Adaptive Band Selection for Robust Speech Detection in Noisy Environments

Mikyong Ji(ICU), Youngjoo Suh(ICU), Hoirin Kim(ICU)

<Contents>

1. Introduction
2. Speech Detection by Adaptive Band Selection
  2.1. Modified Adaptive Time-Frequency Parameter
  2.2. Adaptive Band Selection
    2.2.1. Noise-Centric Band Selection
    2.2.2. Update on Noise-Centric Bands
  2.3. Classification of Frames
    2.3.1. Automation of Threshold Setting
  2.4. Speech Boundary Detection
3. Experimental Results
4. Conclusion
5. References

<Abstract>

Adaptive Band Selection for Robust Speech Detection in Noisy Environments

Mikyong Ji, Youngjoo Suh, Hoirin Kim

One of the important problems in speech recognition is to accurately detect the existence of speech in adverse environments. The speech detection problem becomes severer when recognition systems are used over the telephone network, especially in a wireless network and a noisy environment. In this paper, we propose a robust speech detection algorithm, which detects speech boundaries accurately by selecting useful bands adaptively to noisy environments. The bands where noises are mainly distributed, so called, noise-centric bands are introduced. In this paper, we compare two different speech detection algorithms with the proposed algorithm, and evaluate them on noisy environments. The experimental results show the excellence of the proposed speech detection algorithm.

* Keywords: Speech Detection, Speech Boundary Detection, Endpoint Detection

1. Introduction

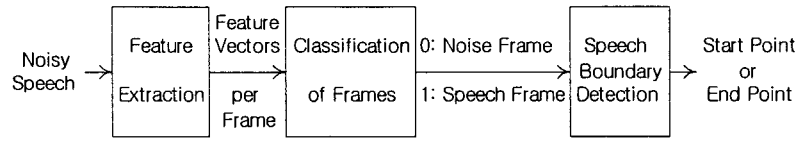
The accurate detection of speech boundaries is crucial to the performance of speech recognition. It is called a robust endpoint location problem. In this paper, we especially focus on reliable real-time speech detection. The importance of the speech detector has been proved out in isolated-word automatic speech recognition. The energy (in time domain), zero-crossing rate, and duration parameters have been usually used to find speech boundaries between speech signal and background noise [1]-[4]. It is even more difficult to accurately locate the start and end point of the speech segment in noisy environments, but it is definitely necessary for robust speech recognition. If the speech detector is able to locate the speech boundary exactly, the recognizer can save the resources that are used in processing non-speech that usually exists between speech boundaries. It also makes the response-time faster.

In this paper, we introduce a robust speech detector, which is frame-based. It classifies each frame as speech or non-speech, and locates a start or end point depending o-n the classification results of a sequence of frames. The feature vector that the propos-ed speech detector uses for frame classification can be produced in the process of computing mel frequency cepstrum coeficients (MFCC) in the feature extraction module for speech recognition. In other words, the feature vectors employed in speech detection can be re-used in speech recognition. It implies that it can have the response-time even faster.

In Section 2, we describe the proposed speech boundary detection algorithm in detail from the feature extraction module to the speech detection module. In Section 3, the performance of three different speech detection algorithms is compared, and evaluated on several noisy databases. And also their experimental results are shown in the same section.

2. Speech Detection by Adaptive Band Selection

All speech boundary detection algorithms considered in this paper are frame-based, that is, they classify noisy speech input as either a speech frame or a noise frame, and after that, they decide speech boundaries based on the classification results of a sequence of frames.



<Figure 1> Block diagram of frame-based speech detection.

The basic structure of frame-based speech detection is depicted in <Figure 1>. The time samples of the observed noisy signal are inputted into the feature extraction module whose outputs are feature vectors. The feature vectors are classified as either speech or non-speech by the classification module, and then the speech boundary detection module detects start/end points of speech depending on the pattern of the sequence of frames classified as speech or non-speech frames.

2.1. Modified Adaptive Time-Frequency Parameter

In this section, we describe a feature vector that is used for robust speech detection and also obtained by the feature extraction module in <Figure 1>. The feature vector is based on an Adaptive Time-Frequency (ATF) parameter [5]. The ATF parameter is composed of the multi-band frequency energy and the logarithm of the rms energy [5]. We only use the modified multi-band frequency energy, which makes clear distinction between speech and noise, so as to locate exact boundaries of speech.

Given a time-domain noisy speech, we obtain the spectrum of the signal, $x_{freq}(m,k)$ by a series of processes such as preprocessing, windowing, and FFT, where m is a frame number, and k is a spectrum index. After that, the only absolute value of the spectrum, $|x_{freq}(m,k)|$, is taken, and the energy of each Mel frequency band, $x(m,i)$, is calculated by applying Mel-scale filter bank, $f(i,k)$, to that, where i is a filter bank index. It is as follows:

$$x(m,i) = \sum_{k=0}^{N-1} |x_{freq}(m,k)| f(i,k) \tag{1}$$

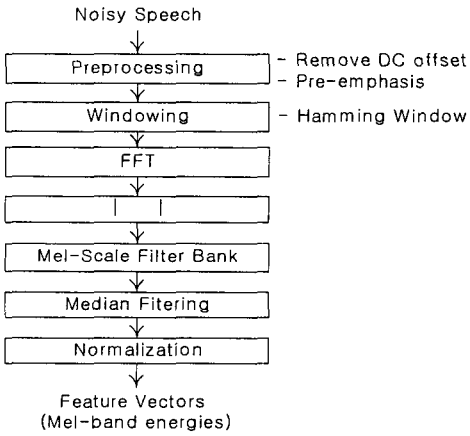
By making use of 3-point Median filtering and Normalization processes in order in Equation (2)-(3), finally, we can get a feature vector, which some undesired impulse

noise and stationary background noise are eliminated from. This feature vector is used for robust speech boundary detection. The block diagram about the entire process to find the feature vector for speech detection is shown in <Figure 2>.

$$\hat{x}(m,i)=\frac{x(m-1,i)+x(m,i)+x(m+1,i)}{3} \tag{2}$$

$$X(m,i)=\hat{x}(m,i)-\frac{\sum_{m'=0}^{S-1}\hat{x}(m',i)}{S} \tag{3}$$

where S is the number of frames that are regarded as silence in initial part of a utterance.



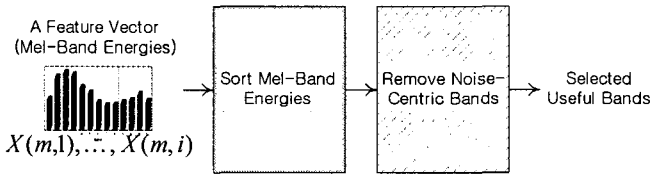
<Figure 2> Block diagram of the feature extraction module.

2.2. Adaptive Band Selection

In the previous section, we described how to obtain a feature vector for robust speech boundary detection. Once the feature vector is given, useful bands are chosen and those selected useful bands are employed to locate exact speech boundaries.

The Mel-Frequency band energies, which are computed by the procedure shown in <Figure 2>, are sorted, and certain bands where noises are mainly distributed, are removed from those bands. According to the type or characteristics of noises, the degree that a noise affects to each band is different and certain noise tends to be

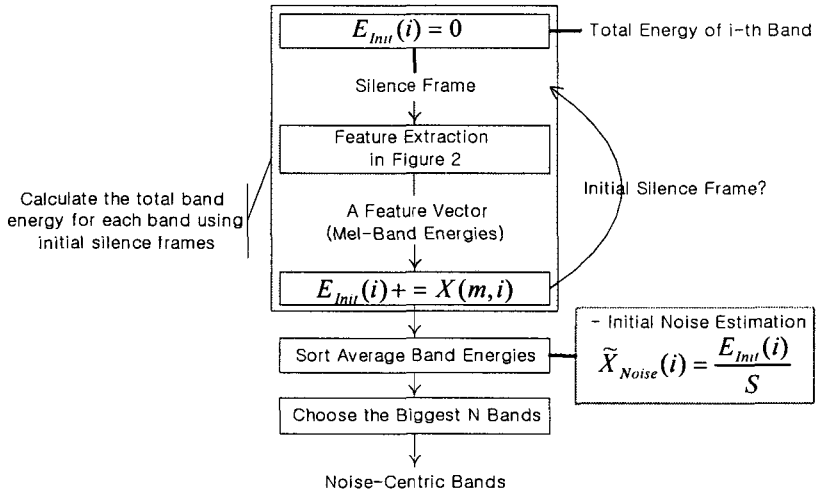
mainly distributed in a specific bands. Therefore, those bands where noises are intensively distributed should be eliminated from useful bands. We call those noisy bands noise-centric bands. Thus, it is possible to detect exact speech boundaries by only using the Mel-Frequency bands which correspond to the selected useful bands within a feature vector (Mel-Frequency band energies).



<Figure 3> Block diagram of adaptive band selection.

2.2.1 Noise-Centric Band Selection

The procedure to acquire the noise-centric bands is described in <Figure 4> in detail. The noise-centric bands, which are utilized in order to select useful bands, are initially obtained from early frames in utterances that are usually considered as silence. The measurement to find the noise-centric bands is based on the noise estimation of each Mel-band. And the noise estimation is updated by frames that are classified as non-speech (noise). The average band energies of early silence frames are computed, and they are used as initial noise estimations. The bands of which energy values belong to the greatest N among those average band energies are picked as initial noise-centric bands.



<Figure 4> Selection of noise-centric bands with initial silence frames.

The computation of the initial noise estimation is shown in Equation (4)-(5).

$$E(i) = \sum_{m=0}^{S-1} |X(m,i)| \tag{4}$$

$$\tilde{X}_{noise}(i) = \frac{\sum_{m=0}^{S-1} |X(m,i)|}{S} \tag{5}$$

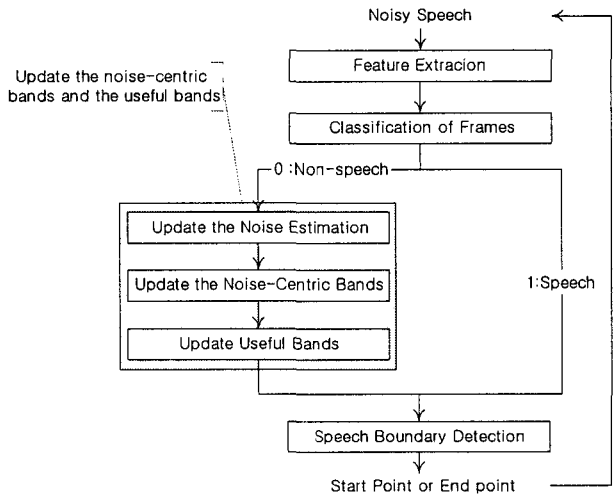
The noise-centric bands are updated depending on the classification results of frames. If the current frame is classified as non-speech, the noise estimation, which is initially estimated in early silence frames, is re-estimated. As a result, the noise-centric bands are updated. The initial noise estimation is shown in Equation (5).

2.2.2 Update on Noise-Centric Bands

The noise-centric bands are updated depending on whether the classification result of the current frame is speech or non-speech. If the current frame is categorized as non-speech, then the noise estimation is updated in Equation (6), and the noise-centric bands are updated by sorting the noise estimations of the bands, and taking the bands whose energy values are among the greatest *N*. In consequence, the useful bands which

noise-centric bands are removed from are adaptively selected and the bands that correspond to the useful bands are only employed for robust speech detection. The detailed description of update on noise-centric bands is shown in <Figure 5>.

$$\tilde{X}_{Noise}(i) = (1-\alpha)\cdot\tilde{X}_{Noise}(i) + \alpha\cdot X(m,i), \quad 0 \le \alpha \le 1 \tag{6}$$



<Figure 5> Block diagram of update on noise-centric bands.

2.3. Classification of Frames

The bands that belong to useful bands are only used in classifying whether a given frame is speech or non-speech. If the percentage of bands whose energies are greater than their thresholds among the useful bands is greater than given threshold, the current frame is classified as a speech frame. Otherwise, it is regarded as non-speech. The algorithm is shown in <Figure 6>.

```
for (i = 0; i < NBand; i++)
    if (i ∈ Useful bands) and (X(m, i) > X_Thrshld(i))
        NBandCount ++;
if ((NBandCount / NBand) × 100) > Percent_Thrshld
    Current Frame ← Speech
else
    Current Frame ← Non - speech
```

<Figure 6> Algorithm of frame-classification.

2.3.1. Automation of Threshold Setting

Most speech detection algorithms set thresholds heuristically through experiments. Here, these thresholds should be adequately changed according to the characteristics of background noises or their experimental environments in order to achieve good performance. However, most algorithms set their thresholds only with early frames that exist in an initial part of utterance and are usually considered as silence.

The proposed algorithm updates a noise estimation of each band when the current frame is classified as non-speech, and also automatically changes its threshold using the updated noise estimation. The update on noise estimation is shown in the previous section, and that on threshold is in Equation (7).

$$X_{Thrshld}(i) = \beta \cdot \tilde{X}_{Noise}(i) \qquad (7)$$

2.4. Speech Boundary Detection

The speech boundary detection module locates speech boundaries (start points or end points) depending on the result of the classification module, that is, the pattern of a set of frames that are classified as speech or non-speech.

3. Experimental Results

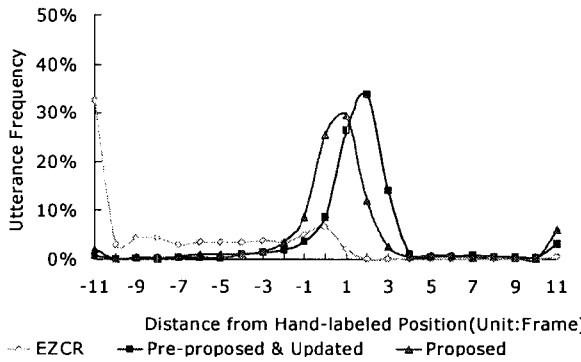
The evaluation of three different speech detection algorithms is performed by measuring the distance from the hand-labeled location to the detected speech boundary. The types of database used in the evaluation are shown in <Table 1>.

<Table 1> Database used in the evaluation

Type of Database	No. of Utterances
TW (wired Telephone Word)	1000
TD (wired Telephone Digit)	1000
MW(wireless Mobile phone Word)	1000
MD(wireless Mobile phone Digit)	1000

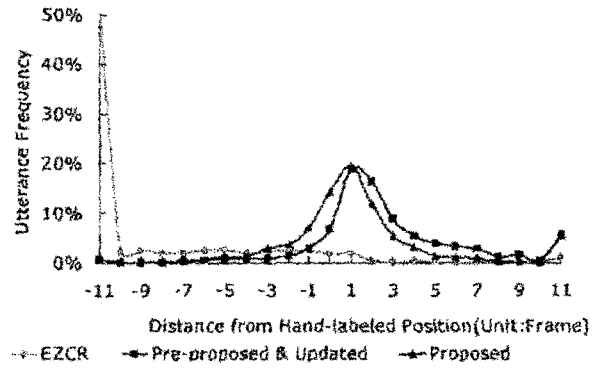
The sampling rate of the databases is 8 KHz, and each utterance was recorded by a different speaker. The SNR of the database are relatively high (more than 20 dB), but their energy levels are very different utterance by utterance, and also many of those utterances include human noises especially in early frames of them. Later, this becomes the reason why the performance of the speech boundary detection algorithms is generally low although the SNR of them are relatively high.

In the paper, the performance of the three different speech detectors, the energy-ZCR based speech detector, the sub-band selection-based one [6], and the proposed one, are evaluated and compared. The experiment was performed with clean DB, and also with noisy DB that the clean DB is mixed with one of a car, subway, and street noise. The Energy-ZCR based speech detector, the sub-band selection-based one, and the proposed one are represented by '*EZCR*', '*Pre-proposed*', and '*Proposed*' respectively.



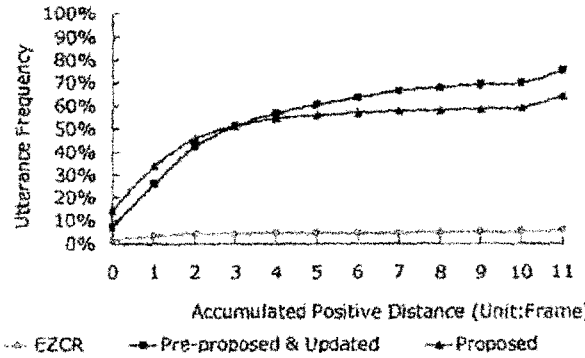
<Figure 7> Distance distribution of the detected start-point from the hand-labeled start-point location.

The distance from the hand-labeled location to the detected start or end point location was measured in frames. It is depicted in <Figure 7> and <Figure 8>. It is shown that the proposed speech detector and the sub-band selection-based one [6] outperform the energy-ZCR based one. In the figures, if the distance of the start point is -3, it means the start point is detected at 3 frames after the exact location. Likewise, if that of the end point is -3, it means the end point is detected at 3 frames before the exact location.

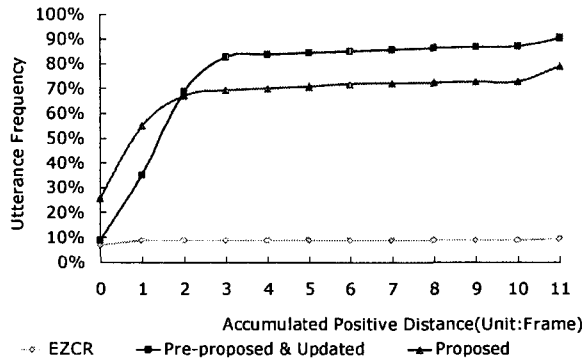


<Figure 8> Distance distribution of the detected end-point from the hand-labeled end-point location.

<Figure 9> and <Figure 10> show the positive accumulated distance from the labeled location to the detected start or end point location respectively. This time, we eliminated the case that a speech detector detects the start point after its exact location or an end point before that. If the distance is 3 in <Figure 9>, it implies the case that start points are detected within 3 frames before their exact locations.



<Figure 9> Positive accumulated distance of the detected start-point from the hand-labeled start-point location.



<Figure 10> Positive accumulated distance of the detected end-point from the hand-labeled end-point location.

<Table 2> shows the accuracy that the start point is detected within 5 frames before its exact location, and the end point is detected within 5 frames after its exact location. Like in <Figure 7-10>, Sub-band Selection based algorithm [6] and the proposed one shows outstanding performance compared with Energy-ZCR based one.

<Table 2> Accuracy that the speech boundaries are detected within +5 frames from labeled location (unit: percent)

	DB / Detector	TW Clean	TW Str15	TD Clean	TD Sub15	MW Clean	MW Car15	MD Clean	TW Car15
Start	EZCR	8.7	4.4	0.2	0.2	0.0	0.0	0.3	1.1
	Pre. Pro.	84.3	50.5	88.5	33.9	74.5	26.2	59.7	11.5
	Pro.	70.7	36.4	77.2	42.7	39.8	22.0	36.4	15.4
End	EZCR	4.8	2.6	0.1	0.2	0.0	0.1	0.8	1.4
	Pre. Pro.	60.5	15.7	65.0	11.5	81.0	9.8	55.2	8.5
	Pro.	55.9	11.6	60.9	11.6	54.3	12.5	50.1	10.4

Clean (As it is), Str15 (Street noise added, SNR 15dB), Sub15 (Subway noise added, SN R 15dB), Car15 (Car noise added, SNR 15dB)

4. Conclusions

In this paper, we compared the performance of three different speech boundary detection algorithms. The performance of those algorithms was evaluated on 4 types of DB, and the noisy DB that one of a car, subway, and street noise was added to them.

We have shown that the proposed speech detector and the sub-band selection-based one [6] outperform the Energy-ZCR based one. And the performance of the sub-band selection-based speech detector shows slightly better performance than the proposed one: the sub-band selection-based speech detector selects useful bands, which are especially effective in speech boundary detection, through training before actual speech detection (See [6]). It is hard to say that the algorithm works in real-time. In case of the proposed algorithm, it adaptively chooses the useful bands that are proper to its noisy environment by removing the noise-centric bands, where noises are mainly distributed, and then detects speech boundaries in real-time. In other words, the processes of band selection and speech detection happen simultaneously. None the less, it shows comparable results.

References

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the end-points of isolated utterances," *Bell Syst. Tech. J.*, vol. 54, pp.297-315, Feb. 1975.

[2] M.H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, vol. 8, pp.45-60, 1989.

[3] L. Lamel et al., "An improved endpoint detector for isolated word recognition," *IEEE ASSP Mag.*, vol. 29, pp.777-785, 1981.

[4] B. Reaves, "Comments on an improved endpoint detector for isolated word recognition," *IEEE Trans., Signal Processing*, vol. 39, pp.526-527, Feb. 1991.

[5] G. D. Wu and C. T. Lin, "Word Boundary detection with Mel-scale frequency bank in noisy environment," *IEEE Trans., Speech and Audio Processing*, vol. 8, no.5, pp.541-554, Sept. 2000.

[6] Mikyong Ji and Hoirin Kim, "Sub-band Selection-based Speech Detector for Robust Keyword Recognition under Noisy Environment," *Proc. of the 4th IASTED Int. Conf. on SIP*, pp.162-166, Aug. 2002.

접수일자: 2004년 5월 3일
게재결정: 2004년 6월 7일

▶ 지미경(Mikyong Ji)
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술연구실
전화: 042) 866-6221
E-mail: lindaji@icu.ac.kr

▶ 서영주(Youngjoo Suh)
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술연구실
전화: 042) 866-6221
E-mail: yjsuh@icu.ac.kr

▶ 김회린(Hoirin Kim)
주소: 305-714 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교(ICU) 음성인식기술연구실
전화: 042) 866-6139
E-mail: hrkim@icu.ac.kr