech case, the problem is to get rid of inter-word silences and any other artifacts as mentioned in the previous case. In the case of a continuous speech recognition engine, an efficient speech segmentation pre-processor can significantly reduce the computational load and power consumption of the recognition engine.

More importantly if the speech segmentation algorithm is inefficient, it deteriorates the accuracy of the overall speech recognizer. In a quiet laboratory environment, the problem of speech segmentation is one of merely picking out the sound from the relative silent background of the rest of the recording/utterance. This however, is not very practical for real-world use.

To be truly effective, an algorithm needs to be robust enough to detect speech even when accompanied by extraneous sounds such as traffic, music, machinery etc. In situations of high noise intensity, speech is much harder to detect. Some of the other proposed algorithms [4, 6 and the references therein] can adequately perform separation of speech from some background noise, but they are computationally intensive and as such cannot be implemented in real-time scenarios unless powerful DSP based systems are employed. Other less computationally intense algorithms [2, 3] are not as accurate at finding true word boundaries and thus impede the ability to recognize speech.

The most commonly used method of endpoint detection is the use of short-time or spectral energy [1-5] followed by a threshold stage. This technique is not very robust against various speech artifacts and tends to cut off the endpoints in some words with fricative sounds even in quieter surroundings. The energy being very sensitive to the amplitude of the speech signal does not yield good speech classification results in noisy environments. However, infusion of pitch and duration information, use of adaptive thresholds, augmentation of zero crossover rate result in somewhat improved performance [4].

A newer promising approach involves the use of entropy to find endpoints [6, 7]. The main advantage of an entropy profile is its reduced sensitivity to the changes in the amplitude of the speech signal, while exhibiting more sensitivity to the presence or absence of a signal. Therefore, using an entropic contrast and an appropriate pre-processing stage for the noisy speech signal, it is possible to design an efficient speech boundary detection algorithm with minimal complexity, yet maximizing accuracy. Further, to make the scheme suitable for real-time hardware implementation, it is desirable to have minimum computational load.

II. WEIGHTED ENTROPY-BASED SPEECH SEGMENTATION ALGORITHM

The proposed algorithm uses entropy of the speech as the key feature for boundary detection. This computation of the entropy estimate is carried out directly in the time domain. The algorithm has been structured; see Fig. 1, so as to provide a drop-in replacement of the energy-based boundary

0-7803-7898-9/03/$17.00 ©2003 IEEE

detection algorithms, while minimizing speech cut-off and artifact inclusion problems.

A. Main Algorithm Implementation

In this section, we illustrate practical implementation of the algorithm for the case of isolated word recognition. The same scheme can be easily extended for other types of recognition engines by a change of the decision parameter values as discussed in this section.

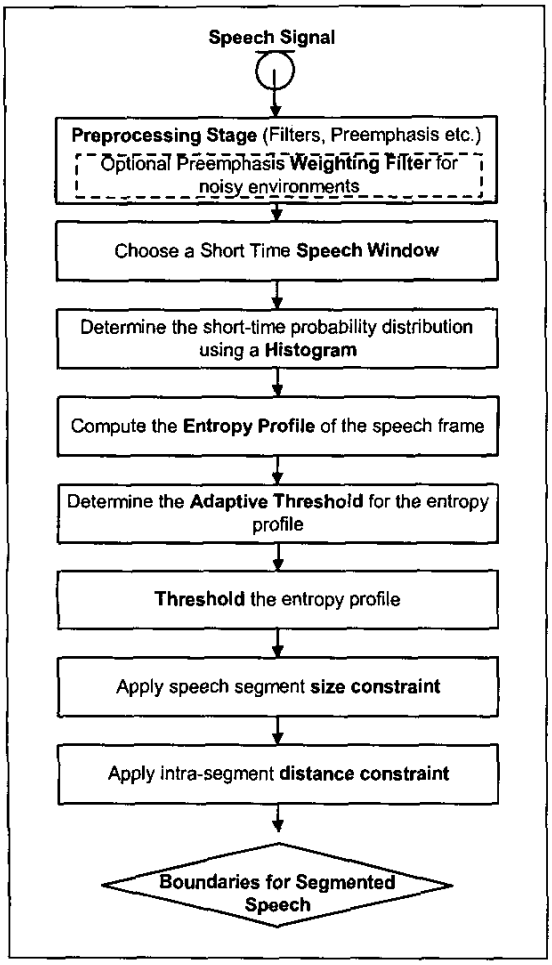


Fig. 1. Block Diagram of the Proposed Speech Segmentation Algorithm

The original incoming speech data is first pre-processed using a pre-emphasis filter. The function of this pre-emphasis is to reduce the effects of the glottal pulses and radiation impedance. It also takes the focus to the spectral properties of the vocal tract [1]. This is followed by a band-pass filter, which removes constant and low frequency components of background, as well as high frequency noise and speech harmonics created due to sampling and nonlinearity of the recording device. The pre-processed speech is divided into overlapping frames with each frame containing approximately 25 milliseconds of speech. These frames for entropy estimation typically have a 25-50 percent overlap.

In order to estimate the probability distribution within each individual frame, a histogram with N bins is constructed. The histogram is then normalized to satisfy the statistical properties of the cdf. Selection of the number of bins (N) for the histogram is a trade-off between sensitivity and computational load. Generally N may be chosen in the range 50-100. The entropy for each frame is then computed as follows [5].

$$H = -\sum_{k=1}^{N} p_k \log p_k \,. \tag{1}$$

where p_k is the normalized mass of the k^{th} bin in the histogram.

The algorithm can be easily implemented online with a unit frame delay, however for clarity of presentation, we assume that we have the entropy profile ξ for the complete speech data available,

where

$$\xi = [H_1 \quad H_2 \quad \cdots \quad H_m] \tag{2}$$

where m is the total frames in the incoming speech.

This entropy profile can then be used to find an appropriate threshold γ to determine the existence of speech regions within the entire speech data. An appropriate threshold in this case, will take the form

$$\gamma = \frac{\max(\xi) - \min(\xi)}{2} + \mu \min(\xi); \quad \mu > 0 \tag{3}$$

The threshold is thus chosen a little higher than the mean entropy profile. Note that the minimum of the profile is a measure of the remnant backdrop noise floor. A selection of threshold as mentioned above minimizes excessive influence of the background noise. Once a threshold has been determined, anything over the threshold is considered to be speech, and anything below the threshold is either silence or noise. i.e.,

$$\xi' = \begin{cases} \xi_i & \text{if } \xi_i \geq \gamma \\ 0 & \text{otherwise} \end{cases}; \quad i=1,2\ldots m \tag{4}$$

In many cases due to possibility of a number of artifacts, parts of the non-speech data are falsely reported as speech; therefore it is necessary to use other classification criteria to eliminate these incorrect results.

The first criterion is the size of the determined speech segment. The thresholded entropy may have several candidate regions, which are not actually speech but relate to some extra-aural or background artifacts. Humans generally do not produce very short duration sounds. Therefore each speech segment should have a certain minimum length λ_i, see Fig. 2. For any i^{th} speech segment

$$\lambda_i = e_i - s_i \tag{5}$$

where

s_i - corresponds to the starting point of the i^{th} frame in the thresholded entropy profile ξ'

e_i - corresponds to the ending point of the i^{th} frame in the thresholded entropy profile ξ'

The lambda corresponds to the shortest phoneme or phone in the vocabulary of the recognition engine and is also a function of the sampling frequency.

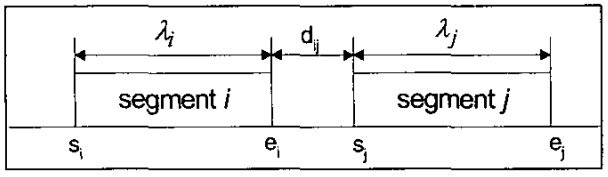


Fig. 2. Relationship between adjacent speech segments

The second criterion is based on the thresholded inter-segment distance d_{ij}. This criterion is required because frequently there may be spurious speech segments that satisfy the first criterion. Also there may be parts of speech, which were separated into separate segments due to the sound or pronunciation of a particular phoneme or phone especially for sub-vocal sounds. It will be necessary to merge two such speech segments into one larger segment. This happens frequently with words. This final test involves a series of criteria, some of the more important ones are

- if $\lambda_i < \kappa$ and $d_{ij} > \delta$, then the i^{th} segment is discarded,
- if $\lambda_j < \kappa$ and $d_{ij} > \delta$, then the j^{th} segment is discarded,
- if $(\lambda_i$ or $\lambda_j) > \kappa$, $d_{ij} > \delta$ and $\lambda_i + \lambda_j < \theta$, then the two segments are merged, and anything between the two segments that was previously ignored, is made part of the speech.

Experimental verification of the algorithm has shown that the proposed setup has a much better ability to detect speech and discard background noise than the counterpart energy algorithms.

B. Extension for Multi-spectral Background Noise

The above instantiation of the algorithm is capable of speech segmentation in environments with no or narrowband frequency separated background noise. While this technique gives better performance than energy based algorithms [1-4], it is still not very robust to real-life situations where the backgrounds can be richer in their spectral content. Therefore, to design a more stable algorithm, it is necessary to incorporate a weighting mechanism to emphasize frequencies that are more common to human speech than backdrop noise.

One way of implementing such an algorithm is presented in [6] where the speech is transformed to the frequency domain using the FFT operation on a sliding time window. Then while in the frequency domain, bandpass filtering is done by discarding some of the frequency bins both below and above the desired speech frequencies. Further, a weighting approach is used to compute the entropy of this truncated FFT sequence. This algorithm gives desirable performance but requires a DSP processor for implementation due to the computational load of computing the FFT. Furthermore, the resolution and efficiency of the algorithm depends on the number of bins used for FFT computation.

Our approach is to combine both the bandpass filtering and the weighting operation for entropy computation into a single filter design problem. This results in computation of a similar number of filter co-efficients as required for the bandpass filtering alone in the primary algorithm, thus maintaining the same operational cost of the algorithm. This design methodology keeps the overall implementation of the algorithm simple and amenable for real-time dedicated hardware implementation using FPGAs, ASICs etc.

The weighting filter, see Fig. 3, is designed to have its peak close to 1 kHz. The weighting filter is also designed to minimize the effects of both low frequency and high frequency noise. In order to practically achieve the weighting characteristics as shown, a combination of separate highpass and a lowpass filters were used. Both the highpass and the lowpass filters were designed as elliptical filters. The final filter is achieved by convolving the design of both component filters.

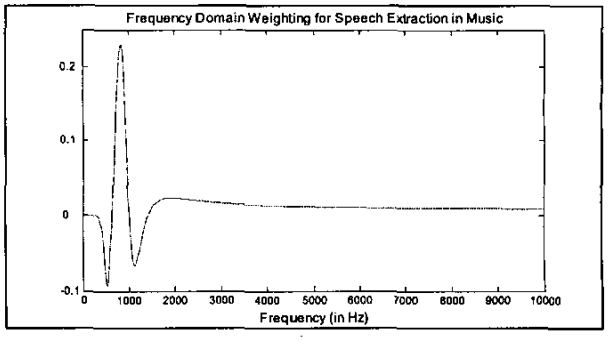


Fig. 3. Graph of the weighting filter in the frequency domain

The effectiveness of this filter lies in the fact that there is no need to transform the signal into the frequency domain for its weighted emphasis. This means that there is no need for FFT calculation, which requires large amounts of memory and processing time.

The computational complexity of the various algorithms is an important decision factor in choosing which algorithm to use. Each speech sample of length S, is broken up into N frames each of length L, which is a function of the sampling frequency. Discarding the commonalities among the various algorithms. The time complexities of the FFT and the time domain algorithm are shown in Table 1, where α, β, μ are

positive constants which depend on the processing power of the platform the algorithms are implemented on. It is evident that for longer speech utterances, the computational time for the frequency domain increases exponentially making the time-domain implementations desirable.

Algorithm	Time complexity
Frequency domain method	$T(L) = \mu NL \log_2(S/L)$
Time domain method	$T(L) = \alpha NL + \beta$

Table 1. Time complexity comparison of the frequency-domain and time-domain entropic contrast algorithms.

It is clear from Table 1, that the time domain algorithm has a reduced computational load as compared to the frequency domain alternative. This advantage becomes significant as the sampling frequency S is increased for better speech fidelity and/or harmonic accuracy, which also may result in a larger window size L. The complexity of the FFT algorithm in such cases makes it impractical for low-cost real time processing.

III. SIMULATION RESULTS

In this section, we present several simulations to illustrate the effectiveness of using entropic contrast for speech boundary detection.

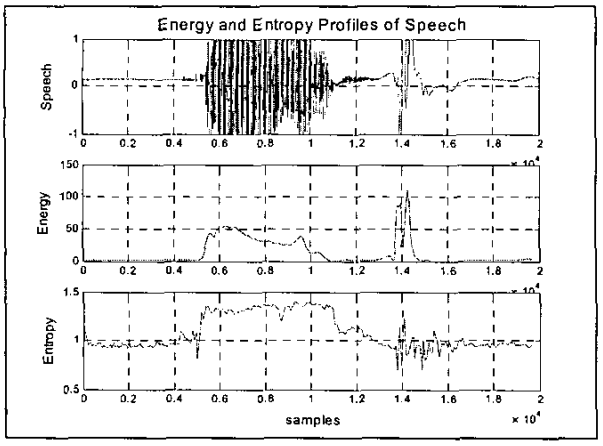


Fig. 4. Digit "five", speech data and corresponding energy and entropy profiles

First of all, we discuss the case of a single word utterance and illustrate the differences in the characteristics of the energy and the entropy profiles of speech data. Fig. 4 shows the speech data for utterance of digit *five*. It is evident that the energy profile has a higher variance as compared to the corresponding entropy profile. Further, the energy falls off towards the end of the speech sample and then rises again towards the end due to the sound "-ve". This makes the use of energy-based techniques difficult for automatic endpoint detection because the lowered level of energy may last long enough to suggest that there is no speech, which leads to false endpoints. The entropy profile on the other hand has smaller variance and is more sensitive to presence of speech. Thus use of entropy results in potentially fewer false endpoints. The overall estimation accuracy is determined by the post-processing stage of the thresholded entropy profile.

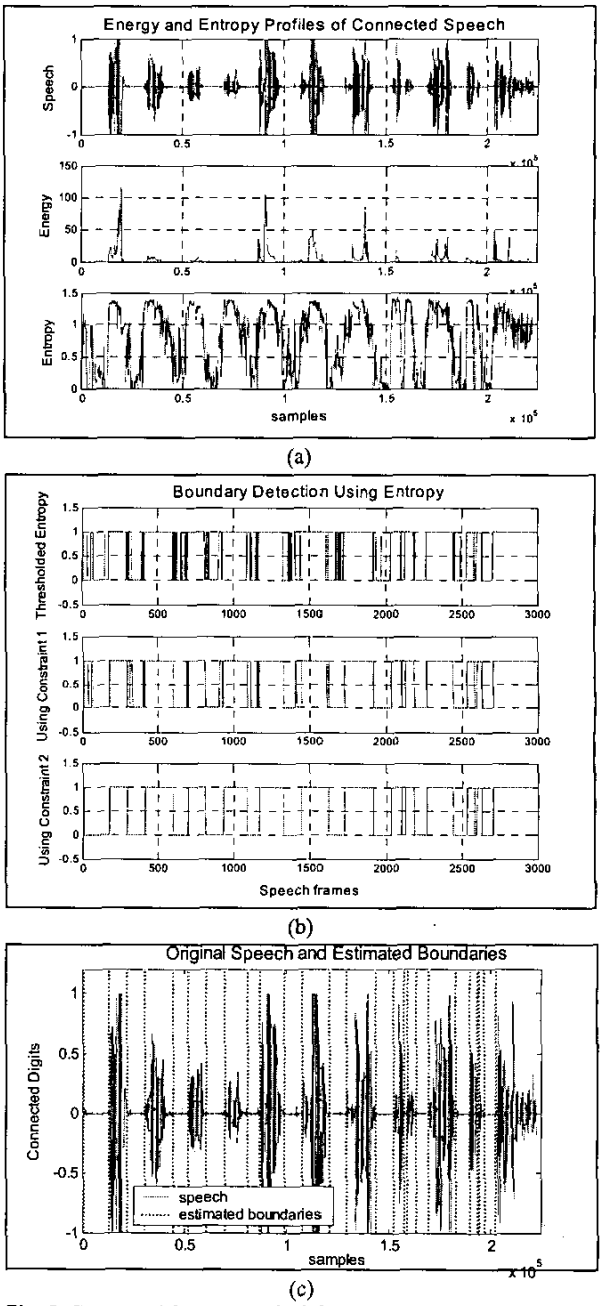
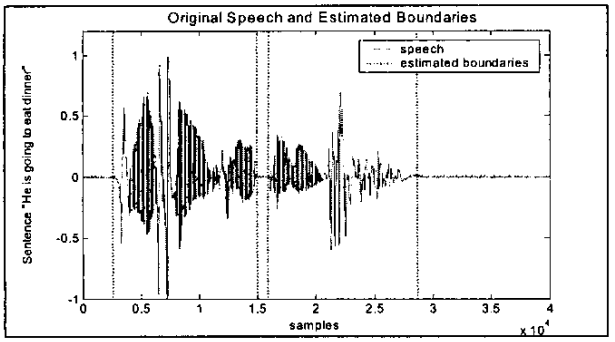


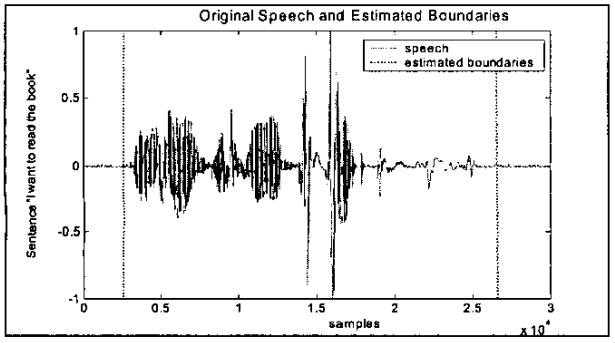
Fig. 5. Connected Sequence of Digits: (a) Speech data, energy and entropy profiles, (b) Thresholded entropy profile and application of post-processing constraints, (c) Identified speech boundaries

In order to demonstrate the word boundary detection algorithm using the entropy-based contrast, we now choose a more interesting case of connected words. Presented below are the simulation results for a sequence of connected digits (0, 1, 2...9, oh) spoken at a slower pace as may be used for speech activated car phones in relatively noiseless background.

As remarked earlier, in Fig. 5.a the entropy profile seems more suited for determination of constituent digits. Fig. 5.b shows application of boundary detection criteria, which are qualitatively similar to those explained in the previous section. Fig. 5.c shows the speech data again with the estimated boundaries, which seem to be in good harmony.



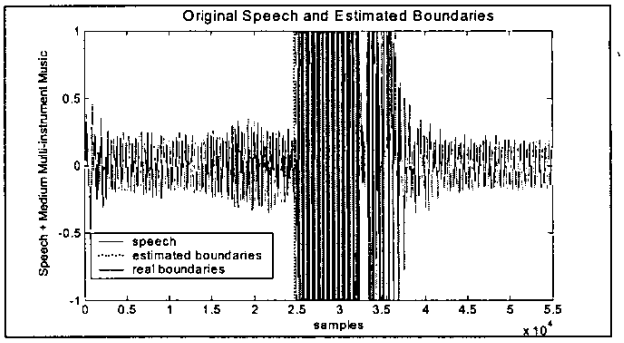
(a)



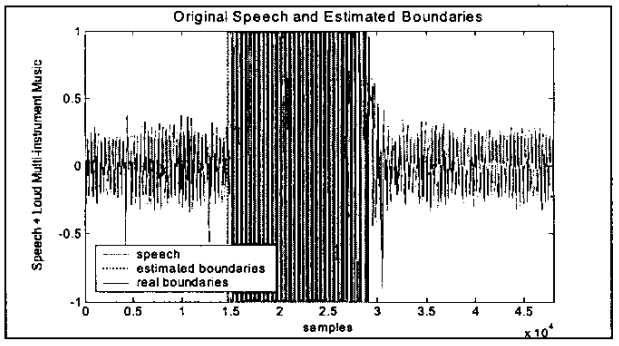
(b)

Fig. 6. Sentence Boundary Detection for the sentence (a) "He is going to eat dinner", (b) "I want to read the book"

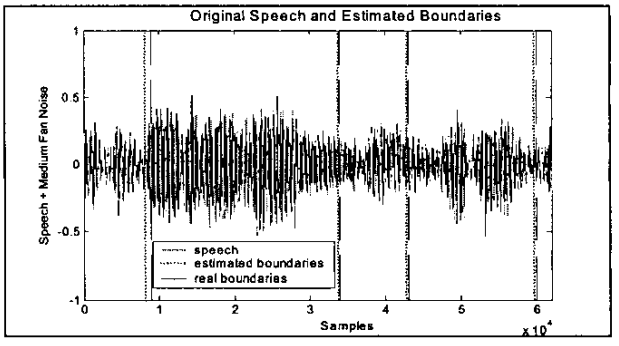
The case of continuous speech, however is different. When speaking continuous sentences, people tend to slur one word into another. The human ear is still able to differentiate among the words in the sentence, but the corresponding problem for an automatic speech recognition engine is quite difficult. The faster the spoken speech is, the less clear it becomes to determine where one word ends and the next begins. The recognition engines for continuous speech and large vocabulary systems are based on HMM models. These engines typically work on smaller units of speech such as the phones or the phonemes and each word or sentence becomes a sequence of these smaller building units. In practical situations this continuous speech can be intermittent, e.g., consider the cases of different people using a bank ATM, a voice operated PC software, a multimedia station, or an interactive toy etc. In these cases the proposed algorithm can be incorporated to determine sentence boundaries and reject periods of silence, see Fig. 6. Please note, that the criteria for post-processing the entropy profile in this case can be slightly more involved. The weighting filter may require tuning based on whether the speech sample consists of a single word, a series of connected words or continuous speech.



(a)



(b)



(c)

Figure 7. Boundary detection of speech in a noisy car environment: (a) medium level music noise, (b) loud music noise, (c) medium car fan and low music noise

It is certainly desirable for a speech segmentation algorithm to deliver in all of these different speech recognition problems. However, it is most important for such an algorithm to be able to perform in noisy environments. Most real world situations requiring speech recognition

systems operate in environments that are less than ideal with regards to background noise. The proposed algorithm provides better noise immunity. As a summary, in case of isolated word recognition against white background noise, the algorithm provides an improvement of 9 % in recognition error rate at an SNR of 15 dB, which rises to 14.3% at an SNR of 10 dB and 15.6 % at an SNR of 5 dB, respectively.

Fig. 7 presents test results for the proposed weighted entropy algorithm in a realistic scenario. The speech samples are recorded using a directional microphone in a moving car, driven at 35 mph. The sampling frequency for the speech recordings is 22050 Hz. The speech segments comprise of commands such as issued to a voice-activated car phone. Various realistic scenarios were simulated while the car was being driven. This includes music and radio noise from car stereo, operating the fan blower, wipers, sliding the car windows up and down, talking in the back seat, as well as driving the car at different speeds etc.

Although due to limited space, only a few results are being presented in this paper, it is evident that the proposed algorithms provide accurate speech boundary detection in the presence of a variety of noise backgrounds. Table 2 presents a comparative summary of speech boundary detection algorithms using the entropic-contrast on a batch of 100 utterances. The deviation of the speech boundary in the table below is defined to be

$$\text{Boundary Deviation} = 100\left((estimated - actual)/f_s\right) \quad (6)$$

		Boundary Deviation	
Speech type	**Method**	**Startpoint**	**Endpoint**
Single word	*Frequency domain*	-1.125	-7
	Time domain	1.375	6
	Time domain with filter	1.09375	2.6875
	Actual	0	0
Single word with noise	*Frequency domain*	6.875	0
	Time domain	-3.125	12
	Time domain with filter	4.71875	5.6875
	Actual	0	0
Connected words	*Frequency domain*	74.125	-10.375
	Time domain	-2.375	33.625
	Time domain with filter	-1.25	-2.6875
	Actual	0	0
Connected words with noise	*Frequency domain*	74.125	-10.375
	Time domain	-21.86875	31.875
	Time domain with filter	-4.0625	8.40625
	Actual	0	0

Table 2. Comparison of entropy-based speech boundary detection algorithms.

IV. CONCLUSIONS

A new weighted entropy based speech segmentation algorithm has been proposed. Complete details on how to implement the algorithm have been provided. Several simulation examples have been demonstrated which show advantages of using entropic contrasts in speech boundary detection problems. The algorithm performance for various cases of isolated, connected and continuous speech are discussed. The main advantage of using entropy based speech-background contrast is for the endpoint detection where the energy based methods fail at times due to sub-vocal or fricative sounds. This suggests that using entropy in endpoint algorithms makes it less likely for important speech information from being discarded.

The new entropy-based algorithm alone shows better performance in monophonic noisy environments. The addition of a weighting filter allows for improved performance even in polyphonic noisy environments. This makes the proposed speech boundary detection algorithms even more attractive than both energy based algorithms and algorithms that require frequency domain computations. We believe that through the use of entropy based speech recognition engines; much higher recognition rates can be achieved especially for small to medium sized vocabulary recognizers. For large vocabulary continuous speech systems, the algorithm is useful in rejecting periods of silence, which can be used to implement power efficient 24/7 online recognition engines.

REFERENCES

[1] J. R. Deller Jr., J. L. H. Hansen and J. G. Proakis, "Discrete Time Processing of Speech Signals", *IEEE Press*, NJ, 2000.

[2] A. Ganapathiraju A., L. Webster, J. Trimble, J. Bush and J. Kornman, "Comparison of Energy-Based Endpoint Detectors for Speech Signal Processing," *Proceedings of the IEEE Southeastcon*, Tampa, Florida, USA, pp. 500-503, April 1996.

[3] L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An Improved Endpoint Detector for Isolated Word Recognition", *IEEE ASSP Magazine*, Vol. 29, pp. 777-785, 1981.

[4] J. C. Junqua, B. Mak, and B. Reaves, "A Robust Algorithm for Word Boundary Detection in the Presence of Noise", *IEEE Trans. on Speech and Audio Processing*, Vol. 2, No. 3, pp. 406-412, Apr. 1994.

[5] L. Rabiner and B.–H. Juang: "Fundamentals of Speech Recognition", Prentice Hall, NJ 1993.

[6] J.-L. Shen., J.-W. Hung and L.-S. Lee, "Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments," , *Proc. Int. Conf. on Spoken Lang. Processing*, Sydney ICSLP-98, CD-ROM, 1998.

[7] K. Waheed, K. Weaver and F. M. Salam, "A Robust Algorithm for Detecting Speech Segments using an Entropic Contrast," in *45th IEEE Int'l Midwest Symposium on Circuits & Systems*, Tulsa, OK, August 4-7, 2002, in press.