A NEW ROBUST ALGORITHM FOR ISOLATED WORD ENDPOINT DETECTION

Lingyun Gu and Stephen A. Zahorian

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, VA, 23529, U.S.A.

ABSTRACT

Teager Energy and Energy-Entropy Features are two approaches, which have recently been used for locating the endpoints of an utterance. However, each of them has some drawbacks for speech in noisy environments. This paper proposes a novel method to combine these two approaches to locate endpoint intervals and yet make a final decision based on energy, which requires far less time than the feature based methods. After the algorithm description, an experimental evaluation is presented, comparing the automatically determined endpoints with those determined by skilled personnel. It is shown that the accuracy of this algorithm is quite satisfactory and acceptable.

1. INTRODUCTION

Endpoint detection, which aims at distinguishing speech and non-speech segments using signal processing and pattern recognition, is considered as one of the key preprocessing components in automatic speech recognition (ASR) systems. The incorrect determination of endpoints for an utterance results in at least two negative effects [1]:

1. Recognition errors are introduced;
2. Computations increase.

There have many attempts to "solve" the endpoint detection problems over the past several decades. Computing the energy of speech signal is a computationally simple operation compared to extracting other features, such as LPC derived cepstrum coefficients (LPCC), mel-frequency cepstrum coefficients (MFCC) and so on, which have been found to work well but are time and computationally intensive [4]. As for energy-based methods, most of the algorithms are based on simple parameters such as energy contours and zero crossings [3]. The Teager Energy approach considers not only energy, but also frequency effects. It can be very effective when the speech signal is very weak but has frequency components different than background noise. As for the Energy-Entropy Features approach, it emphasizes the parts of an utterance, which have a spectrum with high variability. Entropy is a measure of unexpected information. The more "unexpected" information is contained, the larger the entropy value will be. Therefore, it is a very useful tool to take the place of a widely used but unreliable parameter, Zero Crossing Rate (ZCR), to locate beginning and ending for an utterance [1, 2, 3, 5]. Both of these methods have been shown to be effective for endpoint detection. However, sometimes, they still fail, especially in a high noise environment.

In this paper, we integrate the modified Teager approach with the Energy-Entropy (EE) Features. In our new algorithm, the Teager Energy is used to determine crude endpoints, and the EE Features are used to make a final decision. The new algorithm is a simple and accurate one for the detection endpoints for isolated words spoken in a noisy environment. At the same time, it does not use the ZCR, which sometimes is not reliable. Another advantage is that there is no need to estimate the background noise. Therefore, it is very helpful for environments when the beginning or ending noise is very strong or there is not enough "silence" at the beginning or ending of the utterance. In the absence of sufficient beginning or ending silence, algorithms, which require the estimation of background noise, will have great difficulties, and resultant errors in endpoints.

2. THE ALGORITHM DESCRIPTION

2.1.Teager Energy Algorithm

In modeling speech production, Teager [1] presented a new algorithm to compute the energy of the signal. This is the so-called Teager Energy Algorithm. If a signal sample is given as $x_i = A\cos(\Omega i + \phi)$, where A is the amplitude of the signal, Ω is the discrete time frequency and ϕ is the initial phase. In the Teager Energy Algorithm, the instantaneous energy E_i of the sample x_i is given by:

$$\begin{aligned} E_i &= x_i^2 - x_{i+1}x_{i-1} \\ &= A^2 \sin^2(\Omega) \\ &\cong A^2\Omega^2 \end{aligned} \quad (1)$$

From equation 1, we can easily observe that the energy expression is based not only on the signal amplitude, but also on the corresponding frequency component.

A modification to the basic Teager Energy Algorithm is the Teager Frame Energy Algorithm. In this method, instead of calculating the instantaneous energy for each sampling point, a frame-based frequency domain approach is used, which weights each sample in a frame by the square of its associated frequency component. The Teager Frame Energy Algorithm has been found to be very useful for endpoint detection of fricatives and plosives, which have very low amplitude but high frequency [1].

2.2.Energy-Entropy Features Algorithm

Although the Teager Frame Energy works well in the presence of babble noise and background music, it has been found to fail if the non-stationary noise consists of mechanical sounds, such as closing or opening a door or engine shaking noise [2]. In these cases, the EE Features Algorithm is more reliable than pure energy-based methods and is more tolerant to the kinds of noise mentioned above. In addition, it can be really effective in overall high noise environments. In the EE Feature approach, both energy and entropy are first calculated for each frame. Then, average values of each of these components are removed, and adjusted values are multiplied, point-by-point for each frame. The multiplication emphasizes the speech region and attenuates the non-speech region, thus making this overall approach well suited to endpoint detection [2].

2.3.The New Algorithm for Endpoint Detection

In this section, an algorithm for endpoint detection that uses both the modified Teager Energy and modified EE Feature algorithm is presented below.

2.3.1.Pre-emphasizing the input signal

The input signal is first filtered with a bandpass filter from 250Hz to 3750Hz (FIR filter of order 50). This band, very similar to the band of telephone lines, is generally considered to contain the most overall speech information. Thus this type of fixed filtering is reasonably effective for improving the signal to noise ratio of speech to non-speech.

2.3.2.Calculation of the modified Teager Energy

The modified Teager Energy is computed according to the method mentioned in [1]. This algorithm is briefly summarized here. First the signal A is partitioned into overlapping frames (with frame length 20 msec and frame space 8 msec). Denote the window width as w and let s_i denote the i th sample data point. As noted above, the main modification for using the modified Teager Energy, versus the version given in Equation 1, is to more strongly consider the spectrum of the signal. In particular, for each fixed-length frame, we first compute the Fast Fourier Transform (FFT),

$$X(w)=\sum_{i=-\infty}^{\infty} s_i e^{-jwi} \tag{2}$$

Then, the magnitudes of these spectral points are weighted by the square of the corresponding frequency component to denote the i th component in the frequency domain .

$$f_i = w_i^2 X(w_i) \tag{3}$$

Finally, the modified Teager Energy T_i for one frame, is computed as the square root of the sum of the f_k in Equation 3 .

$$T_i = (\sum_{k=1}^{K} f_k)^{1/2} \tag{4}$$

In equation 4, the sum is computed over a range of 250 Hz to 3750Hz.

2.3.3.Computing the modified Energy-Entropy Feature

In this section we summarize the calculation of the modified Energy-Entropy (EE) Features, as given in [2], and also point out some of the details of our implementation. First, for each frame i, the frame energy E_i is computed as the sum of the squares of each point in the frame (using the same frame length and frame spacing as mentioned above).

$$E_i = \sum_{k=1}^{K} s_k^2 \tag{5}$$

Using the results of the DFT already computed for the modified Teager Energy, we compute the pdf (probability distributed function) for the spectrum by normalizing the frequency components:

$$p_i = \frac{X(w_i)}{\sum_{k=1}^{K} X(w_k)} \tag{6}$$

In this equation, $X(\omega_i)$ represents the magnitude of the spectral of frequency component ω_i, p_i is the corresponding probability density, K is the total point number of FFT in each frame.

Then, the entropy H_i for each frame i is defined as following:

$$H_i = \sum_{k=1}^{K} p_k \log p_k \tag{7}$$

In contrast to previously reported uses of the EE Feature algorithm, which usually subtracts the average from the first 10 frames in an attempt to reduce the effects of background noise, we bypass this step, since the modified Teager energy presented above will help to make a crude endpoint detection without the background noise estimation. Finally the EE Feature is defined as:

$$EEF_i = (1+|E_i \times H_i|)^{1/2} \tag{8}$$

2.3.4.Locating the low energy areas

The next, and final step of the endpoint detection, is to apply a set of decision rules to the features mentioned above—Teager Energy (T) and Energy-Entropy Feature (EEF). The first step is

to normalize T (with the result called T1), and EEF (with the result called EEF1). The normalization of each feature is accomplished by a scaling and offset so that each of the normalized parameters has a range of 0 to 1.

The first part of the decision logic is based on two pairs of thresholds, both of which are applied with respect to the normalized Teager Feature, T1. We denote the thresholds for the beginning part of the utterance as (Thres_B1, Thres_B2), and the thresholds for the ending part of the utterance as (Thres_E1, Thres_E2). The idea is that the interval between (Thres_B1, Thres_B2) will be the low area for the beginning of an utterance, and the interval between (Thres_E1, Thres_E2) will be the low area for the ending of an utterance. By "low" interval, we mean the interval that contains the actual endpoints. Thus these two thresholds are intended to find the approximate interval that contains the endpoints. Note the thresholds which end in "2" are greater than the thresholds which end in "1". The search method is summarized below.

Searching all the frames in an utterance from the beginning to the end, we define:

$$t_{b1} = \arg\min_{i} \{T1(i) \geq Thres\_B1\}, 1 \leq i \leq N \quad (9)$$

$$t_{b2} = \arg\min_{i} \{T1(i) \geq Thres\_B2\}, t_{b1} \leq i \leq N \quad (10)$$

In this equation, note that N is the total number of frames in the utterance.

Similarly, searching from the end of an utterance to the beginning, we define :

$$t_{e1} = \arg\max_{i} \{T1(i) \geq Thres\_E1\}, 1 \leq i \leq N \quad (11)$$

$$t_{e2} = \arg\max_{i} \{T1(i) \geq Thres\_E2\}, 1 \leq i \leq t_{e1} \quad (12)$$

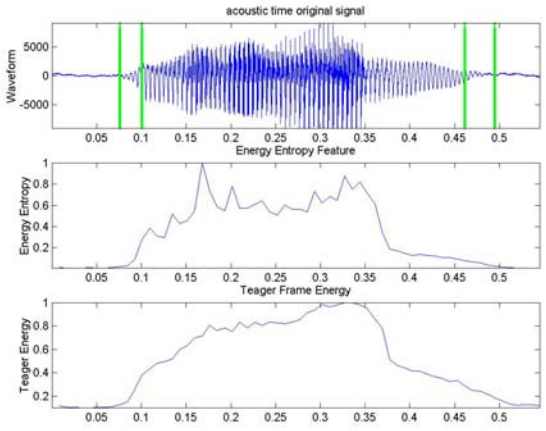

Figure1: Thres_B1, Thres_B2, Thres_E2 and Thres_E1 are represented from left to right respectively

Note that typical values for the thresholds, which apply to the normalized parameters, are 0.14, 0.16 for Thres_B1, Thres_B2 and 0.15, 0.17 for Thres_E1, Thres_E2.

Now, we have the low energy interval at the beginning and ending with (t_{b1}, t_{b2}) and (t_{e2}, t_{e1}). Figure 1 depicts typical results for determination of the two intervals.

Thus, the modified Teager energy is used to find the approximate endpoints, in terms of intervals at the beginning and end of the utterance. The EE Feature is used to find the actual endpoint, as described below.

2.3.5.Final Endpoint Detection

For the two low energy intervals, as mentioned above, we use another set of thresholds (Thres_B, Thres_E), which are applied to $EEF1$ to make the final endpoint decision. Thresh_B is used for the beginning interval, and Thresh_E is used for the ending interval. From the experimental results, we find these two thresholds value are really small, which are little bit sensitive when setting a fixed value to them.

In the beginning low energy interval, searching from t_{b1} to t_{b2}, the beginning endpoint is simply defined as:

$$t_b = \arg\min_{i} \{EEF1(i) \geq Thres\_B\}, t_{b1} \leq i \leq t_{b2}$$

In the end region low energy interval, searching from t_{e2} to t_{e1}, the ending endpoint is obtained as:

$$t_e = \arg\max_{i} \{EEF1(i) \leq Thres\_E\}, t_{e2} \leq i \leq t_{e1}$$

3. EXPERIMENTS

The described algorithm has been tested with a database, which was collected by the Speech Communication Lab in the Electrical and Computer Engineering Department at Old Dominion University. From the database, we chose two male speakers, two female speakers and two children speakers to test 10 digits, 26 alphabets, 12 CVCs and 13 vowels for a total of 732 wave files (each word was repeated twice by the speakers). A sampling rate of 22.5 kHz was used for all data. All recordings were made in a "normal" computer lab environment; for some of the recordings there is significant background and/or breath noise.

For these experiments, the frame length was 20 msec, and the frame spacing was 8 msec (or frame overlap of 12 msec. With the sampling rate used, there are 450 points in each frame, and the frame spacing is 170 points. In the results presented here, the endpoint detection is defined to be correct if the time difference between the visual and auditory test, and the algorithm results are less than 50 msec.

The following figures and table depict the performance of this algorithm. The algorithm is especially good even when the

utterance begins or ends with fricatives or plosives, and for the noise present in this data. Note that for about 10% of the recordings, the background noise (mainly breath noise), was larger than the offset or onset of the speech. The table below gives an error summary of the algorithm, both in terms of total errors and distribution of errors.

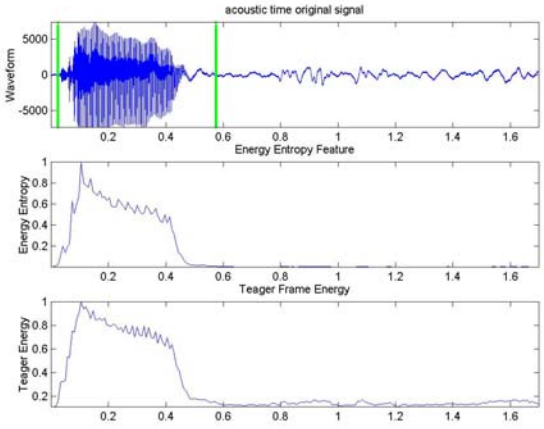

Figure2: Illustration of the endpoints selected for "UE", indicating the algorithm correctly ignores the noise burst between 0.8 and 1.6seconds

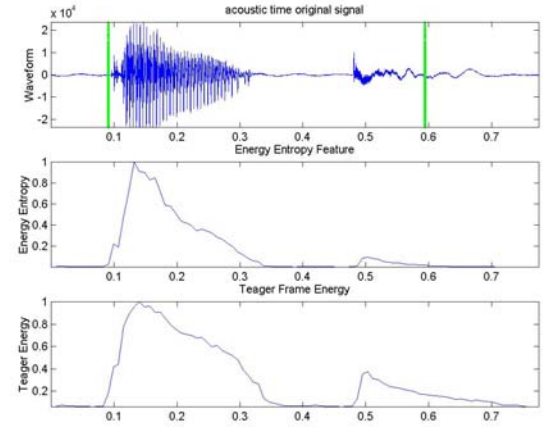

Figure3: Illustration of the algorithm for "eight", indicating that the algorithm correctly preserved the final "t".

	Male speaker	Female speaker	Children speaker
Number of files	244	244	244
Incorrect number	5	9	8
Error at beginning	3	6	6
Error at ending	2	3	2
Error missing signal	2	4	3
Error with too much silence	3	5	5

Table1: Distribution of errors in endpoint detection

	Number of files	Incorrect number	Accuracy
Male speaker	244	5	97.95%
Female speaker	244	9	96.31%
Children speaker	244	8	96.72%

Table2: Overall accuracy summary for entire database

4. CONCLUSION

The new algorithm for isolated words endpoint detection has been proposed in this paper. This algorithm combines the modified Teager Energy operator and EE Feature approach, to achieve good overall results. In particular this method is able to reliably detect the onset and offset of speech even for weak beginnings and endings, in the presence of noise, which has greater energy than the initial and final speech.

5. REFERENCES

[1] G.S.Ying, C.D. Mitchell, L.H. Jamieson, *Endpoint Detection of Isolated Utterances Based on A Modified Teager Energy Measurement*. In Proc. IEEE ICASSP-92, pp.732-pp.735, 1992

[2] Liang-Sheng Huang, Chung-Ho Yang, *A Novel Approach to Robust Speech Endpoint Detection in Car Environments*. In Proc. IEEE ICASSP-00, pp.1751-pp.1754, 2000

[3] Evangelos S. Dermatas, Nikos D. Fakotakis, George K. Kokkinakis, *Fast Endpoint Detection Algorithm for Isolated Word Recognition in Office Environment*.
In Proc. IEEE ICASSP-91, pp.733-pp.736, 1991

[4] Yiying Zhang, Xiaoyan Zhu, Yu Hao, Yupin Luo, *A Robust and Fast Endpoint Detection Algorithm for Isolated Word Recognition*, IEEE ICIPS-97, pp. 1819-1822, 1997

[5] Jean-Claude Junqua, Brian Mark, Ben Reaves, A Robust Algorithm for Word Boundary Detection in The Presence of Noise. IEEE Transactions on Speech and Audio Processing, VOL. 2. No. 3. July 1994, pp.406-412, 1994.

6. ACKNOWLEDGEMENT

This work was partially supported by NSF grant BES-9977260.