# A Robust and Sensitive Word Boundary Decision Algorithm

*Jong Uk Kim, SangGyun Kim and Chang D. Yoo*

Department of Electrical Engineering and Computer Science
Korea Advanced Institute of Science and Technology
jukim@kaist.ac.kr, zom@eeinfo.kaist.ac.kr, cdyoo@ee.kaist.ac.kr

## Abstract

A robust and sensitive word boundary decision algorithm for automatic speech recognition (ASR) system is proposed. The algorithm uses a time-frequency feature to improve both robustness and sensitivity. The time-frequency features are passed through a bank of moving average filters for temporary decision of word boundary in each band. The decision results of each band are then passed through a median filter for the final decision. The adoption of time-frequency feature improves the sensitivity, while the median filtering improves the robustness. Proposed algorithm uses an adaptive threshold based on the signal-to-noise ratio (SNR) in each band which further improves the decision performance. Experimental result shows that the proposed algorithm outperforms the Q.Li et al's robust algorithm.

## 1. Introduction

An accurate decision of word boundary is one of the most important factors in the design of high performance speech recognition systems, and its importance is well addressed in various literatures [1,2]. The word boundary can be accurately obtained in a clean environment; however, its accuracy is greatly reduced in noisy environment.

Recently, a robust word boundary decision algorithm was proposed by Q.Li et al [4], which adopted a ramp edge decision algorithm [3] used in image processing. Although more robust than previously suggested algorithms, the algorithm lacks sensitivity. As a way to improve the sensitivity, an algorithm using time-frequency features which is based on mel-frequency scale is incorporated in [5]; however, the role of the time-frequency feature is only to remove low SNR parts in each band in order to obtain less contaminated feature. Therefore, the improvement of the accuracy is limited. So is it in [6], where the feature is the weighted sum of wavelet coefficients.

Improving both robustness and sensitivity in word boundary decision is a difficult problem when only one feature per frame is used. To improve sensitivity, we propose a time-frequency feature. In order to process the feature, we use a bank of moving average filters so that the temporary decision of word boundary can be made in each band independently. To improve the robustness, the decision of each band is passed through the median filter. The median filter interrelates the independently decided word boundary with that of neighboring bands and frames, since all of them are related together.

The organization of the paper is as follows. A robust algorithm proposed in [4] is briefly explained in Section 2. To improve both robustness and sensitivity, we propose a time-frequency algorithm in Section 3. The experimental results are given in Section 4. Finally, we concludes in Section 5.

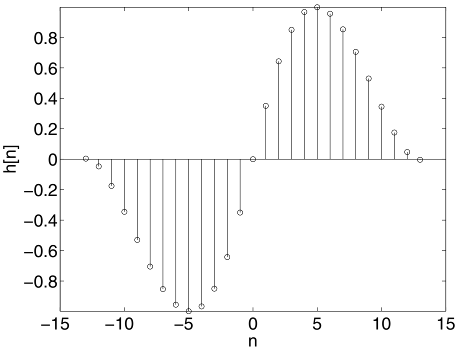

Figure 1: *Moving average filter $h[n]$ for robust algorithm with* $W = 13$.

## 2. A robust algorithm

In this section, a robust algorithm based on log energy and optimal filtering in [4] is explained. A log energy of $n$-th frame is given by

$$g[n] = 10 \log_{10} \sum_{i=0}^{I-1} x_n^2[i] \tag{1}$$

where $x_n[i]$ and $I$ denote data sample at $n$th frame and frame size respectively. This feature is then passed through the moving average filter of the form

$$h[n] = \begin{cases} -h_+[n] & -W \leq n \leq 0 \\ h_+[n] & 1 \leq n \leq W \end{cases} \tag{2}$$

where

$$\begin{aligned} h_+[n] = {} & e^{An}[K_1 \sin(An) + K_2 \cos(An)] \\ & + e^{-An}[K_3 \sin(An) + K_4 \cos(An)] \\ & + K_5 + K_6 e^{sn}. \end{aligned} \tag{3}$$

For $W = 13$, $s = 7/W = 0.5385$, $A = 0.41s = 0.2208$, and $[K_1, \cdots, K_6] = [1.583, 1.468, -0.078, -0.036, -0.872, -0.56]$ as given

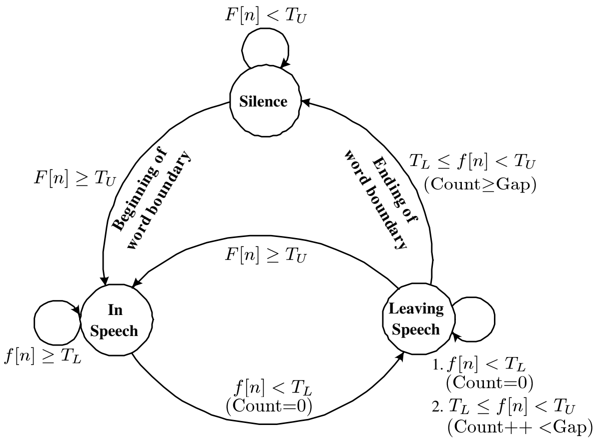

Figure 2: *State transition diagram for robust algorithm. The state transition diagram consists of three states, namely, 'Silence', 'In Speech' and 'Leaving Speech'.*

in [3,4]. The filter shape of $h[n]$ for $W = 13$ is given in Fig. 1. The filtering equation is given by

$$f[n] = \sum_{i=-W}^{W} h[i]g[n+i]. \tag{4}$$

Observing the filter, we know that $h[n]$ acts as a lowpass filter followed by a highpass filter. The word boundary is detected using a state transition algorithm shown in Fig 2 [4], and appropriately choosing upper threshold ($T_U$), lower threshold ($T_L$) and Gap. In the diagram, three states called '*Silence*', '*In Speech*' and '*Leaving Speech*' are given. The state transits from *Silence* state to *In Speech* state when $f[n]$ exceeds $T_U$, again transits to *Leaving Speech* state when $f[n]$ undergoes $T_L$. From the *Leaving Speech* state, the state can transit either to *In Speech* state or to *Silence* state depending on the value of $f[n]$ and counter according to the conditions given in the Fig 2.

## 3. Time-frequency algorithm

There exists a tradeoff between robustness and sensitivity in word boundary decision algorithm. In the robust algorithm [4], it is affected by the filter length and threshold. Increasing the robustness decreases the sensitivity and vice versa. To increase sensitivity, we proposed an algorithm based on time-frequency feature instead of simply log energy algorithm. To increase robustness, we propose median filtering. The median filtering suppresses the sensitivity which has been emphasized by using time-frequency feature.

### 3.1. Time-frequency filtering

The time-frequency feature based on mel-frequency scale for word boundary decision is introduced in [5], where the time-frequency feature is only used to get rid of the frequency band with low SNR. The final feature is just a frame energy. In this paper, a time-frequency feature is processed independently in each frequency band. After independent processing, the output of each band is passed through median filter to consider the effect of the neighboring band and neighboring frame.

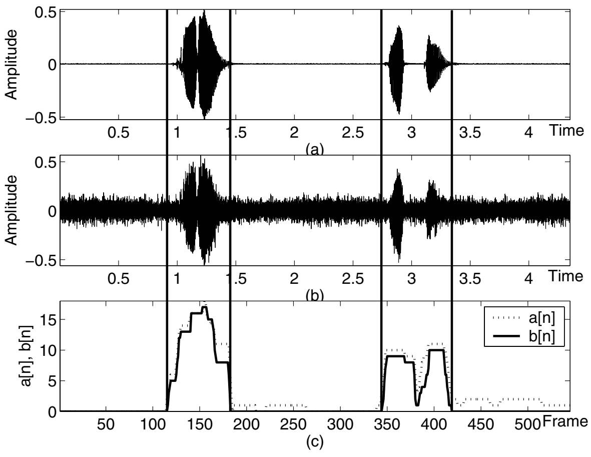

Figure 3: *Word boundary decision example through median filtered output $b[n]$. (a) The time waveform of clean speech. (b) The time wave form of noisy speech contaminated by 0dB helicopter noise (c) The summation of flags, $a[n]$ (dotted line) and $b[n]$ (solid line). The vertical lines passing (a) through (c) indicate the detected beginning and ending positions of speech.*

The time-frequency feature is defined as

$$G[m,n] = \frac{|X[m,n] - \bar{w}[m,n]|}{\bar{w}[m,n]} \tag{5}$$

where $X[m,n]$ and $\bar{w}[m,n]$ are the energy and normalization factor in $m$th band and $n$th frame respectively. Throughout this paper, (i) when double indices are used, the first and second index denote frequency band number and frame number respectively, and (ii) when single index is used, it denotes frame number. Note that $G[m,n]$ is actually an estimate of SNR. The energy $X[m,n]$ is defined as

$$X[m,n] = \sum_{k=(m-1)p}^{(mp-1)} |X_n[k]|^2, \quad m = 1, 2, \cdots, M \tag{6}$$

where $X_n[k]$, $M$ and $p(= I/2M)$ are the discrete Fourier transform (DFT) of $x_n[i]$, the number of bands in each frame and the number of Fourier coefficients in each band respectively. The normalization factor in Eqn. (5) is given by averaging the $J$ smallest energies of the recent non-speech frames up to $n$th frame in $m$th band. So $\bar{w}[m,n]$ can be considered as the energy of noise.

By filtering $G[m,n]$ through $h[i]$, we obtain the output

$$F[m,n] = \sum_{i=-W}^{W} h[i]G[m,n+i]. \tag{7}$$

Setting $T_U$ and $T_L$ appropriately, we can determine if the frame in each band is speech frame or non-speech frame, following the state transition diagram of Fig. 2. The output by the state transition diagram we denote $A[m,n]$. The element of $A[m,n]$ is either 1 for speech frame or 0 for non-speech frame, that is $A[m,n]$ is a matrix whose elements are composed of speech

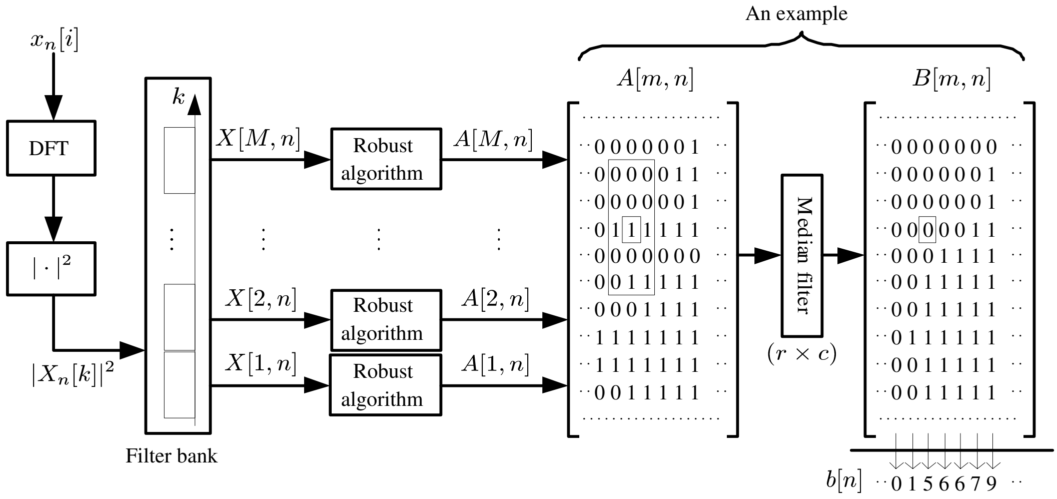

Figure 4: *The Overall system of proposed algorithm. An example using $A[m,n]$ and $B[m,n]$ are given to illustrate the effect of median filtering.*

and non-speech flag. Passing $A[m,n]$ through median filter of size $(r \times c)$, we obtain $B[m,n]$ as

$$B[m,n] = \underset{(r\times c)}{\text{median}}\{A[m,n]\}. \quad (8)$$

By setting both $r$ and $c$ to be odd, $\frac{r-1}{2}$ and $\frac{c-1}{2}$ are even. $B[m,n]$ is set to 1 when the number of 1's enclosed by 4 points

$$A\Big[m-\frac{r-1}{2}\quad,\quad n-\frac{c-1}{2}\Big],$$
$$A\Big[m-\frac{r-1}{2}\quad,\quad n+\frac{c-1}{2}\Big],$$
$$A\Big[m+\frac{r-1}{2}\quad,\quad n-\frac{c-1}{2}\Big],$$
$$A\Big[m+\frac{r-1}{2}\quad,\quad n+\frac{c-1}{2}\Big]$$

is greater than the number of 0's. Otherwise $B[m,n]$ is set to 0. The column sum of $A[m,n]$ and $B[m,n]$ we denote $a[n]$ and $b[n]$ respectively ,that is,

$$a[n] \quad = \quad \sum_{m=1}^{M} A[m,n], \quad (9)$$

$$b[n] \quad = \quad \sum_{m=1}^{M} B[m,n], \quad (10)$$

which is the summation of flags in each band.

The decision rule becomes as;

$$\boxed{\textit{n-th frame} = \begin{cases} \textit{Speech frame}, & \textit{if}\ \ b[n] \geq 1 \\ \textit{Non-speech frame}, & \textit{if}\ \ b[n] = 0 \end{cases}}$$

Proposed algorithm poses an equal priority regardless of whether the band has small energy or large energy. So the algorithm can sensitively pick the low SNR parts of the speech, which has been very difficult problem in the robust algorithm of [4]. Examples of median filtered output is given in Fig. 3. In the figure vertical solid lines indicate the beginning and ending position of word boundary obtained by decision rule. Note that the falsely alarmed frame in some filter bank is disappeared after median filtering.

The overall system of the proposed algorithm is given in Fig. 4. The two matrices at the right side of the system are given to illustrate the effect of the median filtering. Note that after median filtering, either the speech frames or the non-speech frames conglomerate together.

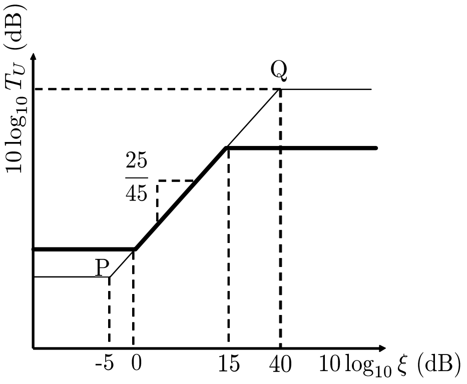

Figure 5: *The relation between input SNR $\xi$ and upper threshold $T_U$.*

### 3.2. Threshold adjustment

Setting an appropriate value of the threshold in robust word boundary decision algorithm [4] is always difficult. And this is usually done heuristically. In this paper, a rule for adjusting the threshold according to SNR is proposed. According to the experiment, the relation between input SNR $\xi$ and upper threshold $T_U$ is shown in Fig. 5. The horizontal and vertical axes are set to dB scale. Two points at the terminal of linear region are $P(-5, 10\log_{10} T_U(-5))$ and $Q(40, 10\log_{10} T_U(40))$ re-

Table 1: *Frame error rate. Both false alarm and false rejection rate (in %) are given separately.*

| Algorithm | | Type of noise | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | WGN | | | HEL | | | PS2 | | |
| | | 5dB | 10dB | 15dB | 5dB | 10dB | 15dB | 5dB | 10dB | 15dB |
| Robust algorithm | False alarm | 0.36 | 0.50 | 0.58 | 0.44 | 0.44 | 0.50 | 0.29 | 0.47 | 0.53 |
| | False rejection | 5.58 | 4.17 | 3.71 | 6.76 | 4.83 | 4.13 | 5.88 | 4.27 | 3.87 |
| Proposed algorithm | False alarm | 0.27 | 0.24 | 0.27 | 0.70 | 0.63 | 0.46 | 1.58 | 1.26 | 0.83 |
| | False rejection | 4.13 | 3.00 | 2.25 | 4.27 | 3.27 | 2.65 | 2.77 | 2.61 | 2.45 |

spectively. When the input SNR varies from -5dB to 40dB the threshold varies about 25∼35dB, which we will approximate to 25dB to take the minimum extreme. So the slope is $\frac{25}{45}$. Thus the upper threshold is given by

$$T_U = T_U(0)\xi^{\frac{25}{45}}. \tag{11}$$

As shown in Fig. 5 as thick solid line, the minimum and maximum threshold that $T_L$ can have is bounded by the value at 0dB and 15dB. So $T_U$ must hold $0 \leq 10\log_{10} T_U \leq 15$. Once $T_U$ is determined, the lower threshold is set to $T_L = -0.8T_U$. The number $0.8$ is determined experimentally.

Now, in order to find $T_U(0)$ which is the upper threshold at input SNR 0dB, we must consider the ramp edge model indicated in [3,4]. The ramp edge is modelled by

$$c[n] = \begin{cases} 1 - e^{-sn}/2, & n \geq 0 \\ e^{sn}/2, & n \leq 0 \end{cases} \tag{12}$$

where $s = 7/W$ with $W$ being set to 13, and $n$ is the frame number. This model can be considered to be the 0dB model where the speech energy is equal to noise energy. Filtering $c[n]$ through $h[n]$ gives a bell shaped curve with maximum value 6.5715 which corresponds to the edge in ramp edge model. To find edge, we set $T_U(0) = 6.5715$.

## 4. Experimental results

For evaluation, speech samples of 10 Korean speakers' (5 males' and 5 females') sampled at 16KHz are collected in office environment. The environmental noise consists only of very low fan and computer noise. We call these as clean speech. In the experiment, we artificially contaminate the speech samples with three kinds of noise, namely, white Gaussian noise (WGN), helicopter noise (HEL), IBM PS2 fan noise (PS2) with varying input SNRs; 5dB, 10dB and 15dB for each type of noise. The median filter size is set to $r = 9$, $c = 5$. The frame error rate is shown in Table 1. Note two aspects of the results; (i) The proposed algorithm works well not only in high SNR environment but also in low SNR environment. This justifies the robustness. (ii) The reduction of false rejection rate is higher than that of false alarm rate. This justifies the sensitivity of the proposed algorithm.

## 5. Conclusions

A robust and sensitive word boundary decision algorithm for isolated ASR is proposed. In the robust algorithm, log energy feature is used. However, to improve both robustness and sensitivity, we propose a time-frequency feature. Energies extracted on a frame-by-frame basis and band-by-band basis are used as time-frequency features. The time-frequency feature thus obtained is passed through the moving average filter independently to obtain word boundary in each band. The word boundary of each band is passed through the median filter, which is based on dominant decision. The reason why we apply the median filtering is that the output of each band is interrelated with neighboring bands and neighboring frames, which has been overlooked in the band based independent processing. The proposed algorithm uses an adaptive threshold based on the input SNR at each band to further improve the decision performance. Experimental result shows that the proposed algorithm outperforms the robust algorithm of Q.Li et al. The proposed algorithm shows low false rejection rate which is a desirable characteristic for ASR.

## 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 Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 4, pp. 777–785, August 1981.

[2] J.Taboada, S.Feijoo, R.Balsa, C.Hernandez, "Explicit estimation of speech boundaries," *IEE Proceedings on Science, Measurement and Technology*, vo. 141, no. 3, pp. 153–159, May 1994.

[3] M.Petrou and J.Kittler, "Optimal edge detectors for ramp edges," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 5, pp. 483–491, May 1991.

[4] Q.Li, J.Zheng, A.Tsai and Q.Zhou, "Robust endpoint detection and energy normalization for real-time speech and speaker recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 3, pp. 146–157, March 2002.

[5] G.D.Wu and C.T.Lin, "Word boundary detection with mel-scale frequency bank in noisy environment," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 541–554, September 2000.

[6] J.W.Seok and K.S.Bae, "A novel endpoint detection using discrete wavelet transform," *IEICE Transactions on Information and Systems*, vol. E82-D, no. 11, pp. 1489–1491, November 1999.