A ROBUST ALGORITHM FOR DETECTING SPEECH SEGMENTS USING AN ENTROPIC CONTRAST

Khurram Waheed, Kim Weaver and Fathi M. Salam

Circuits, Systems and Artificial 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 quiet and noisy environments. We propose to use an entropy based contrast function between the speech segments and the background noise. A simplified data based scheme of computing the entropy of the speech data is presented. The entropy-based contrast exhibits better-behaved characteristics as compared to the energy-based methods. An adaptive threshold is used to determine the candidate speech segments, which are subjected to word/sentence constraints. Experimental results show that this algorithm outperforms energy-based algorithms. The improved detection accuracy of speech segments results in at least 25 % improvement of recognition performance for isolated speech and more than 16% for connected speech. For continuous speech, a preprocessing stage comprising of the proposed speech segment detection makes the overall HMM based scheme more computationally efficient by rejection of silence periods.

1. INTRODUCTION

The automatic segmentation of speech especially in real-world noisy environments is a challenging problem. Most importantly, the efficiency achieved in automatic detection of speech boundaries largely determines the accuracy of the following recognition engine. Even minor improvement in speech boundary detection front-end greatly influences the overall system accuracy in the long run. For the isolated word recognition in a limited vocabulary scenario, this problem boils down to the determination of the correct isolated word boundary and the rejection of the speech artifacts such as breath, mouth and lip clicks etc. For the connected speech case, the problem is to get rid of intra-word silences and any other artifacts as mentioned in the previous case. For a continuous speech recognition engine, efficient automatic speech segmentation pre-processor can reduce the computational load and power consumption of the engine.

The most commonly used method of endpoint detection is the use of short-time or spectral energy [1,2,3,4]. Typically an adaptive threshold is employed based on the features of the energy profile to differentiate between the speech segments and the background noise. This is not a very good method, as it tends to cut off the ends of some words in quieter surroundings. The energy being very sensitive to the amplitude of the speech signal, will not result in good classification results in noisy environments.

A newer promising approach involves the use of entropy to find endpoints. The main features of an entropy profile include less sensitivity to the changes in the amplitude of the speech signal, which directly results in retention of more detail as compared to the corresponding energy profiles.

In order for a speech recognition system to be truly practical, it must be able to perform satisfactorily in noisy environments. This can be a problem since endpoint detection programs typically fail in the presence of noise. Even in the presence of some noise, the entropy profile retains greater detail and leads to better accuracy as compared to energy-based methods.

2. ENTROPY-BASED SPEECH SEGMENTATION ALGORITHM

Conventional short-time or spectral energy based endpoint detection algorithms are very sensitive to speech artifacts and break down quickly in the presence of noise. Infusion of pitch and duration information, use of adaptive thresholds, augmentation of zero crossover rate result in somewhat improved performance [4].

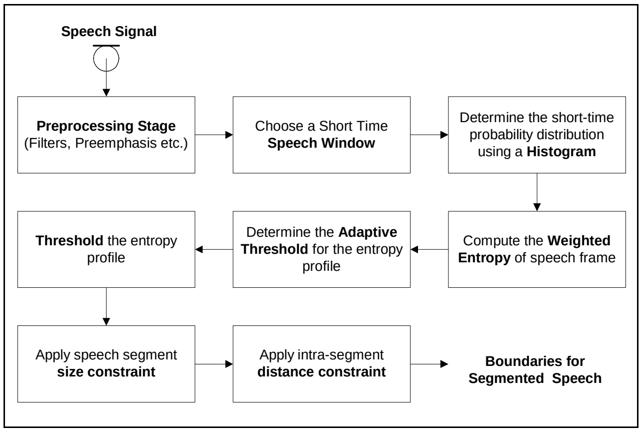

Speech Signal
Preprocessing Stage (Filters, Preemphasis etc.)
Choose a Short Time Speech Window
Determine the short-time probability distribution using a Histogram
Compute the Weighted Entropy of speech frame
Determine the Adaptive Threshold for the entropy profile
Threshold the entropy profile
Apply speech segment size constraint
Apply intra-segment distance constraint
Boundaries for Segmented Speech

Figure 1. Block Diagram of the Proposed Speech Segmentation Algorithm

The proposed algorithms, replaces entropy of the speech as the key feature for boundary detection. This computation of the entropy 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 detection algorithm.

The proposed speech segmentation algorithm gives improved boundary conformance between manual and automatically determined word/speech boundaries in high SNR experiments. This minimizes problems of speech chop-off and inclusion of artifacts. For medium SNR experiments (i.e., typically in the range: 10 to 15 dB), the accuracy is much better than the counterpart algorithms. While, for lower SNR problems, the performance improvement is a function of the type of noise, the worst case being the case of white noise.

2.1 Implementation Example

In this section, we illustrate practical implementation of the algorithm for the case of isolated word recognition. The same scheme can be adapted to various other speech recognition problems by a change of the parameters as discussed in this section.

The original incoming speech data is pre-processed first 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 [6]. 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 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 determine 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].

$$H = -\sum_{k=1}^{N} p_k \log p_k \ . \tag{2.1}$$

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

$$\xi = \begin{bmatrix} H_1 & H_2 & \cdots & H_m \end{bmatrix} \tag{2.2}$$

assuming m total frames in the incoming speech.

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

$$\gamma = \frac{\max(\xi) - \min(\xi)}{2} + \mu \min(\xi); \quad \mu > 0 \tag{2.3}$$

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 noise floor. A biased selection of the 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.,

$$\xi' = \begin{cases} \xi_i & \text{if } \xi_i \ge \gamma \\ 0 & \text{otherwise} \end{cases}; \quad \text{i=1,2}\ldots m \tag{2.4}$$

In many cases due to possibility of a number of artifacts, parts of the non-speech data are falsely reported as speech; also some valid speech data may be rejected due to its physio-vocal characteristics; 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 vocal or background artifacts. Humans generally do not produce very short duration sounds. Therefore each speech segment should have a certain minimum length λ_i . For any i^{th} speech segments

$$\lambda_i = e_i - s_i \tag{2.5}$$

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.

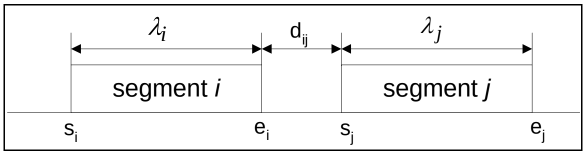


Figure 2. Relationship between adjacent speech segments

The second criterion is based on the intra-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, similarly if $\lambda_j < \kappa$ and $d_{ij} > \delta$, then the j^{th} segment is discarded

- if $(\lambda_i \text{ or } \lambda_j) > \kappa$, $d_{ij} > \delta$ and $\lambda_i + \lambda_j < \theta$, then the two segments are merged, and anything between the two segments that was previously left, 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 based algorithms.

3. SIMULATION RESULTS

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

First we discuss the case of a single word utterance to illustrate differences in the characteristics of the energy and the entropy profiles of speech data. Fig. 3 shows the speech data for utterance of digit *five*

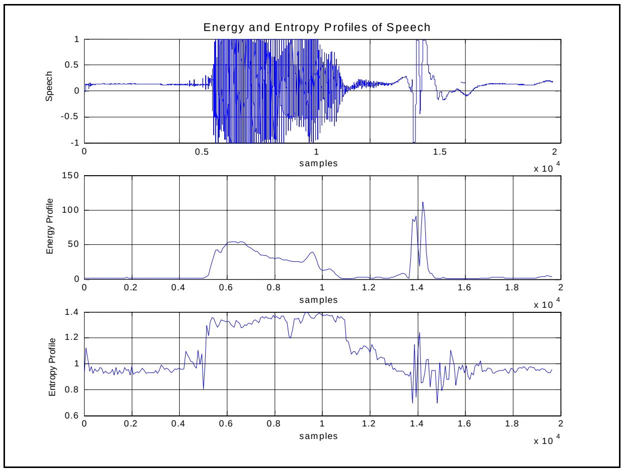
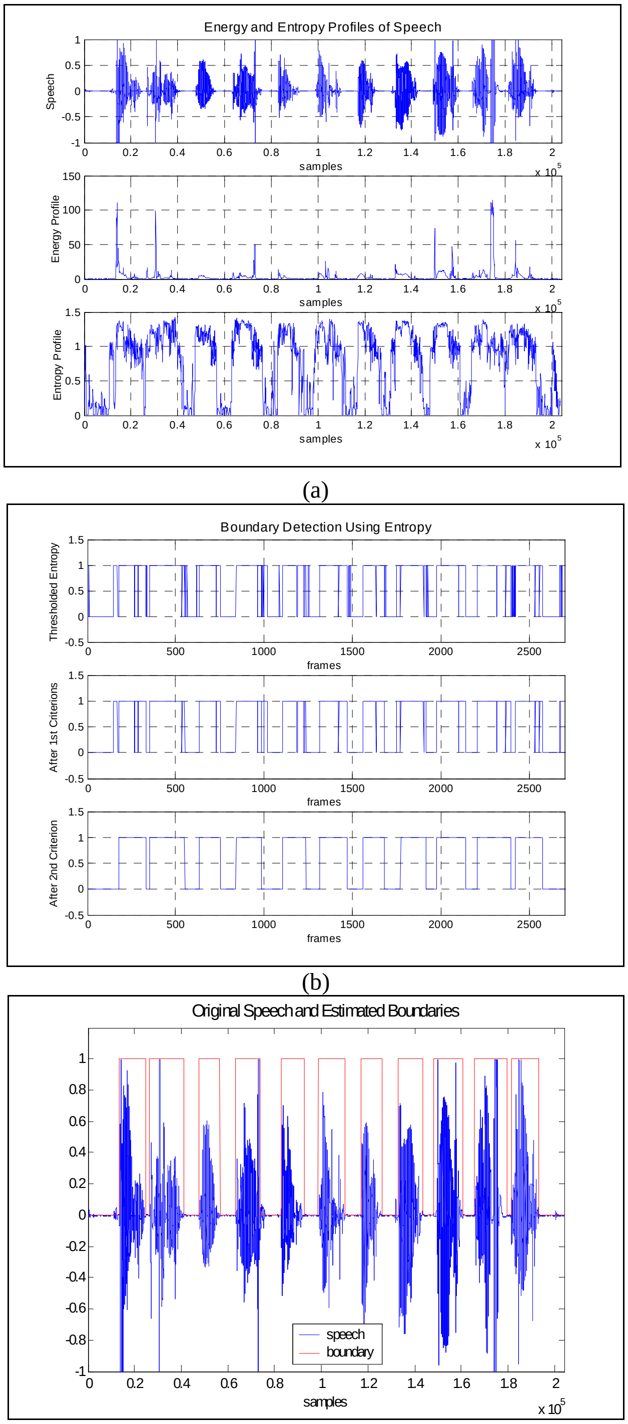


Figure 3. Digit “five”, speech data and corresponding energy and entropy profiles

It is evident that the energy profile has a higher variance as compared to the corresponding entropy profile. Further, due to the finite window effects, the energy falls off towards the end of the speech sample and then rises sharply again towards the end due to the sound “ve”. This makes the use of energy difficult for automatic endpoint detection because the lowered level of energy is long enough to suggest that there is no speech, this 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 accuracy is determined by the post-processing stage of the thresholded entropy profile.

In order to demonstrate the endpoints detected using entropy, we now choose a more interesting case of connected words. Presented below are the simulation results for a sequence of digits (1, 2…9, oh) spoken at a slower pace as may be used for speech activated car phones.

As remarked earlier, in Fig. 4.a the entropy profile seems more suited for determination of constituent digits. Fig. 4.b shows application of boundary detection criteria, which are qualitatively similar to those explained in the previous section. Fig. 4.c shows the speech data again with the estimated boundaries, which seem to be in good harmony.



(a)

(b)

(c)

Figure 4. Connected Sequence of Digits (a) Speech data, energy and entropy profiles, (b) Thresholded entropy profile and application of post criteria, (c) Identified speech boundaries

The case of continuous speech is however different. When speaking continuous sentences, people tend to slur one word into another. The human ear is still able to differentiate between the words in the sentence, but the corresponding problem for automatic speech recognition for a machine is very difficult. The faster the spoken speech is, the lesser clear it becomes to determine where one word ends and the next begin.

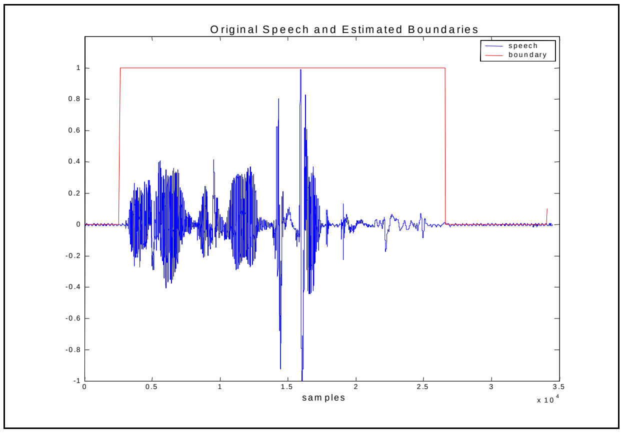


Figure 5 (a). Sentence Boundary Detection for the sentence “I want to read the book”

However, the recognition engines for continuous speech and large vocabulary systems are based on HMM models. The engines work on smaller units of speech and each word or sentence becomes a sequence of these smaller units. In practical situations this continuous speech can be intermittent, e.g., consider the case of different people using a bank ATM, or a voice operated PC software, or an interactive toy etc. In this case the proposed algorithm can be incorporated to determine the appropriate sentence boundaries and rejection of the inter-sentence periods of silence. Please, note that the criteria for post-processing the entropy profile in this case can be slightly more involved.

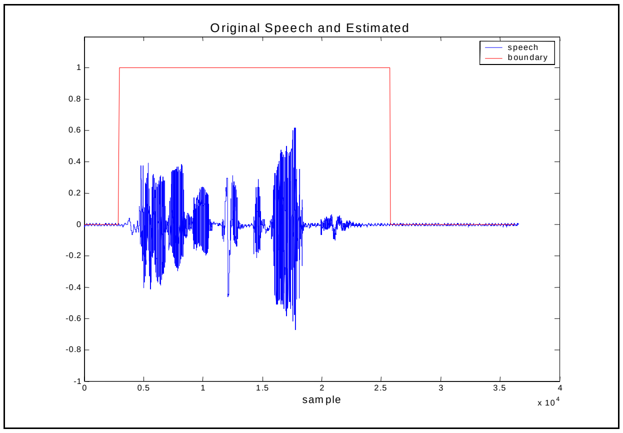


Figure 5 (b). Sentence Boundary Detection for the sentence “Where did you put the box? ”

It is certainly very good for a speech segmentation algorithm to deliver in all these different speech recognition problems. However, it is more 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. In case of isolated word recognition against white background noise, the algorithm provides an improvement of 9 % in recognition error rate at a SNR of 15 dB, which rises to 14.3% at an SNR of 10 dB and 15.6 % at an SNR of 5 dB as compared to energy based methods. For colored noise, the recognition error rate is reduced by approximately 30% at 10dB SNR and 20% at a SNR of 5dB. The details of these results will be provided in a subsequent publication.

4. CONCLUSIONS

A new 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 for 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 also shows better performance in noisy environments, which makes it even more attractive. We believe that through the use of entropy based speech recognition engines; much higher recognition rates can be achieved especially for small to medium vocabulary problems. For large vocabulary continuous speech systems, it is useful in rejecting silence periods, which may lead to power efficient recognition engines.

5. REFERENCES

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

[2] Rabiner L. and Juang B.–H.: “Fundamentals of Speech Recognition”, Prenticae Hall, NJ 1993

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

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

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

[6] Deller J. R. Jr., Hansen J. L. H. and Proakis J. G.: “Discrete Time Processing of Speech Signals”, IEEE Press, NJ, 2000.