# A MATCHED FILTER APPROACH TO ENDPOINT DETECTION FOR ROBUST SPEAKER VERIFICATION

*Qi Li and Augustine Tsai*

Multimedia Communications Research Laboratory
Bell Labs, Lucent Technologies
Murray Hill, NJ 07974
{qli,atsai}@bell-labs.com

## ABSTRACT

Fast and accurate endpoint detection of spoken utterance in noise environment is important to robust speech and speaker recognition. The detection algorithm can affect system performance, including accuracy and speed, significantly. In this paper, we propose a fast algorithm to locate endpoints by combining the information from matched filter responses and statistical measures of the spectral energy. Experiments showed that the algorithm can locate endpoints accurately even for the utterances containing voice phrases surrounded by heavy breath, clicks, dial tone, or other noise. The algorithm has been tested in a database for speaker verification and has been found to perform well in a real language-independent voice controller with embedded speaker verification capability.

## 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 from the various types of non-speech events and background noise is called endpoint problem or speech detection. Fast, accurate endpoint detection is important for two reasons. First, the accuracy of speech or speaker recognition depends on the accuracy of endpoint detection. For example, cepstral mean subtraction (CMS), as a popular algorithm for robust speaker and speech recognition, needs accurate endpoints, therefore, the mean of voice data can be computed precisely. Second, the computation of speech recognition can be significantly reduced if endpoints can be accurately located such that the non-speech signal can be removed before speech modeling and decoding. As pointed out in previous studies (e.g. [1]), endpoint detection is a difficult problem. The non-speech events and background noise complicate the endpoint detection problem considerably. For example, the beginning or end of speech is often obscured by speaker generated artifacts such as clicks, pops, heavy breathing, or dial tone signal. Similar types of artifacts and background noise are also introduced by long-distance telephone transmission system.

In this paper, we intend to apply the theory developed for edge detection to the endpoint problem. The filter outputs and statistical information from the spectral energy are then combined to locate accurate endpoints. A fast, robust, and language-independent speaker verification system is then proposed based on the proposed endpoint algorithm.

## 2. ENDPOINT DETECTION ALGORITHM

We consider that one utterance may have several voice segments. Each of the segments is determined by a pair of endpoints, called *beginning* and *ending points*. When we plot the cepstral energy of an utterance, the edges corresponding to the beginning and ending points are named as *beginning* and *ending edges*, respectively. We will first introduce the matched filters for edge detection, then a simple statistical model for spectral energy following by the proposed algorithm with several examples.

### 2.1. Optimal Filter for Edge Detection

The foundation of the theory of optimal edge detector was first set by Canny [2] for image processing. Canny's optimal step edge detector was developed based on three criteria: good signal to noise ratio, good locality, and maximum suppression of false responses. Petrou and Kittler then extended Canny's work to ramp edge detection [3]. Since the edges in spectral energy are closer to the ramp edge than the ideal step edge, we applied Petrou and Kittler's filter to endpoint detection.

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

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

where $s$ is some positive constant. Also, we assume that the edges are emersed with white Gaussian noise. Following

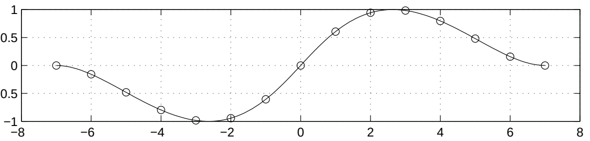

Figure 1: The function of the matched filter for beginning edge detection, plotted as $h_b(-x)$, with $W = 7$ and $s = 1$.

Canny's criteria, Petrou and Kittler proposed a 1-D convolution filter for ramp edge detection [3]. The function of the filter within the range $-W \le x \le 0$ is

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

where $A$ and $K_i$ are filter parameters. The entire function of the filter for beginning edge detection is

$$h_b(x) = \{-f(-W \le i \le 0),\ f(-1 \le i \le -W)\}, \tag{3}$$

where $-W \le x \le W$. Given the profile of beginning edge, we choose $s = 1$ and $W = 7$. Other filter parameters provided in [3] are $A = 0.41$, and $K_1 ... K_6 = \{1.583, 1.468, -0.078, -0.036, -0.872, -0.56\}$. In order to show that the function is consistent to the beginning edge, we plot it as $h_b(-x)$ in Fig. 1. On the other hand, the filter for ending point detection is defined as $h_e(x) = -h_b(x)$ to ensure positive responses at edge locations. Since the last ending edge in an utterance is usually wider than others, we have $W = 35$, $s = 0.2$ and $A = 0.082$. Other parameters are the same as above.

### 2.2. Spectral Energy Model

We assume the distribution of cepstral energy in an utterance to be approximately represented by a Gaussian mixture model with two mixtures representing voice and background energy level respectively,

$$p(x) = c\mathcal{N}_1(x; \mu_1, \sigma_1) + (1 - c)\mathcal{N}_2(x; \mu_2, \sigma_2), \tag{4}$$

where $\mathcal{N}_i$ is a normal distribution, $\mu_i$ and $\sigma_i$ are the mean and stand deviation respectively, and $c$ is a weighting parameter. The means for voice and background are $\mu_v = \max\{\mu_1, \mu_2\}$ and $\mu_n = \min\{\mu_1, \mu_2\}$ with the corresponding standard deviations, $\sigma_v$ and $\sigma_n$. The thresholds for voice and background are $\theta_v = \mu_v - \sigma_v$ and $\theta_n = \mu_n + \sigma_n$, respectively. When the value of cepstral energy is above $\theta_v$, we consider it as voice. When the value of cepstral energy is below $\theta_n$, we consider it as background noise. To obtain fast and explicit parameter estimation, we applied a moment algorithm instead of the popular EM algorithm. Detail of the estimation algorithm can be found in [4].

### 2.3. Proposed Algorithm

We use the example in Fig. 2 to present the concept of the proposed algorithm. The utterance, "Call office", is first converted to log spectral energy, $g(x)$. The energy level is normalized to have the largest value be zero. We first estimate $\mu_v$, $\theta_v$, $\theta_n$, and $\mu_n$. The results are shown in Fig. 2 as the horizontal dashed lines from top to bottom respectively. Then, we compute the convolution, $y_b(x) = h_b * g(x)$, for beginning point detection. The filter output $y_b$ is shown in Fig. 3 as a solid line. Each peak in the output can be a candidate of the beginning edge of a voice segment. After comparing the values of the peaks with a predetermined threshold, the two largest peaks were determined as the locations of the centers of beginning edges. From the first beginning point, we search for the location where the energy level is lower than $\theta_n$ as the corresponding ending point. For this example, we got two pairs of endpoints corresponding to two voice segments, as shown in Fig. 2, from line E to F and from line G to H, respectively. As we can see from Fig. 2, the last segment in between line G and H including the heavy breath. The energy signal in that segment is then fed into the ending-edge filter, $h_e$. The filter output is shown in Fig. 3 as the dashed line. The ending point for the last segment is located by shifting the frame index of the largest peak to the right for about a half of the size of the ending edge filter. The final voice segments are from line E to line F and from line G to line I, respectively, as shown in Fig. 4.

We now summarize the proposed algorithm for endpoint detection as follows.

1. Compute log energy of the given utterance, $g(x)$, and normalize it to have the highest value be 0. We assume that the speech is surrounded by silence and various kinds of noise.

2. Remove dial tone signal from $g(x)$. The dial tone can be detected at $g(x) > -1.5$, $x = n...i$, when $i - n > 8$. These two parameters are determined

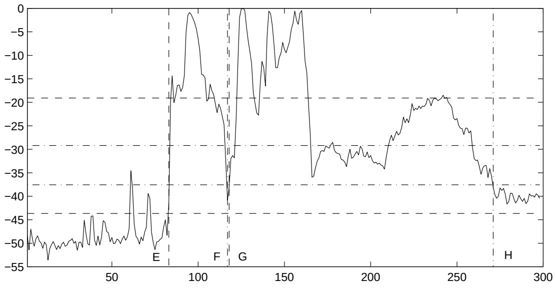

Figure 2: Normalized log energy of "Call office" with heavy breath in the end.

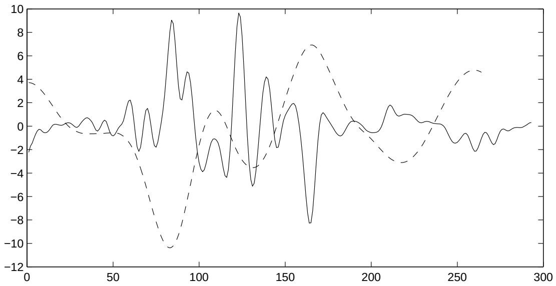

Figure 3: The outputs of the beginning-edge filter (solid line) and ending-edge filter (dashed line).

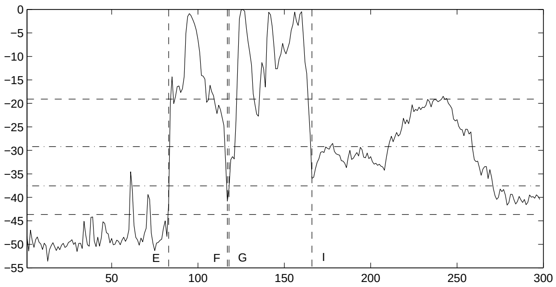

Figure 4: The last ending point was adjusted from Line H to I by applying the ending-edge filter.

based on the minimal length and minimal energy level of dial tones.

3. Estimate $\mu_v$, $\mu_n$, $\sigma_v$, and $\sigma_n$ from $g(x)$ using the moment algorithm [4], then determine two thresholds $\theta_v = \mu_v - \sigma_v$ and $\theta_n = \mu_n + \sigma_n$ for voice and background energy, respectively. Voice energy should be above the value of threshold $\theta_v$ and silence/background noise energy should be lower than $\theta_n$.

4. If there is no silence frames in the beginning of $g(x)$, add $W$ frames of silence with the value of $\mu_n$, where $W = 7$ for the beginning edge. Compute the convolution $y_b(x) = h_b * g(x)$, where $h_b$ is the matched filter for beginning edge detection. Search for the locations of all peaks $R(k)$, from the filter output. However, not all the peaks are associated with beginning points. A peak associated with a beginning point should have the following properties: $y'(R(k)) = 0$, $y''(R(k)) < 0$, and $y(R(k)) > 0.2\max\{y_b\}$. The actual beginning point is $B(m) = R(m) - 2,\ m = 1, ..., M$, where $M$ is the total number of beginning edges in the utterance. The shift is due to the offset between the center of the beginning edge and the actual beginning point.

5. From the first beginning point $B(m),\ m = 1$, search for corresponding ending point $E(m)$, which should satisfy the following conditions: (1) $g(E(m)) \geq \theta_n$and $g(E(m)+1) < \theta_n$; (2) $E(m) - B(m) \geq 6$; (3) 60% frames of $g(x)$, $B(m) \leq x \leq E(m)$, should have the values above $\theta_v$; and (4) $E(m) < B(m+1)$. Here, (2) and (3) is to ensure that the segmentation is voice but a click or breath noise. The parameters are independent to utterance contents. This gives totally $M$ pairs of endpoints and $M$ voice segments. The segment that can not meet the above conditions is not considered as a voice segment.

6. Search for the last ending point. Compute the response of the ending-edge filter in the last segment, $y_e(x) = h_e * g(x)$, $B(M) \leq x \leq E(M)$. Search for the last peak of $y_e$ at $x = T$, and $y_e(T) \geq 0.6\max\{y_e(x)\}$. Then, shift the peak point located in the center of ending edge to the last ending point. The offset should be about the half of the filter size. We choose16 frames. Then, $E(M) = T + 16$, if $g_e(T+16) \geq \theta_n$, otherwise, $E(M) = \ell$ where $g(\ell) \geq \theta_n$ and $g(\ell + 1) < \theta_n$.

## 3. SPEAKER VERIFICATION APPLICATIONS

To evaluate the effectiveness, we first compared the endpoints detected by the proposed algorithm with the endpoints detected by HMM approach, using manually detected endpoint as references. The experiments, on a database with 100 speakers and 4741 utterances, showed that the proposed approach has the similar accuracy as the HMM approach on locating endpoints while the proposed one is much faster.

Then, we apply the proposed algorithm to develop fast, robust, and language-independent speaker verification systems. The system front-end is shown in Fig. 5 [5]. After LPC cepstral extraction, the proposed algorithm is applied to detect endpoints on cepstral energy, then silence, breath, dial tone, and other non-speech signals are removed from the feature set. Given the original feature observation of $\mathcal{O}$, after silence removal, the feature set becomes $\mathbf{O}$ which is a subset of $\mathcal{O}$, i.e., $\mathbf{O} \subset \mathcal{O}$. CMS is then performed on $\mathbf{O}$.

The speaker verification performance was evaluated on a database consisting of 38 speakers, 18 male and 20 female [6] for speaker verification. The common pass-phrase for all speakers is "Call Janice at her office phone." Each true speaker is tested with the same pass-phrase from all impostors. The feature vector is composed of 12 cepstrum and 12 delta cepstrum coefficients. The cepstrum is derived from a 10th order LPC analysis over a 30 ms window. The feature vectors are updated at 10 ms intervals.

In a training session, 5 utterances collected from an enrollment phone call are used to train a left-to-right HMM, called target model, $\lambda_t$. Due to the limitation on the model size, the number of states is estimated based on 10 frames per sate. There are 4 Gaussian components associated with

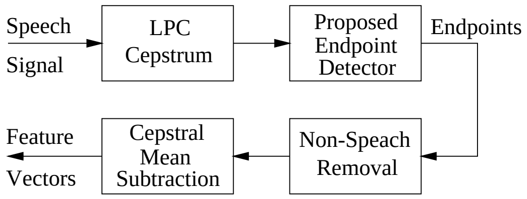

Figure 5: The front-end processor

each state. Also, due to unreliable variance estimates from limited amount of training data, a global variance estimate is used as a common variance to all Gaussian components [7].

The evaluation was separated into two groups, male and female, and speakers were tested within the same gender. For each male speaker, ten utterances from the true speaker collected from 5 different sessions and $17 \times 5 = 85$ utterances from impostors were used for testing. For each female speaker, there are 10 utterances from the true speaker while the number of impostors' utterances are $19 \times 5 = 95$.

In a language-independent configuration, after CMS, the feature observation **O** is decoded by the target model to get a likelihood score

$$L(\mathbf{O};\lambda_t) = \frac{1}{N}\log P(\mathbf{O}|\lambda_t), \tag{5}$$

where $N$ is the total number of feature vectors, $P(\mathbf{O}|\lambda)$ is the accumulative likelihood score computed by forced decoding. The decision on acceptance or rejection is made by comparing the score with a pre-determined threshold value.

When an application needs a higher level of security, in addition to the target model, a background model with likelihood ratio test can be applied to speaker verification [8, 7]. Since lexicon is needed for accurate HMM decoding, the configuration is language dependent.

The evaluation results are shown in Table 1. The accuracy is in the same level as the speaker verification system [7] where HMMs were applied for endpoint detection. The system performance can be further improved if model adaptation is allowed.

Table 1: **Speaker Verification Performance on Average Individual Equal-Error Rates**

| System | Language- | Language- |
|---|---|---|
| Configurations | Independent | Dependent |
| | (w/o BK models) | (with BK models) |
| 18 Males | 3.6% | 2.0% |
| 20 Females | 4.4% | 3.5% |
| Average | 4.0% | 2.8% |

## 4. CONCLUSIONS

We have proposed a fast, efficient algorithm for locating the endpoints of an utterance with variety of noise, such as heavy breath, dial tone, clicks, etc. The detection decision is based on matched filter responses and statistical measures of spectral energy. The experiments showed that proposed algorithm provided the similar verification accuracy as the systems using HMM for endpoint detection, while the proposed algorithm is much faster and can support language-independent applications. The algorithm has been applied to develop a real system for language-independent voice control with embedded speaker verification [5]. The proposed algorithm can be applied to provide accurate endpoints for CMS in speech recognition and for other applications. Also, the proposed algorithm can be extended to real-time endpoint detection.

## 5. ACKNOWLEDGMENT

The authors wish to thank Donald W. Tufts, Weon-Goo Kim, Chin-Hui Lee, and Frank K. Soong for useful discussions.

## 6. REFERENCES

[1] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoustics, speech, and signal process.*, vol. ASSP-29, August 1981.

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

[3] 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.

[4] B. S. Everitt and D. J. Hand, *Finite mixture distributions*. New York: Chapman and Hall, 1981.

[5] Q. Li and A. Tsai, "A language-independent personal voice controller with embedded speaker verification," in *Eurospeech'99*, (Budapest, Hungary), Sept. 1999.

[6] R. A. Sukkar and C.-H. Lee, "Vocabulary independent discriminative utterance verification for non-keyword rejection in subword based speech recognition," *IEEE Trans. Speech and Audio Process.*, vol. 4, Nov. 1996.

[7] S. Parthasarathy and A. E. Rosenberg, "General phrase speaker verification using sub-word background models and likelihood-ratio scoring," in *Proceedings of ICSLP-96*, (Philadelphia), October 1996.

[8] A. E. Rosenberg and S. Parthasarathy, "Speaker background models for connected digit password speaker verification," in *Proceedings of ICASSP*, May 1996.