# A ROBUST, REAL-TIME ENDPOINT DETECTOR WITH ENERGY NORMALIZATION FOR ASR IN ADVERSE ENVIRONMENTS

*Qi Li, Jinsong Zheng, Qiru Zhou, and Chin-Hui Lee*

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974, USA
{qli,jszheng,qzhou,chl}@research.bell-labs.com

## ABSTRACT

When automatic speech recognition (ASR) is applied to hands-free or other adverse acoustic environments, endpoint detection and energy normalization can be crucial to the entire system. In low signal-to-noise (SNR) situations, conventional approaches of endpointing and energy normalization often fail and ASR performances usually degrade dramatically. The goal of this paper is to find a fast, accurate, and robust endpointing algorithm for real-time ASR. We propose a novel approach of using a special filter plus a 3-state decision logic for endpoint detection. The filter has been designed under several criteria to ensure the accuracy and robustness of detection. The detected endpoints are then applied to energy normalization simultaneously. Evaluation results show that the proposed algorithm significantly reduce the string error rates on 7 out of 12 tested databases. The reduction rates even exceeded 50% on two of them. The algorithm only uses one-dimensional energy with 24-frame lookahead; therefore, it has a low complexity and is suitable for real-time ASR.

## 1. INTRODUCTION

Speech processing is based on the premise that the signal in an utterance consists of speech, silence or other background noise. The detection of the presence of speech embedded in various types of non-speech events and background noise is called endpoint detection or speech detection. Real-time endpoint detection is a continuous time process requiring a short time delay.

As is well known, endpoint detection is crucial to automatic speech recognition (ASR) because it can affect the performance of an ASR system in terms of accuracy and speed for several reasons. First, cepstral mean subtraction (CMS), as a popular algorithm for robust speaker and speech recognition, needs accurate endpoints to compute the mean of voice frames precisely in order to improve recognition accuracy. Second, if silence frames can be removed prior to recognition, the accumulated utterance likelihood scores will focus more on the speech portion of an utterance instead of scoring both noise and speech. Therefore it has the potential to increase the recognition accuracy. Third, it is hard to model noise and silence accurately. This effect can be limited by removing background noise frames in advance. Last, one can significantly reduce the computation time by removing non-speech frames.

A problem related to endpoint detection is real-time energy feature normalization. In ASR, we usually normalize the energy level such that the largest energy level in a given utterance is close to or slightly below zero. This is not a problem in batch-mode processing, but it can be a crucial problem in real-time mode since it is difficult to estimate the maximal energy within a limited data buffer while the acoustic environment is changing. It is especially hard in adverse acoustic environments.

Endpoint detection has been studied for several decades and many papers have been published about various applications (e.g. [1, 2, 3, 4]). A lookahead approach on energy normalization can be found in [5]. In recent years, ASR is applied to hands-free, wireless, IP phone, and other adverse acoustic environments. The source speech is often with very low signal-to-noise ratio (SNR). In these cases, the ASR performance often degrades dramatically due to unreliable endpoint detection and energy normalization. This paper is to propose a combined approach for both endpoint detection and energy normalization to benefit real-time ASR in adverse environments.

## 2. THE PROPOSED ALGORITHM

To ensure the low complexity requirements, we choose the one-dimensional (1-D) short-term energy in dB from cepstral feature as the feature for endpoint detection. We first design a filter to detect all possible endpoints on the energy feature, then develop a 3-state decision logic for final, reliable decisions.

### 2.1. Optimal Filter for Endpoint Detection

We need to design a filter with the following criteria: (i) invariant outputs to various background energy levels; (ii) capability to detect both beginning and ending points; (iii) limited length or short lookahead; (iv) maximum output SNR at endpoints; (v) accurate location of detected endpoints; and finally, (vi) maximum suppression of false detection.

Fortunately, the last three criteria are very similar to the criteria in optimal edge detection in image processing. The foundation of the theory of the optimal edge detector was first set by Canny [6]. He derived an optimal step-edge detector. Petrou and Kittler then extended the work to ramp-edge detection [7]. Since the edges corresponding to endpoints in the energy feature are closer to the ramp edge than the ideal step edge, Li and Tsai applied Petrou and Kittler's filter to batched-mode endpoint detection for speaker verification [1]. We now need to extend the batch-mode algorithm to real-time mode and add real-time energy normalization in the task.

Assume that the beginning edge in the energy is a ramp edge that can be modeled by the function

$$c(x) = \begin{cases} 1 - e^{-sx}/2 & \text{for } x \geq 0 \\ e^{sx}/2 & \text{for } x \leq 0 \end{cases} \tag{1}$$

where $s$ is some positive constant. The problem is to find a filter profile $f(x)$ which maximizes a mathematic representation of the Criteria (iv), (v), and (vi) [7][1]. By optimizing the criteria with the boundary conditions as discussed in [7] and in above Criterion (i), a solution for the filter profile is:

$$\begin{aligned} f(x) \quad = \quad & e^{Ax}\left[K_1 \sin(Ax) + K_2 \cos(Ax)\right] \\ & + e^{-Ax}\left[K_3 \sin(Ax) + K_4 \cos(Ax)\right] \\ & + K_5 + K_6 e^{sx}, \end{aligned} \tag{2}$$

where $A$ and $K_i$ are filter parameters. Since $f(x)$ is only half of the filter from $-w$ to 0, the actual function of the filter for the edge detection is

$$h(i) = \{-f(-w \leq i \leq 0),\ f(-1 \leq i \leq -w)\}. \tag{3}$$

For a filter with $s = 1$ and $w = 7$, its parameters are $A = 0.41$, and $(K_1, ...K_6) = (1.583, 1.468, -0.078, -0.036, -0.872, -0.56)$ [7]. In our case, we need to chose a single filter to obtain reliable responses to both beginning and ending points. After investigating the beginning and ending edges of a few utterances, we chose $w = 13$. We then rescaled the original filter by $s = 7/13$ and $A = 0.41s = 0.2208$. All other parameters are the same. The profile of the designed filter is shown in Fig. 1, with a simple normalization, $h/13$. The profile indicates that the filter response will be positive to a beginning edge, negative to an ending edge, and close to zero to silence. The response is basically invariant to background noise at different energy levels. For real-time detection, let $H(i) = h(i - 13)$, and the filter actually has a 24-frame lookahead. So far, we have met all 6 criteria.

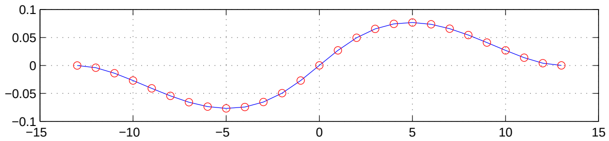

Figure 1: The profile of designed optimal filter.

### 2.2. Decision Logic

The filter operates as a moving-average filter:

$$F(t) = \sum_{i=1}^{W=25} H(i)E(t + i - 1), \tag{4}$$

where $E(.)$ is the energy feature, and $t$ is the current frame number. The output $F(t)$ is then evaluated in a 3-state transition diagram for final endpoint decisions.

As shown in Fig. 2, the diagram has three states: *silence*, *in-speech* and *leaving-speech* states, respectively. Either the Silence or the In-Speech state can be a starting state, and any state can be a final state. In this paper, we assume the Silence state is the starting state. The input is $F(t) \in \mathbf{R}$, and the output is the detected frame numbers of beginning and ending points. The transition conditions are labeled on the edge between states, and the actions are listed in parentheses. "Count" is a frame counter, $T_L$ and $T_U$ are two thresholds, and "Gap" is an integer indicating the required number of frames from a detected endpoint to the actual end of speech.

We use Fig. 3 as an example to illustrate the state transition. The raw energy is in Fig. 3 (A) (bottom line) and the filter output is in Fig. 3 (B) (solid line). The diagram stays in the Silence state until $F(t)$ reaches point A in Fig. 3 (B), where $F(t) \geq T_U$ means that a beginning point is detected. The actions are to output a beginning point (corresponding to the left vertical solid line in Fig. 3 (C)) and to move to the In-Speech state. It stays in the In-Speech sate until reaching point B in Fig. 3 (B), where $F(t) < T_L$. The diagram then moves to the Leaving-Speech state and sets Count = 0. After it stays in the Leaving-speech state for Gap = 30 frames, an actual endpoint is detected and the diagram moves back to the Silence state at point C (corresponding to the left vertical dashed line in Fig. 3 (C)).

### 2.3. Energy Normalization

Suppose the maximal energy value in an utterance is $E_{\max}$. Energy normalization is to normalize the utterance energy $E(t)$, such that the largest value of energy is close to zero by performing $E'(t) = E(t) - E_{\max}$. In real-time mode,

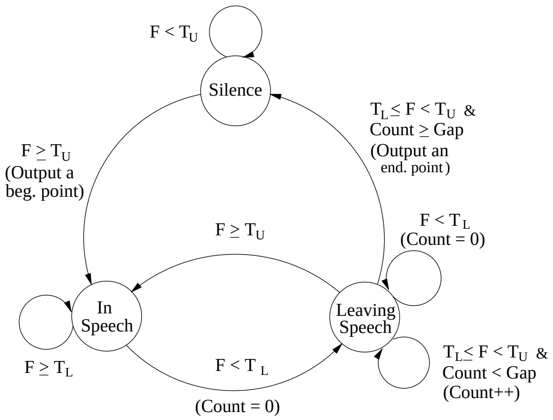

Figure 2: The state transition diagram for endpoint decision.

we have to estimate the maximal energy $E_{\mathsf{max}}$ sequentially while the data collection is going on. Here, the estimated maximum energy becomes a variable, i.e. $\hat{E}_{\mathsf{max}}(t)$. Nevertheless, we can use the detected endpoints to perform a better estimation.

We first initialize the maximal energy to be $E_0$, and use it for normalization until we detect the first beginning point A, i.e. $\hat{E}_{\mathsf{max}}(t) = E_0, \forall t < \mathsf{A}$. If the average energy $\bar{E}(t) = \mathsf{E}\{E(t); \mathsf{A} \leq t < \mathsf{A} + W\} \geq E_m$, where $E_m$ is a pre-selected threshold, we then estimate the maximal energy as $\hat{E}_{\mathsf{max}}(t) = \max\{E(t); \mathsf{A} \leq t < \mathsf{A} + W\}$, where $W = 25$ is the length of the filter. From now on, we update $\hat{E}_{\mathsf{max}}(t)$ as, $\hat{E}_{\mathsf{max}}(t) = \max\{E(t + W - 1), \hat{E}_{\mathsf{max}}(t - 1); \forall t > \mathsf{A}\}$, recursively.

For the example in Fig. 3, the energy features of two utterances with 20 dB SNR (bottom) and 5 dB SNR (top) are plotted in Fig. 3 (A). The 5 dB one is generated by artificially adding car noise to the 20 dB one. The filter outputs are shown in Fig. 3 (B) for 20 dB (solid line) and 5 dB (dashed line) SNRs, respectively. The detected endpoints and normalized energy for 20 and 5 dB SNRs are plotted in Fig. 3 (C) and Fig. 3 (D), respectively. We note that the filter outputs to 20 and 15 dB cases are almost invariant around $T_L$ and $T_U$, although their background energy levels have a 15 dB difference. This ensures the robustness in detection. We also note that the normalized energy profiles are almost the same as the original one, although the normalization is done in a real-time mode.

## 3. LARGE DATABASE EVALUATION

The proposed algorithm was further evaluated on 12 databases collected from the telephone networks with 8 KHz sampling rates in various acoustic environments. LPC feature and short-term energy were used. The HMMs are consisted of 1 silence model, 41 mono-phone models, and 275 head-body-tail units for digit recognition. It has a total of 79 phoneme

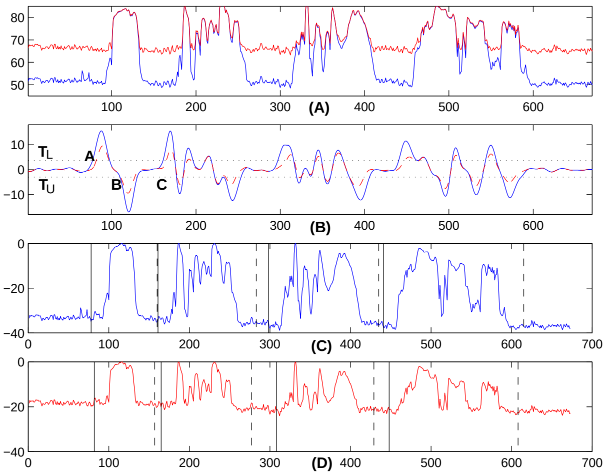

Figure 3: (A) Energy features of “4-327-631-Z214" from original utterance (bottom, 20 dB SNR) and after adding car noise (top, 5 dB SNR). (B) Filter outputs to 5 dB (dashed line) and 20 dB (solid line) SNR cases. (C) Detected endpoints and normalized energy for the 20 dB SNR case, and (D) for the 5 dB SNR case.

symbols, 33 of which are for digit units. Twelve databases, DB1 – DB12, were used for the evaluation, where DB6 – DB11 contain digit strings, and DB12 contains alphabet strings. No digit or alphabet string length constraint has been applied to the experiments. All the other databases contain digits, alphabet, and word strings. Finite-state grammars were used to specify the valid forms of recognized strings. In the evaluation, both endpoint detection and energy normalization were performed in real-time mode and only the detected voice portions of an utterance were sent to the recognition back-end.

In the baseline system, endpoints are detected by a 6-state diagram with multiple absolute thresholds on absolute energy values. The energy normalization in the baseline system is done separately by estimating the maximal and minimal energy values, then comparing their difference to a fixed threshold for decision. Since the energy values change with acoustic environments, the baseline approach causes an unreliable endpoint detection and energy normalization in low SNR cases.

In the proposed system, we set the parameters as $E_0 = 80.0$, $E_m = 60.0$, $T_U = 3.6$, $T_L = -3.0$, and Gap = 30. These parameters were selected by investigating the energy and filter output of a few utterances. The parameters were unchanged throughout the evaluation on all 12 databases to show the robustness of the algorithm, although the parameters can be adjusted according to signal conditions in different applications. The evaluation results are listed in Tab. 1. It shows that the proposed algorithm provided significant string error reduction in 7 out of 12 databases. The

Table 1: DATABASE EVALUATION RESULTS (%)

| Database IDs | Word Error Rate | | String |
|---|---|---|---|
| (Number of strings, | Base- | Pro- | Error |
| Number of words) | line | posed | Reduction |
| DB1 (232, 1393) | 13.7 | 11.3 | 10.8 |
| DB2 (671, 1341) | 14.6 | 6.9 | 50.9 |
| DB3 (1957,1957) | 4.5 | 4.8 | -6.7 |
| DB4 (272, 1379) | 10.0 | 10.1 | 8.5 |
| DB5 (259, 2632) | 15.8 | 16.0 | -2.1 |
| DB6 (576, 1738) | 2.8 | 1.1 | 51.5 |
| DB7 (583, 1743) | 1.7 | 1.5 | 12.2 |
| DB8 (664, 2087) | 0.9 | 0.6 | 27.6 |
| DB9 (619, 8194) | 1.0 | 1.5 | 11.7 |
| DB10 (651, 8452) | 5.7 | 6.8 | -6.2 |
| DB11 (707, 9426) | 1.6 | 1.9 | -4.3 |
| DB12 (661, 3681) | 40.7 | 38.5 | 0.0 |

string error reductions even exceeded 50% on two of them.

To analyze the improvement, the original energy feature of an utterance, “1 Z 4 O 5 8 2", in DB6 is plotted in Fig. 4 (A). The detected endpoints and normalized energy using conventional approach are shown in Fig. 4 (B) while the results of the proposed algorithm are shown in Fig. 4 (C). The filter output is plotted in Fig. 4 (D). From Fig. 4 (B), we can observe that the normalized maximal energy of the conventional approach was about 10 dB below zero, which gave a wrong recognition result: “1 Z 4 O 5 8". On the other hand, the proposed algorithm normalized the maximal energy close to zero, and the utterance was recognized correctly as “1 Z 4 O 5 8 2".

The above evaluation is based on telephone data which have over 15 dB SNR. In a separate evaluation using a dataset with 10 and 5 dB SNRs, the baseline system failed to work due to its real-time energy normalization algorithm while the proposed real-time algorithm still gave the same recognition accuracy as using the batch-mode energy normalization.

## 4. CONCLUSIONS

We have proposed a real-time algorithm for combined endpoint detection and energy normalization with a 24-frame lookahead. All possible endpoints are first detected by an optimal filter designed to provide accurate and robust response to endpoints. The output from the filter is then evaluated by a 3-state transition diagram for final endpoint decisions. The energy normalization utilizes the endpoint detection results. Since the entire algorithm only uses 1-D energy feature, it has very low complexity and is fast in computation. Furthermore, since the decision is made on the filter output, which is almost invariant to background noise levels, the endpoint detection is reliable and robust, even in very low SNR situations. The evaluation on 12 databases showed that 7 of them have significant string error rate reduction and 2 of them exceed 50% string error reduction.

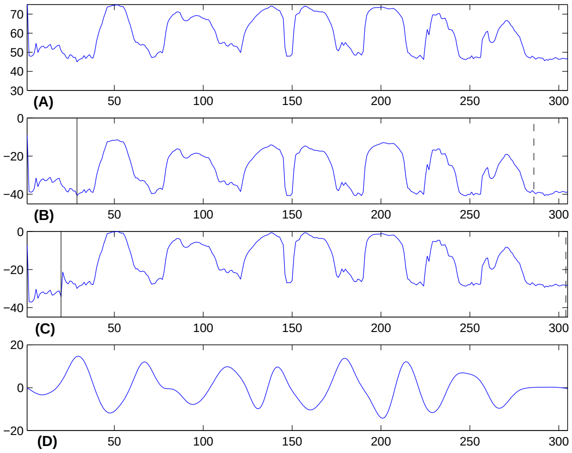

Figure 4: (A) Energy feature of the 523th utterance in DB6: “1 Z 4 O 5 8 2". (B) Endpoints and normalized energy from the baseline system. It was recognized as “1 Z 4 O 5 8". (C) Endpoints and normalized energy from the proposed system. It was recognized correctly as “1 Z 4 O 5 8 2". (D) The filter output.

## 5. ACKNOWLEDGMENT

The authors wish to thank Dr. D. W. Tufts for a very useful lecture and Drs. O. Siohan, F. K. Soong, and F. Korkmazskiy for useful discussions.

## 6. REFERENCES

[1] Q. Li and A. Tsai, “Fast, efficient endpoint detection for robust speaker verification,” in *Proceedings of IEEE Workshop on Automatic Identification*, Oct. 1999.

[2] S. G. Tanyer and H. Özer, “Voice activity detection in nonstationary noise,” *IEEE Trans. on Speech and Audio Processing*, vol. 8, pp. 478–482, July 2000.

[3] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, “An improved endpoint detector for isolated word recognition,” *IEEE Trans. on Acoustics, speech, and signal processing*, vol. ASSP-29, pp. 777–785, August 1981.

[4] J. G. Wilpon, L. R. Rabiner, and T. Martin, “An improved word-detection algorithm for telephone-quality speech incorporating both syntactic and semantic constraints,” *AT&T Bell Laboratories Technical Journal*, vol. 63, pp. 479–498, March 1984.

[5] R. Chengalvarayan, “Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition,” in *Proceedings of Eurospeech’99*, (Budapest), pp. 61–64, Sept. 1999.

[6] J. Canny, “A computational approach to edge detection,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, pp. 679–698, Nov. 1986.

[7] M. Petrou and J. Kittler, “Optimal edge detectors for ramp edges,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 483–491, May 1991.