ENDPOINT DETECTION OF ISOLATED UTTERANCES BASED ON A MODIFIED TEAGER ENERGY MEASUREMENT

G.S. Ying, C.D. Mitchell, L.H. Jamieson

School of Electrical Engineering

Purdue University

West Lafayette, IN 47907-1285

ABSTRACT

Zero crossing rate and energy of the speech signal have been the two most widely used features for locating the endpoints of an utterance. We propose a new energy measure, based on Teager's energy algorithm. This new energy measure simplifies the process of endpoint detection. We present examples showing that the new measure is more effective than traditional measures in capturing some speech events, and present experimental results demonstrating that the measure can be used to improve the performance of endpoint detection algorithms.

1 INTRODUCTION

In either isolated word or continuous speech recognition systems, incorrect endpoint detection of an utterance can produce two negative effects:

1. introduce recognition errors because of the incorrect boundaries;

2. increase the computation needed if the detected boundaries incorrectly non-speech events in the utterance.

Most endpoint detection algorithms have been based on a combination of two features: zero-crossing rate (ZCR) and energy measured on the acoustic signal [1, 3, 6, 7, 8, 9]. Energy is the basic measure used to distinguish between voiced speech and either voiceless speech or background silence. ZCR provides a rough spectral measure of the frequency of major energy concentration, and aids in distinguishing between voiced speech (low ZCR), unvoiced fricatives or stop bursts (high ZCR), and silence (ZCR typically in the middle of the signal bandwidth).

In modeling speech production, Teager developed a new algorithm for computing the energy of a signal [10]; this algorithm has been presented by Kaiser [4, 5] as *Teager's Energy Algorithm.* We use this algorithm as the basis for a new energy measure that replaces the two traditional features, ZCR and energy, in an endpoint detection algorithm.

In this paper we present the new energy measure for endpoint detection. In Section II we briefly review Teager's energy algorithm. In Section III we show the adaptation of the measure for use in endpoint detection. Section IV evaluates the performance of the endpoint detector using the new measure. A summary is given in Section V.

2 TEAGER'S ENERGY ALGORITHM

2.1 RMS Energy

The most common way to calculate the energy of a speech signal is the *root mean square energy (RMSE),* which is the square root of the average of the sum of the squares of the amplitude of the signal samples. Using a window of width W to segment the speech into frames, letting $s_n(i)$ denote the i^{th} windowed speech sample in frame number n, and letting E_n be the energy of frame n, the *RMSE* of E_n is given in equation 1:

$$E_n = [\frac{1}{W}\sum_{i=1}^{W} s_n^2(i)]^{\frac{1}{2}} \tag{1}$$

2.2 Teager's Energy Algorithm

In modeling speech production, Teager developed a new algorithm for computing the energy of a signal [10]; this algorithm has been presented by Kaiser [4, 5] as *Teager's Energy Algorithm.* Given a signal with the motion of an oscillatory body, sample $x_i = Acos(\Omega i + \phi)$, where A is the amplitude of the oscillation, Ω is the digital frequency, and ϕ is the initial phase. In Teager's algorithm, the instantaneous energy E_i of the sample x_i is:

$$\begin{aligned} E_i &= x_i^2 - x_{i+1}x_{i-1} && (2)\\ &= A^2\sin^2(\Omega) \\ &\approx A^2\Omega^2 && (3)\end{aligned}$$

From Equation 3, the output of Teager's algorithm is affected not only by the amplitude of the signal samples,

but also by the oscillation frequency. This new energy measure is therefore capable of responding rapidly to changes in both A and Ω. Kaiser [5] has noted not only the ability of the measure to track rapid changes, but also the qualitatively different character of the Teager energy when measured on signals of different types. We use this property as the basis for a frame-based energy measure applied to the problem of endpoint detection.

2.3 Teager's Algorithm and the Source Energy of a Speech Signal

The fact that the Teager energy algorithm reflects both the amplitude and frequency of a signal suggests that it may be a more suitable measure for some speech events than the RMSE, which reflects only the amplitude of the signal. From the point of view of speech production, the amount of energy to produce noise-like fricatives should not be an order of magnitude less than the amount of energy to produce periodic voiced sounds, yet this is the typical difference in RMSE measured on the acoustic signal. Fricatives ($/s, \int, f, \theta, .../$) and plosives ($/p, t, k, .../$) have very low amplitude, but, unlike most vowels, these sounds have energy distributed in the frequency range above 5 KHz. A more suitable way to calculate the energy of producing the fricatives (the *source energy*) should be to consider not only the amplitude of the acoustic signal (the *acoustic energy*), but also the frequency at which the acoustic energy is located.

3 THE ENERGY MEASURE FOR ENDPOINT DETECTION

To apply Teager's algorithm to the problem of endpoint detection, we observe that the expression for the instantaneous energy in equation 2 can be related to the square of the samples of the derivative signal:

$$E_i = x_i^2 - x_{i+1}x_{i-1} \tag{4}$$

When calculating the RMSE for the samples as $x_i = A\cos(\Omega i + \phi)$, the result is proportional to A^2 only. However, if we calculate the RMSE on the derivative of x_i, then the result is proportional to A^2 and Ω^2, as is the Teager instantaneous energy. Therefore, instead of calculating the instantaneous energy for each signal sample using equation 4, we develop the following algorithm to compute the power spectrum of the *derivative* samples for each frame of speech data:

1. Calculate the power spectrum;
2. Weight each sample in the power spectrum with the square of the frequency,
3. Take the square root of the sum of the weighted power spectrum.

The result of the summation is the measure used to represent the energy of one frame; we call this the *Frame-based Teager Energy Measure*.

4 EXPERIMENTAL EVALUATION

4.1 Database

The endpoint detection algorithms were evaluated using isolated word speech data from the Texas Instruments' "Command-Digit" vocabulary [2]. The TI data was recorded in a quiet environment, sampled at a rate of 12.5 kHz, and quantized to 12 bits per sample. The utterances tested were drawn from eight speakers, five male and three female. Preliminary testing concentrated on vocabulary words starting or ending with low (acoustic) energy fricatives or plosives: *'zero', 'three', 'five', 'start', and 'rubout'*. Each of the five words was spoken ten times by each of the eight speakers, resulting in a total of 400 test files.

4.2 Computation

The speech data is segmented into 20 ms frames (250 samples per frame) with a frame overlap of 10 ms. Three energy measures are computed for each utterance: the frame-based Teager energy measure, the instantaneous Teager energy measure (equation 4), and the RMSE. The algorithm for estimating the endpoint locations from the energy measures follows that proposed by Rabiner and Sambur [8]. This algorithm employs multiple thresholds and decision logic to eliminate spurious starts and breath noise. The Rabiner and Sambur algorithm makes preliminary endpoint estimates using energy, then incorporates the ZCR to make the final endpoint estimates. In our experiments, we use only the logic pertaining to energy. For the RMSE measure, our results therefore correspond to the preliminary endpoints using the Rabiner and Sambur algorithm; we do not show the final endpoints after incorporation of the ZCR. The same endpoint detection logic was used for all of the energy measures. The silence energy level is determined by the first 100 ms of the speech samples in the utterance. The experiments were conducted using MATLAB^{TM}.

4.3 Experimental Results

To evaluate performance, we visually compare the estimated locations of the beginning and ending points using the different energy measures. ZCR is also computed, and is used aid visually in determining the "correct" endpoints. The detected endpoints are also evaluated via audio playback, by listening to the speech signals windowed at the selected endpoints.

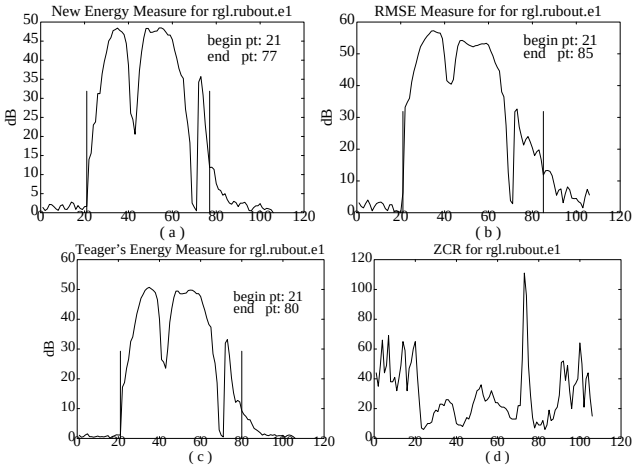

Figure 1: Results of the endpoint estimates for the word "*Rubout*" by speaker *RGL*: (a) Frame-based Teager Energy Measure, (b) RMSE Measure, (c) Instantaneous Teager Energy Measure, and (d) Zero Crossing Rate. The horizontal axis shows frame number.

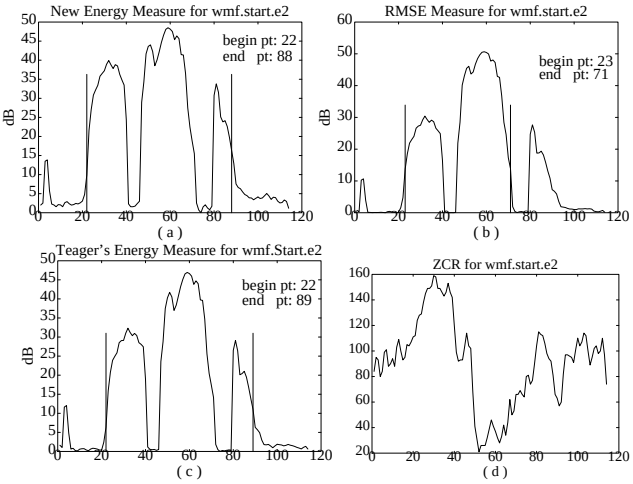

Figure 2: Results of the endpoint estimates for the word "*Start*" by speaker *WMF*: (a) Frame-based Teager Energy Measure, (b) RMSE Measure, (c) Instantaneous Teager Energy Measure, and (d) ZCR. The horizontal axis shows frame number.

The following summarizes the conclusions from the experiments on the 400 test utterances.

1. The frame-based Teager energy measure locates the end of final plosives more accurately than RMSE:
   This is typified in figure 1, word *"rubout"*. The estimate of the beginning point is the same for both energy measures. However, the estimate of the ending point using only RMSE is much later than the estimate using the new energy measure: frame 85 vs. frame 77, representing a difference of 80 ms. The ZCR plot and audio evaluation verify that the frame 77, as selected using the new energy measure, is the correct endpoint.

   In figure 2, the two energy measures identify approximately the same starting point. The new energy measure correctly locates the final endpoint. However, the RMSE measure misses the final plosive /t/ entirely, placing the ending point at frame 71, which is before the stop gap of the plosive. In testing, the RMSE measure frequently truncated the final plosive.

2. The frame-based Teager energy measure locates initial endpoints in the presence of a weak initial fricative (/f/, /θ/) more accurately than RMSE:
   The example in figure 3, word *"five"*, typifies the results for initial weak fricatives. Using only the RMSE measure, the estimate for the beginning point is at frame 30; the instantaneous and frame-based Teager energy measures place the beginning point at frame 22. This represents a difference of 80 ms. The ZCR plot indicates that the fricative most likely starts around frame 22. Audio evaluation, performed by listening to the speech signals windowed at the selected endpoints, confirms that the endpoints detected by the two Teager energy measures are correct: audio playback of the word *five* sounds like *vive* using the endpoints from the RMSE analysis. Similar results are observed for the word *three*, which becomes *dree* when played back using the endpoints based on the RMSE measure.

In addition to these differences in the overall performance of the energy measures, analysis of the 400 test utterances reveals several general properties of the frame-based Teager energy measure:

1. The new measure reports a higher energy level than RMSE for fricatives and plosives: the higher the ZCR, the larger the energy, using the Teager based measures. This is illustrated in figure 1, where the high zero crossing rate occurring at the release of the final /t/ in *"rubout"* is reflected in the new energy measure.

2. Compared to RMSE, the new measure decreases the energy difference between voiced and voiceless sound. This is illustrated in figure 2, where the difference in energy between the /s/ and /ar/ portions of the word *"start"* is approximately 20 dB using RMSE, but is less than 10 dB using the new energy measure.

3. The frame-based Teager energy measure decreases the energy difference between vowels and fricatives/plosives more than the instantaneous Teager energy measure. This is shown in the example in figure 2.

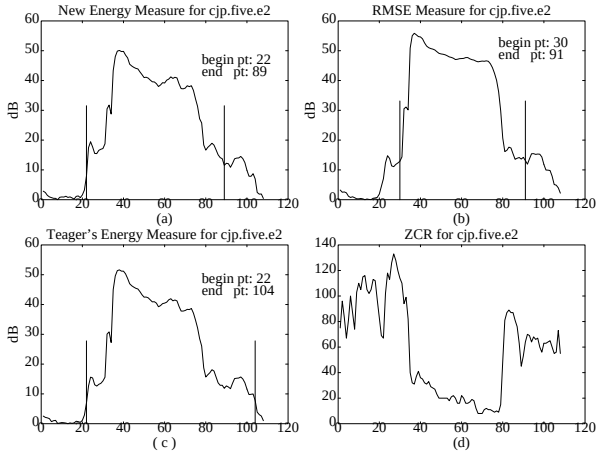

Figure 3: Results of the endpoint estimates for the word "*Five*" by speaker *CJP*: (a) Frame-based Teager Energy Measure, (b) RMSE Measure, (c) Instantaneous Teager Energy Measure, and (d) ZCR. The horizontal axis shows frame number.

4. The frame-based Teager energy measure suppresses the energy level of background silence. Both the RMSE and instantaneous Teager energy measure are more sensitive to background noise than the frame-based Teager energy measure. For the case of the instantaneous Teager energy measure, this is shown in the detection of the final endpoint in figure 3, where the instantaneous measure mistakenly includes approximately 15 frames of background noise (silence) at the end of the word.

5 CONCLUSION

We have developed an energy measure based on Teager's energy algorithm, and have applied the new measure to the problem of endpoint detection. The energy measure is important in that it (a) appears to be more suitable for describing the *source energy* associated with the production of speech sounds than the *acoustic energy* typically measured, and (b) explores a new way of viewing and using Teager's energy algorithm [4, 5, 10]. Experiments were conducted on 400 utterances on which endpoint detection was expected to be difficult. We summarized the results of the testing by presenting typical examples that show that this new measure is more effective than traditional measures in capturing speech events such as initial and final fricatives and plosives. Whereas traditional endpoint detectors have used both (acoustic) energy and zero crossing rate, the new measure effectively combines this information into a single measure. The experimental results demonstrate that the measure can be used to improve the performance of endpoint detection algorithms and should be effective for the detection of speech in noisy environments.

Further experiments will include performing endpoint detection using the frame-based Teager energy measure on speech to which different types of the noise have been added, in order to test the robustness of this measure. We will also evaluate the endpoint detection algorithm in the context of an HMM-based word recognizer, in order to evaluate the usefulness of the new energy measure as a component of a complete system.

References

[1] B.S. Atal and L.R. Rabiner. A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition. *IEEE Trans. ASSP*, ASSP-24:201–212, June 1976.

[2] G.R. Doddington and T.B. Schalk. Speech recognition: Turning theory to practice. *IEEE Spectrum*, pp. 26–32, Sept. 1981.

[3] M.Hahn and C.K. Park. An improved speech detection algorithm for isolated korean utterance. In *Proc. IEEE ICASSP-92*, pp. I525–528, Mar. 1992.

[4] J. F. Kaiser. On a simple algorithm to calculate the 'energy' of a signal. In *Proc. IEEE ICASSP-90*, pp. 381–384, Apr. 1990.

[5] J. F. Kaiser. On Teager's energy algorithm and its generalization to continuous signals. In *Proc. 4th IEEE Digital Signal Processing Workshop*, Mohonk, NY, Sept. 1990.

[6] L.F. Lamel, L.R. Rabiner, A.E. Rosenberg, and J.G. Wilpon. An improved endpoint detector for isolated word recognition. *IEEE Trans. ASSP*, ASSP-29:777–785, Aug. 1981.

[7] D. O'Shaughnessy, *Speech Communication*, Addison-Wesley, Reading, MA, 1987.

[8] L.R. Rabiner and M.R. Sambur. An algorithm for determining the endpoints of isolated utterances. *Bell System Tech. Journal*, 54:297–315, Feb. 1975.

[9] L.J. Siegel and A.C. Bessey. Voiced/unvoiced/mixed excitation classification of speech. *IEEE Trans. ASSP*, ASSP-30:451–460, June 1982.

[10] H. M. Teager. Some observations on oral air flow during phonation. *IEEE Trans. ASSP*, ASSP-28:599–601, Oct. 1980.