THE JOURNAL OF THE ACOUSTICAL SOCIETY OF KOREA Vol.22, No.3E 2003. 9. pp. 113~123

Robust Speech Detection Based on Useful Bands for Continuous Digit Speech over Telephone Networks

Mi-Kyong Ji*, Young-Joo Suh*, Hoi-Rin Kim*, Sang-Hun Kim**
*School of Engineering, Information and Communications University
**Speech Information Technology Center, ETRI
(Received February 17 2003; accepted August 20 2003)

Abstract

One of the most important problems in speech recognition is to detect the presence of speech in adverse environments. In other words, the accurate detection of speech boundary is critical to the performance of speech recognition. Furthermore the speech detection problem becomes severer when recognition systems are used over the telephone network, especially wireless network and noisy environment. Therefore this paper describes various speech detection algorithms for continuous digit recognition system used over wire/wireless telephone networks and we propose a algorithm in order to improve the robustness of speech detection using useful band selection under noisy telephone networks. In this paper, we compare some speech detection algorithms with the proposed one, and present experimental results done with various SNRs. The results show that the new algorithm outperforms the other speech detection methods.

Keywords: *Robust speech detection, Speech boundary detection, Endpoint detection, Endpoint of continuous speech*

I. Introduction

The accurate detection of speech boundaries is crucial to the performance of speech recognizer. It is called robust endpoint location problem. In this paper, we especially focus on reliable real-time speech detection for continuous digit speech over telephone networks. 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 the boundary between the word signal and background noise[1-4]. If the speech detector is able to locate the boundary of the speech exactly, the recognizer can save the resources that are used in processing silence that is usually included before and after speech. It also makes the response-time faster.

Real-time speech detection is necessary for real-time digit recognition over telephone networks. In this paper, we introduce a speech detector that can extract well the speech boundaries especially for digit speech. It is difficult to accurately locate the start and end point of the speech segment in such an environment, but it is definitely necessary for robust speech recognition.

Thgrefore we propose a real-time speech detector based on useful bands in Mel frequency bands. The new algorithm is composed of two algorithms, the baseline energy-ZCR based method and ATF method. This paper provides some solutions in order to improve the robustness of continuous digit speech detection over noisy wire/wireless telephone networks.

In Section II, we describe two speech detection algorithms and analyse their characteristics, advantages, and disadvantages. In Section III, the proposed speech

Corresponding author: Mi-Kyong Ji (lindaji@icu.ac.kr)
School of Engineering, Information and Communications University
58-4, Hwaam-dong, Yuseong-gu, Daejeon, Korea

detection algorithm is presented. The experimental results and their comparison are provided in Section 4. Finally, in Section 5 we present conclusions.

II. Various Speech Detection Algorithms

2.1. Baseline Energy-zcr Based Speech Detection

The algorithm is based on signal energy, zero-crossing rate, a set of complex decision rules, and threshold settings. It continuously looks at the input samples and detects the start and end points of speech without a prior knowledge of the input signal[5]. And it is also capable of real-time processing[5]. The start point of the speech is classified into two categories: fricative-like speech and vowel-like speech. To reduce the confusion between a pause and the real speech end, and to deal with temporary noises, the history about the past a few frames is used. The detection of the real speech end is based on the number of frames classified into noise or noise category during past a few frames[5]. As long as the SNR is high enough, this algorithm is fast, accurate, and practical. However, this algorithm is not appropriate to get reliable speech boundaries in case of low SNR environment. It doesn't work well in such an environment. And it cannot make sure that they're the exact boundaries. It extracts islands of reliability. Therefore it always includes extra a few frames before the start point and after the end point not to miss the important part of the start and end of speech. It causes the recognizer to consume unnecessary time and to waste resource.

2.2. ATF (Adaptive Time and Frequency) Algorithm

In this section, we discuss the speech detection algorithm with ATF (Adaptive Time and Frequency) parameter which consists of Time parameter and Frequency parameter. This algorithm analyses the entire utterance before the speech detector starts speech boundary detection and selects useful bands which the frequency parameter is based on. After that, it actually starts the speech detection process. Therefore it is difficult to locate the speech boundary in real-time. Because it has to go through the entire utterance to select bands every utterance, it also causes heavy computation. However, it selects the useful bands every utterance in advance and use this selected useful bands depending on each utterance to get the frequency parameter. The band selection makes the speech detector work noise-robust in noisy telephone environment.

The time parameter is the logarithm of root-mean square (rms) energy of time-domain speech signal. It is smoothed and normalized. The procedure to calculate the time parameter is shown in equation (1)-(3):

$$x_{rms}(m) = log\sqrt{\frac{\sum_{n=0}^{L-1} x_t^2(m,n)}{L}} \tag{1}$$

$$\hat{x}_{rms}(m) = \frac{x_{rms}(m-1) + x_{rms}(m) + x_{rms}(m-1)}{3} \tag{2}$$

$$T(m) = \hat{x}_{rms}(m) - \frac{\sum_{m=0}^{S-1} \hat{x}_{rms}(m)}{S} \tag{3}$$

Let the given time-domain speech signal be $x_t(m, n)$. This represents the magnitude of the nth point of the mth frame. L is a window size and S is the number of initial frames considered as silence. The overall procedure to compute ATF parameter is shown in Figure 1.

As shown in Figure 1, Mel-band energies are calculated per frame. Each Mel-band energy is computed in equation [4]:

$$x(m,i) = \sum_{k=0}^{N-1} |x_{fre}(m,k)| f(i,k), 0 \le m \le M-1 \tag{4}$$

Let the magnitude of the kth point of the spectrum of the mth frame be $x_{fre}(m, k)$ and the magnitude of the kth point of the ith filter bank be $f(i, k)$. Mel-band energies only corresponding to pre-selected useful bands are chosen and each selected Mel-band energy is smoothed and normalized like equation (5)-(6):

$$\hat{x}(m,i) = \frac{x(m-1,i) + x(m,i) + x(m+1,i)}{3} \tag{5}$$

$$X(m,i) = \hat{x}(m,i) - \frac{\sum_{m=0}^{S-1} \hat{x}(m,i)}{S} \quad (6)$$

Finally the selected Mel-band energies are summed, and this is the Frequency parameter. Finally the time parameter and the frequency parameter are combined. the ATF parameter is shown in equation[4], and the overall procedure to get ATF parameter is shown in Figure 1.

$$ATF(m) = smoothing\,(F(m) + C^{*}\,T(m)) \quad (7)$$

The useful bands are computed the same way as the frequency parameter. Each band energy is summed and averaged with time. Consequently the useful bands are determined by the magnitude of that. Concerning decision rule to locate speech boundary, the same rule as the baseline energy-ZCR based speech detector is applied in this paper.

III. Speech Detection Based on Useful Bands

The new algorithm is based on the two speech detectors described previously in Section II. The new algorithm takes advantages of those two speech detectors and gets over the disadvantages of them. It simplifies ATF parameter, uses this as its parameter, and applies the decision rule similar to the baseline energy-ZCR based speech detector. It selects useful bands in advance while training. Hence the speech detection is done in real-time and also it makes use of the unique characteristics of continuous digit speech through the band selection. In addition, it don't have to select useful bands every utterance, not like the ATF algorithm. It takes less computation than the ATF method.

3.1. Selection of Useful Bands

The goal of the new algorithm is to exactly detect digit speech over telephone networks by using pre-selected useful bands through training previously so that the speech detection is done in real-time. In this section, we put emphasis on how to select useful bands. The procedure for useful band selection is shown in Figure 2.

As shown in Figure 2, we put one more procedure next to the normalization process because we assume that the ratio of each Mel-band energy represents the unique characteristics of digit speech. Therefore each smoothed

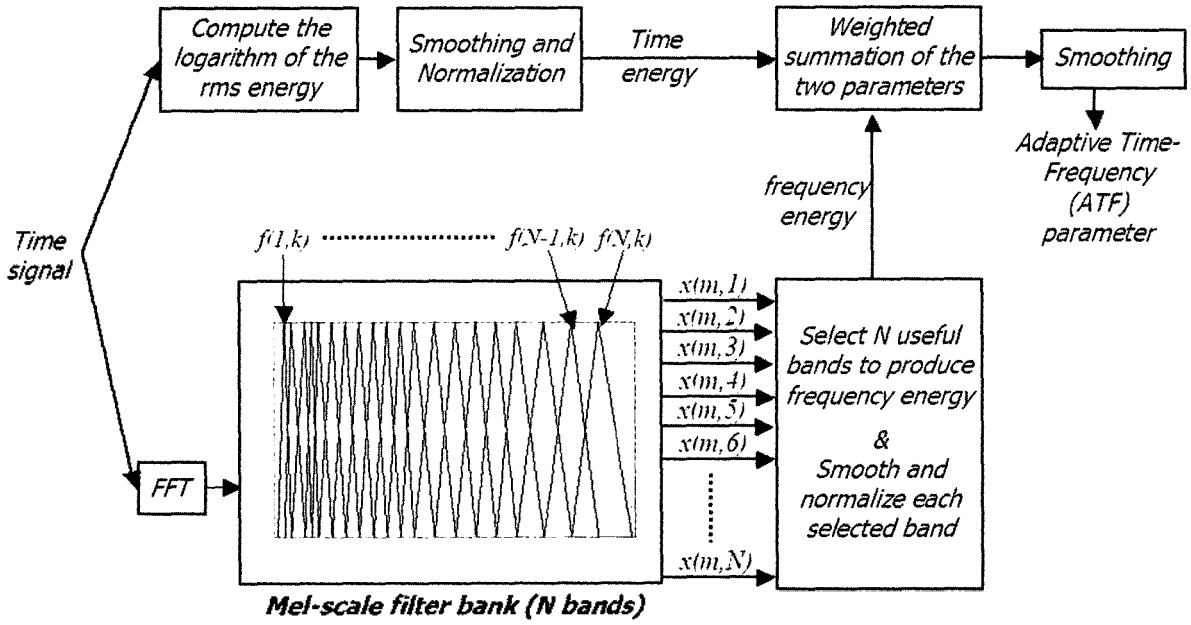


Figure 1. Procedure to compute ATF parameter.

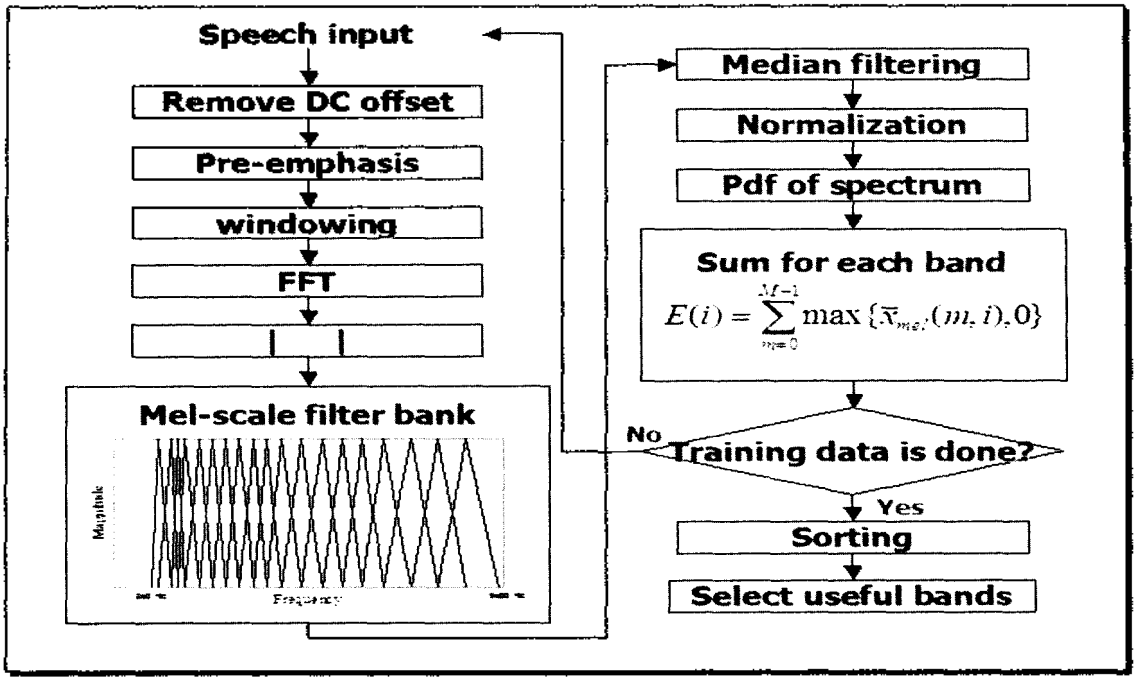


Figure 2. How to select useful bands.

and normalized Mel-band energy is divided by total sum of Mel-band energies. In consequence, the probability density function (pdf) for the spectrum was estimated by normalization over all Mel-band energies. The equation are follow as:

$$pdf\ of\ x_{mel}(m,i) = \frac{x_{mel}(m,i)}{\sum_{n=0}^{P} x_{mel}(m,i)} \tag{8}$$

Total energy of each Mel-band over all training data is calculated and sorted. Useful bands for continuous digit speech over telephone network are selected by the order of each band energy.

3.2. Basic Decision Rule for Speech Detection

The basic decision rule to detect the speech boundary is described in this section. First the new algorithm computes Mel-band energies from each frame, eliminates impulse noise by 3-point median filtering and gets rid of stationary noise estimated from the initial frames of each utterance. And it decides whether the current frame is speech or non-speech depending on thresholds. Here it compares the energies of the only bands corresponding to pre-selected useful bands. with their thresholds. If the number of bands that are greater than their thresholds is bigger than the pre-defined number, the current frame is considered as speech, otherwise it is considered as non-speech. Finally it decide the start of speech depending on the pattern of recent frames. The endpoint of speech is decided the same way.

3.3. Decision Rule 2 for Speech Detection

In case of vowel, we all know that it is usually distributed in low frequency area, and its energy is much bigger than that of the consonant, therefore it is less affected by noise than the consonant, so that we try to apply one more rule to the basic decision rule additionally. In this paper, 13 and 20 Mel frequency bands are applied between 0 and 4 kHz. Here we limit the frequency band between 0 and 2 kHz and apply the basic decision rule to this low frequency area again and we call it Decision rule 2. For 13 Mel-band structure, there exit about 9 bands between 0 and 2 kHz, and for 20 Mel-band structure, there are 13 bands.

IV. Experiment

4.1. Training

4 continuous digit utterances (ETRI DB) are used to select useful bands. The utterances are sampled with 8 kHz and they are quantized into 16 bit value. 16,000 utterances that are recorded from 70 speakers over telephone networks are used for training.

As shown in Figure 3, the normalized and averaged energies per Mel-band obtained by training is presented. And it shows that the band energies are concentrated on some Mel frequency bands.

4.2. Test DB

4 continuous digit utterances (ETRI DB) are used as baseline DB (Clean) for test. 76,000 utterances are recorded from 300 speakers over telephone networks. The average SNR is about 25 dB. We made 2 kinds of noisy DB (10 dB, 15 dB) by contaminating the baseline DB (Clean) adding AWGN and use them for test.

4.3. Evaluation

The proposed method is compared with the baseline energy-ZCR based method and the ATF method. The new algorithm is tested by changing the number of Mel frequency bands between 0 and 4 kHz. And also the performance is evaluated by whether the decision rule 2 is applied or not. In this paper, the numbers of Mel frequency band between 0 and 4 kHz are fixed as 13 and 20. The speech detection are tested for 3 kinds of DB (Clean, 10 db, and 15 dB). The performance of the new algorithm is compared with those of 2 speech detectors previously mentioned from the different point of view.

4.4. Results

The experimental results are shown in Figure 4 through Figure 11. UB13, UB13_L, and UB20_L are proposed methods. They are different depending on what is the number of Mel frequency bands or whether the decision rule 2 is applied or not. For UB13, the number of Mel frequency bands between 0 and 4 kHz is 13 and the decision rule 2 is not applied. For UB13_L, the number of Mel frequency bands are the same as the UB13 but the decision rule 2 is applied additionally. In case of UB20_L, the number of Mel frequency bands are 20 and the decision rule 2 is applied additionally. ATF stands for the experiment of ATF algorithm. Finally Energy-ZCR means the experiment of the baseline energy-ZCR based algorithm. The number inside the parentheses is the number of useful bands used.

As described in Figure 4, the distribution of UB13_L

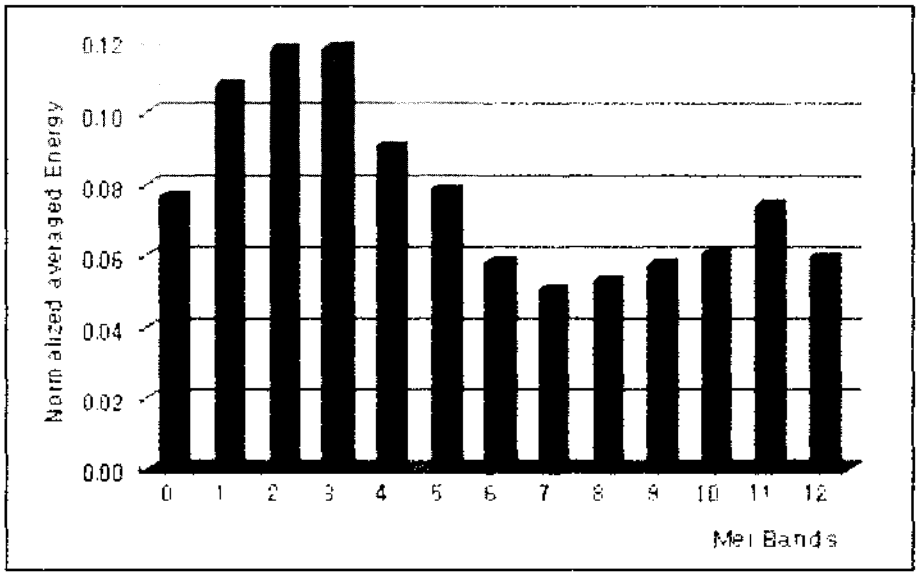


Figure 3. Normalized and averaged Mel-band energies from training DB.

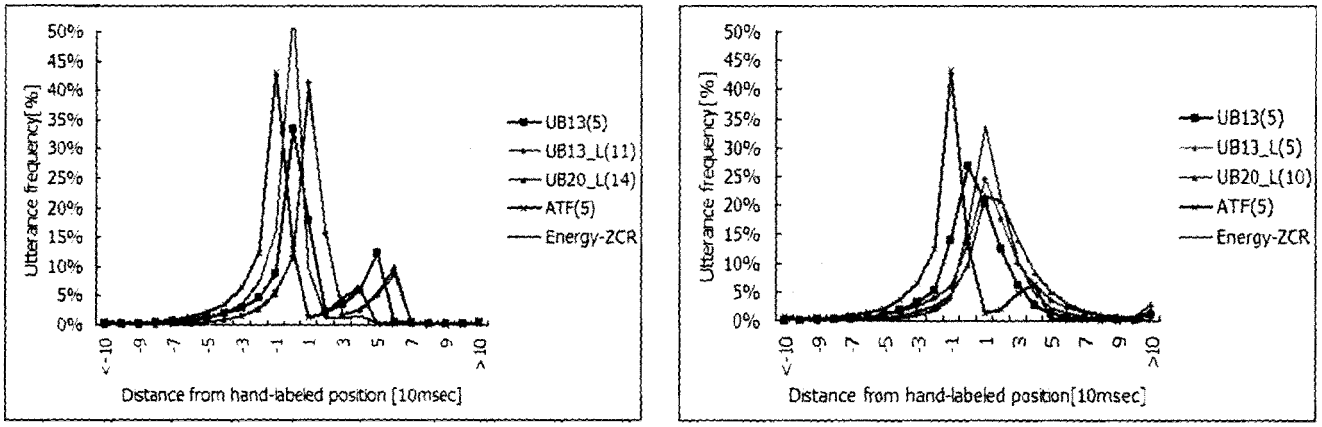
and UB20_L are concentrated around 10 *msec*, and that of Energy-ZCR is distributed more around 0 *msec* for the start point (The window size used in the experiment is 20 *msec* and the windows are overlapped). This means that the UB13_L and the UB20_L are more sensitive than the Energy-ZCR in detecting the speech. At least 10 *msec* cannot be avoidable. And the distributions are not like Gaussian. There is the other peak around 60 *msec*. This explains that the labeling information used in the experiment is not perfect. And actually there exist quite a lot of data that are not matched with the labeling information in DB.

Figure 5 represents the absolute distance of the detected speech from the hand-labeled position. For example, when the distance from hand-labeled position is 60 *msec*, the utterance frequency is about 98% for the start point. This means that the 98% of the test DB are located within 60 *msec* from the labeled position. The Energy-ZCR shows the best performance.

In Figure 6, we measure only the positive distance of the detected speech from the labeled position. That is, the only detected speech without cutting the speech part, is considered. Although the labeling information is not quite perfectly matched with real position, the UB13_L and UB20_L shows the best performances for the start point detection, but the Energy-ZCR is the best for the end point detection.

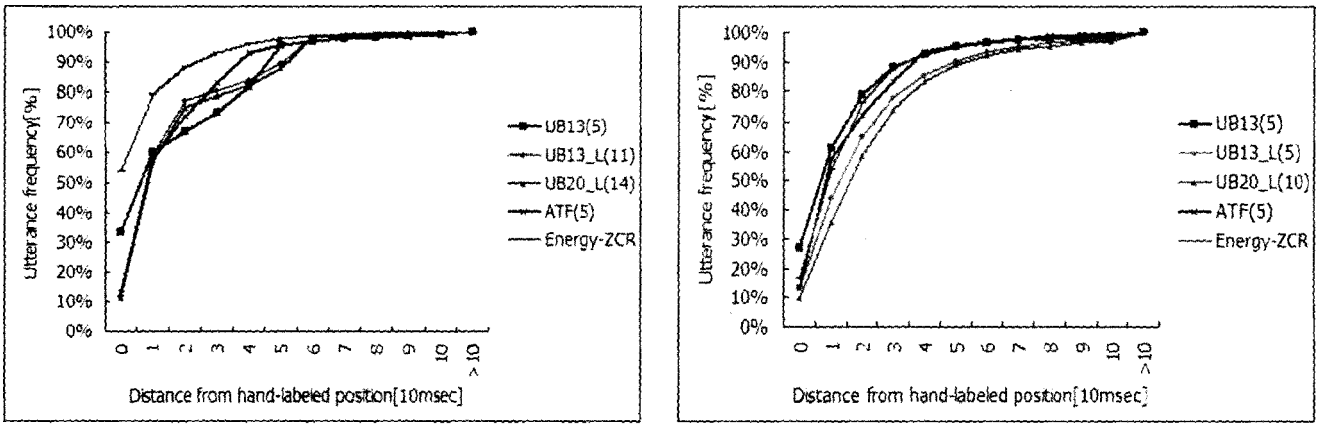
We tested 15 dB noisy DB with the same way as the baseline DB (Clean). The UB13_L and UB20_L shows the best performances for both start and end point detection.

As described in Figure 8, the performances of U13_L and UB20_L are better than the others not like the baseline



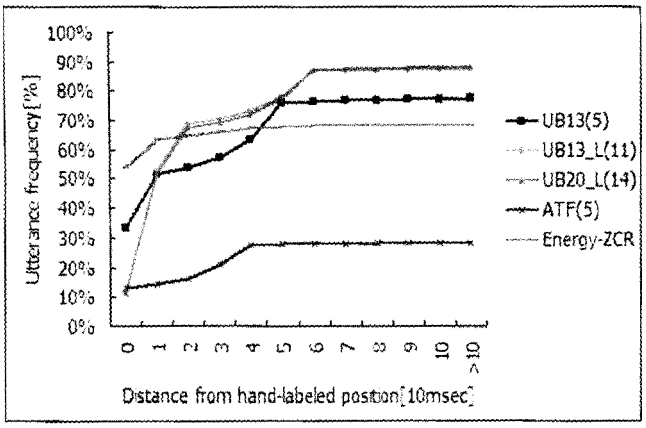
(a) For start point (b) For end point

Figure 4. Average distance distribution of the start and end point detection for the baseline clean DB (From the top). *Plus and minus sign represent additional inclusion of non-speech and omission of speech, respectively.

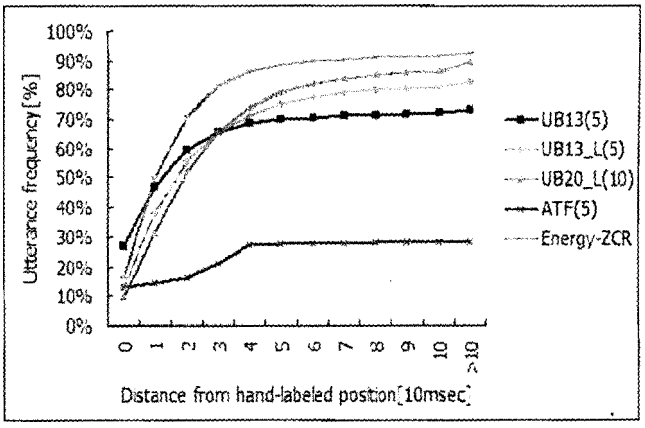


(a) For start point (b) For end point

Figure 5. Accuracy to detect the start and end point within the specified distance for the Baseline DB with SNR 25 dB (From the top).

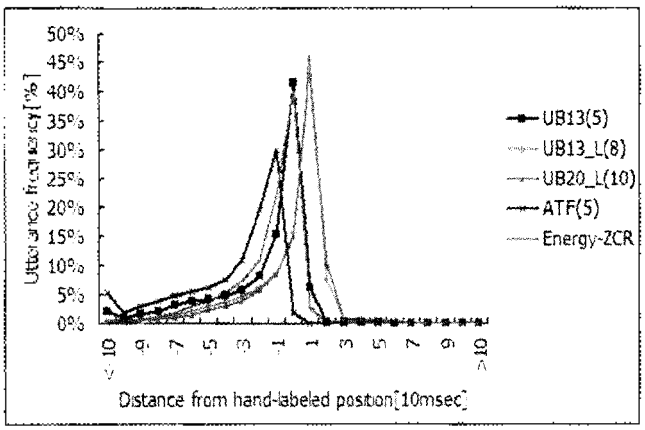


(a) For start point

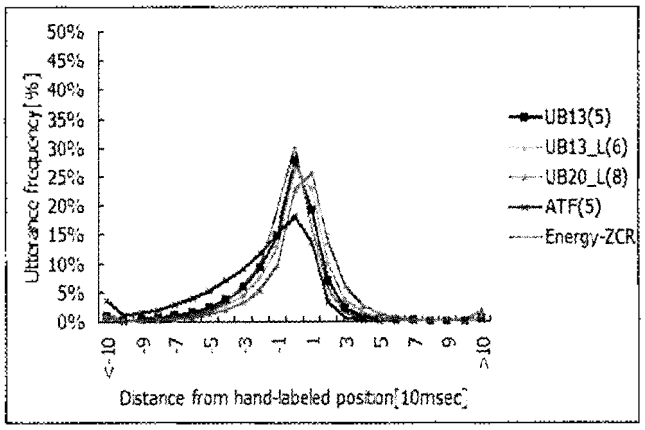


(b) For end point

Figure 6. Accuracy to detect the start and end point between 0 and the specified distance for Baseline DB with SNR 25 dB (From the top).

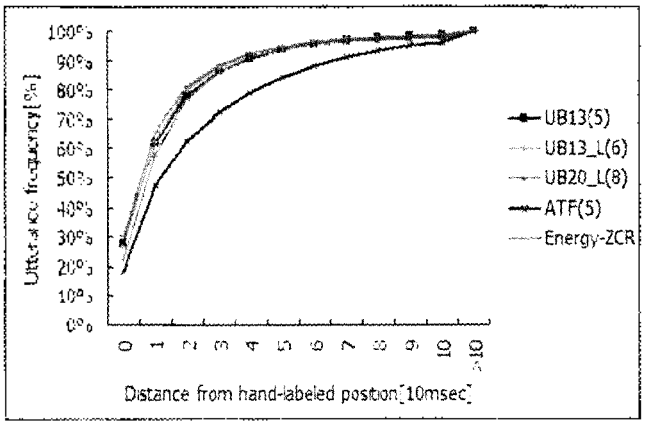


(a) For start point

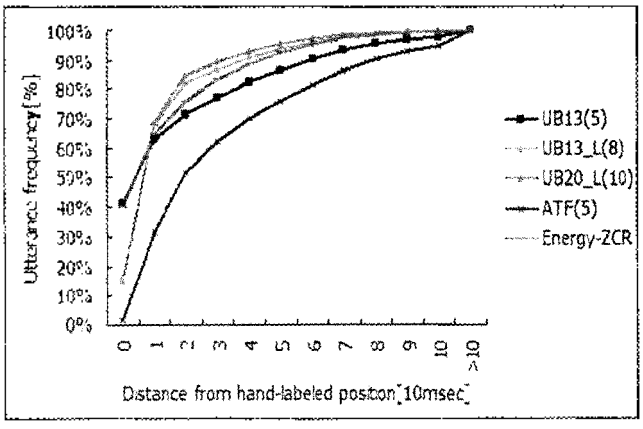


(b) For end point

Figure 7. Average distance distribution of the start and end point detection for the noisy DB with SNR 15 dB (From the top).



(a) For start point



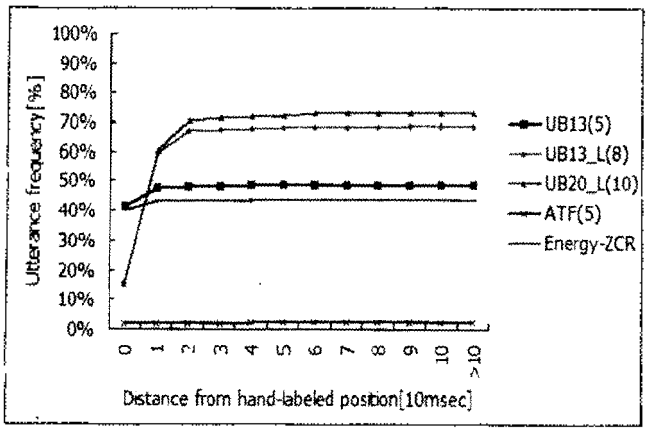
(b) For end point

Figure 8. Accuracy to detect the start and end point within the specified distance for the noisy DB with SNR 15 dB (From the top).

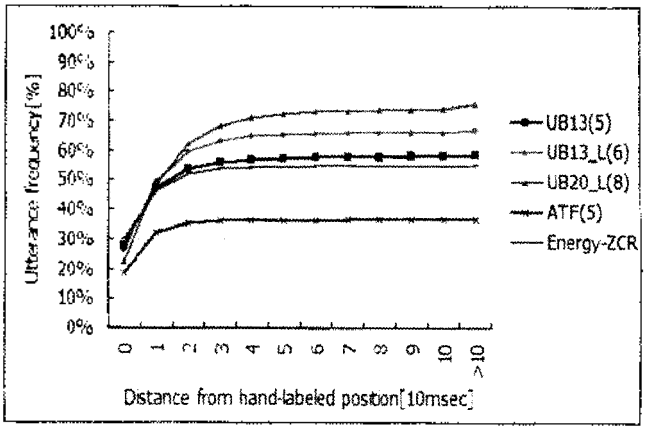
experiment.

For 10 dB noisy DB, the UBL13_L and UB20_L present the best performances to detect speech boundary. Both of them outperforms the others. This is shown in Figure 10 through 12.

The performances of the new algorithm by changing the number of useful bands is shown in Figure 12 and 13. As described in Figures, the performance of the new

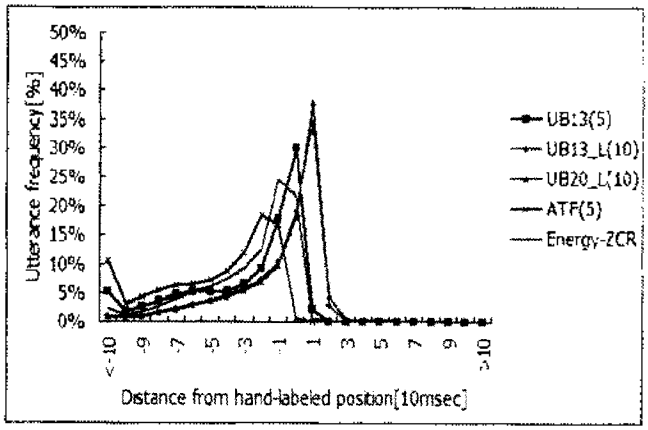


(a) For start point

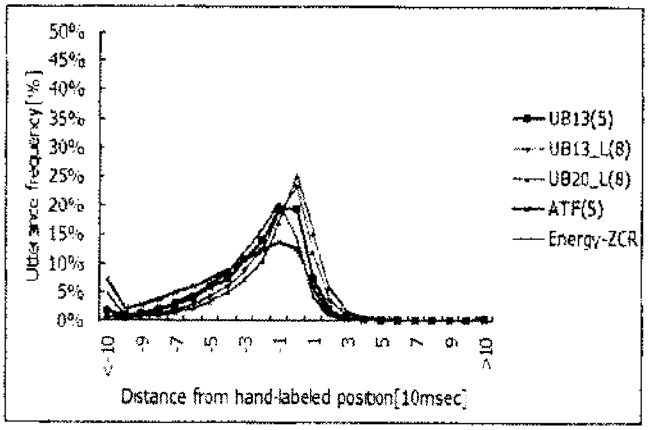


(b) For end point

Figure 9. Accuracy to detect the start and end point between 0 and the specified distance for the noisy DB with SNR 15 dB (From the top).

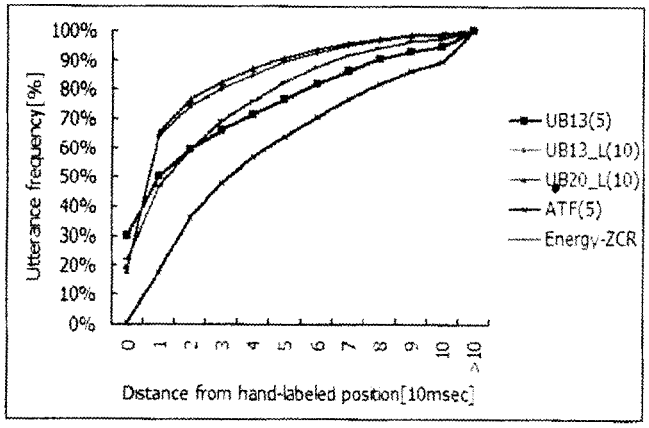


(a) For start point

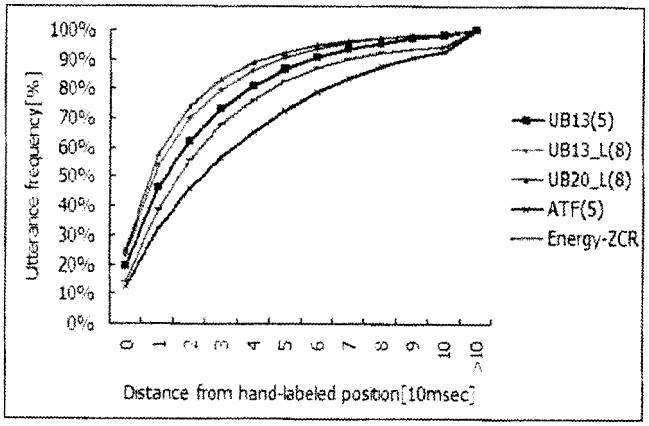


(b) For start point

Figure 10. Average distance distribution of the start and end point detection for the noisy DB with SNR 10 dB (From the top).



(a) For start point

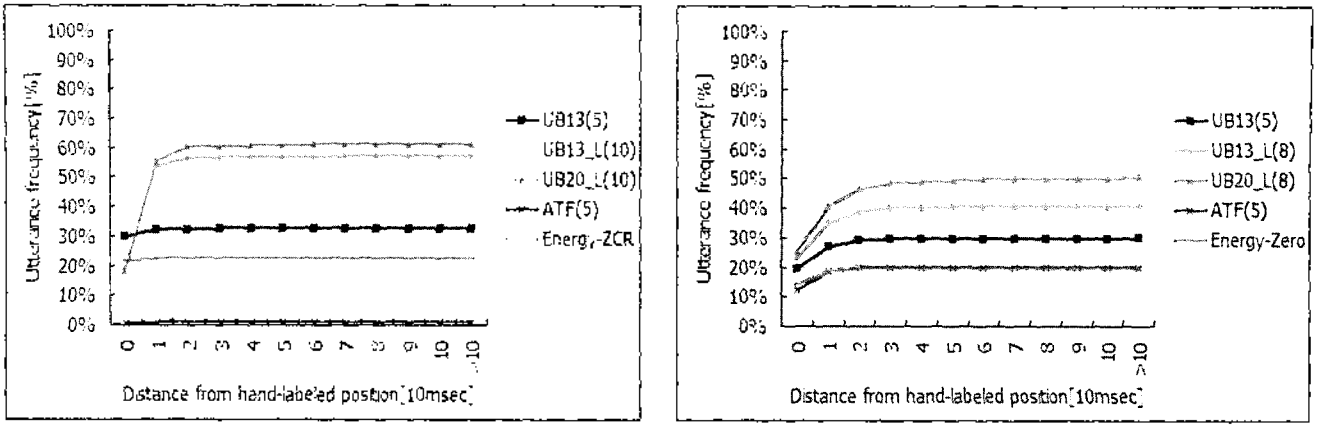


(b) For end point

Figure 11. Accuracy to detect the start and end point within the specified distance for the noisy DB with SNR 10 dB (From the top).

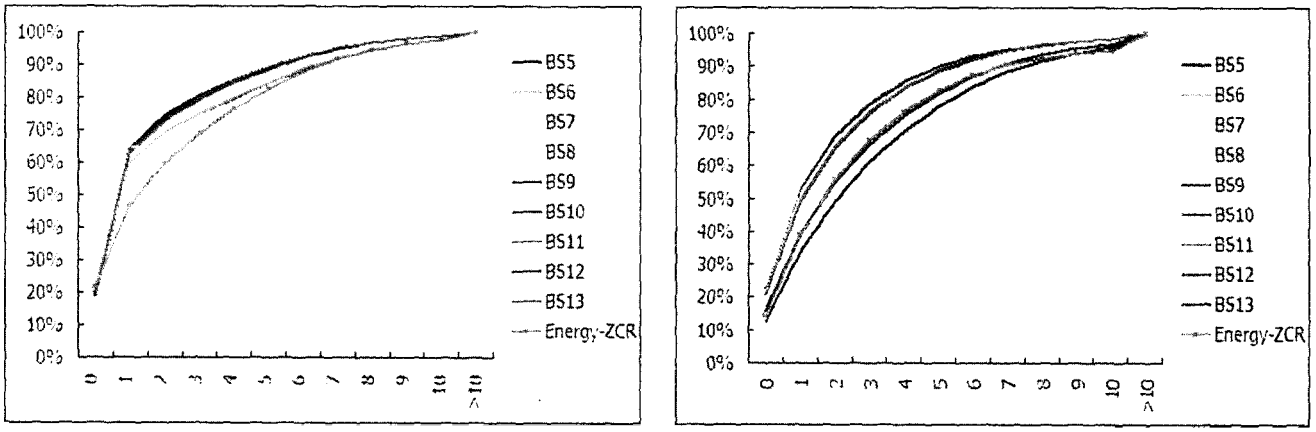
'gorithm is better than the baseline energy-ZCR based gorithm regardless of the number of useful bands for the rt point. For the end point, all experiments shows good 'ormance except when the number of useful bands is 13. (BS13 means that the number of selected useful bands is 13.)

Table 1 represents the number of utterances whose start and end points are detected only between -50 *msec* and



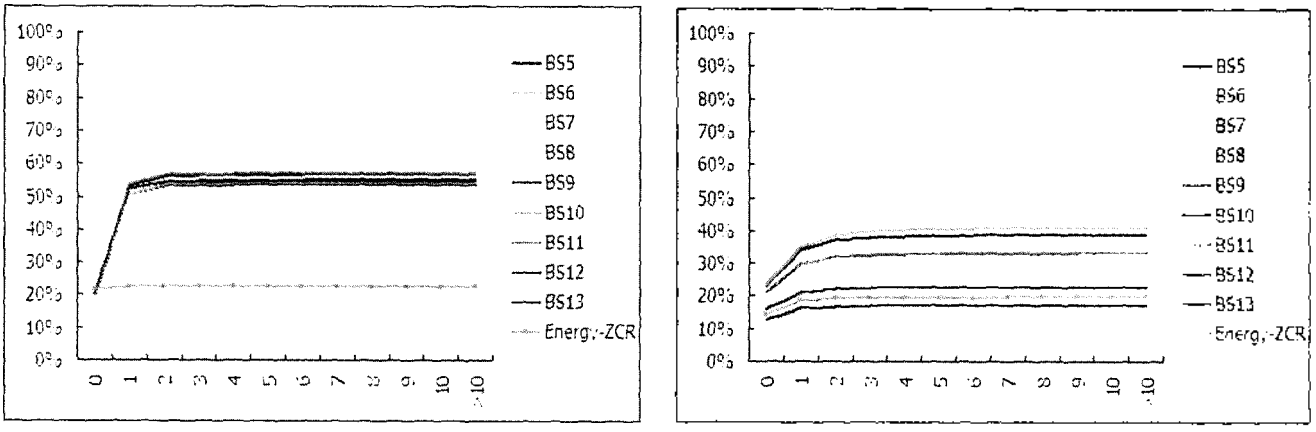
(a) For start point (b) For end point

Figure 12. Accuracy to detect the start and end point between 0 and the specified distance for the noisy DB with SNR 10 dB (From the top).



(a) For start point (b) For end point

Figure 13. Accuracy for the new algorithm to detect the start and end point (From the top) within the specified distance by changing the number of useful bands with the noisy DB with SNR 10 dB.



(a) For start point (b) For end point

Figure 14. Accuracy for the new algorithm to detect the start and end point (From the top) between 0 and the specified distance by changing the number of useful bands with the noisy DB with SNR 10 dB.

+50 msec from the labeled position in percentage. In this table, the performance of the baseline DB (25 dB) is not good. The reason is the mismatch of the labeling location and real location. Nonetheless, the performances of 15 dB and 10 dB shows that UB20_L and UB13_L are much better than the rest of them

Table 1. Performance comparison of the detected speech between -50 msec and +50 msec.

Method / Test DB	Energy-ZCR		ATF		UB13		UB13_L		UB20_L	
	Start	End	Start	End	Start	End	Start	End	Start	End
25 dB	***98.00***	95.38	95.50	***95.50***	95.48	95.09	89.41	90.29	87.87	88.81
15 dB	92.42	***94.86***	75.97	84.20	86.46	93.83	93.84	94.81	***95.49***	93.98
10 dB	82.34	82.76	63.87	72.72	76.64	86.79	89.22	90.90	***90.73***	***92.54***

Table 2. Performance comparison of the detected speech between 0 msec and +50 msec.

Method / Test DB	Energy-ZCR		ATF		UB13		UB13_L		UB20_L	
	Start	End	Start	End	Start	End	Start	End	Start	End
25 dB	67.95	***88.45***	27.64	27.64	75.66	69.93	77.93	75.30	***77.29***	79.20
15 dB	43.46	54.35	1.99	36.27	48.52	57.42	68.17	65.32	***72.46***	***72.24***
10 dB	22.48	19.55	0.73	19.83	32.52	29.55	56.90	40.46	***60.97***	***49.39***

Table 2 represents the number of utterances whose start and end points are detected only between 0 *msec* and +50 msec from the labeled position in percentage. In other word, it shows the frequency of utterances whose start and end points are located within 50 *msec* without cutting speech. The performance shows that UB20_L and UB13 outperforms the others. The performance of UB20_L is the best, and also that of UB13_L is comparable with it.

V. Discussion and Conclusion

In this paper, we proposed a robust speech detector for continuous digit speech over telephone networks, compared the new algorithm with 2 different speech detector and evaluated them. The new algorithm seems to show much better performance than the rest in noisy environment rather than in clean environment. It improves the robustness of the speech detector by making use of the unique characteristics of continuous digit speech, and increases the performance by applying the basic decision rule to the limited frequency area again (decision rule 2). It also makes it possible for the speech detector to locate speech boundary in real-time. And we compared the performances by changing the number of selected useful bands with 10 dB noisy DB. The performance was good except when the number of useful bands was 13. This can be avoided by reducing the number of selected bands in noisy environment. The new method was more reliable and robust for continuous digit speech over noisy telephone network. However, the new algorithm selects useful bands for continuous digit speech through training, and detects speech boundary. Therefore the algorithm is somewhat domain-specific. For the further work, we will focus on this problem later on.

Acknowledgments

This study was supported by ETRI in 2002.

References

1. L. R. Rabiner and M. R. Sambur, "An algorithm for determining the end-points of isolated utterances," *Bell Syst. Tech. J.*, **54**, 297-315, Feb. 1975.
2. M. H. Savoji, "A robust algorithm for accurate endpointing of speech," *Speech Commun.*, **8**, 45-60, 1989.
3. L. Lamel, L. Rabiner, A. Rosenberg, and J. Wilpon, "An improved endpoint detector for isolated word recognition," *IEEE ASSP Mag.*, **29**, 777-785, 1981.
4. B. Reaves, "Comments on an improved endpoint detector for isolated word recognition," *IEEE Trans., Signal Processing*, **39**, 526-527, Feb. 1991.
5. J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Trans. Speech Audio Processing*, **2**, 406-412, July 1994.
6. J. B. Allen, "Cochlear modeling," *IEEE Acoust., Speech, Signal Processing Mag.*, **2**, 3-29, 1985.
7. J. L. Shen, J. W. Hung, L. S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy

environments," *Proc. Int. Conf. on Spoken Lang. Processing*, 3, 1015, Sydney, 1998.
8. J. G. Wilpon and L. R. Rabiner, "Application of hidden Markov models to automatic speech endpoint detection," *Computer Speech and Language*, 2, 321-341, 1987.

[Profile]

• Mi-Kyong Ji

Mikyong Ji received the B. S. degree in information engineering from Hansung University, Seoul, Korea, in 2000. She received the M. S. degree in the school of engineering from ICU, Daejon, Korea, in 2002. Her research interests include speech detection, speech recognition, and keyword spotting.

• Young-Joo Suh

Youngjoo Suh received the B. S. and M. S. degrees in electronics engineering in 1991 and 1993 from Kyungpook National University, Taegu, Korea. From 1993 through 1998 he was a research member in Electronics and Telecommunications Research Institute (ETRI), Taejon, Korea. From 1999 to 2000, he was an invited professor at Yeongjin Junior College, Taegu, Korea. From 2000 to 2002, He worked as a senior member of technical group at the Corevoice, Inc., Taejon, Korea. Since 2002, he has been a Ph.D. candidate at Information and Communications University, Taejon, Korea. His current research interests include speech enhancement, robust endpoint detection and speech recognition.

• Hoi-Rin Kim

Hoirin Kim received the B. S. degree in Electronic engineering from Hanyang University, Seoul, Korea, in 1984. He received the M. S. and Ph. D. degree in Electrical engineering from KAIST in 1987 and 1992, respectively. From 1994 to 1995, he was a visiting researcher in ATR Interpreting Telecommunication Research Laboratories (ATR-ITL), Japan. From 1987 to 1999, he worked for Electronics and Telecommunications Research Institute (ETRI) and he was involved in key technology development for automatic interpreting telephone, speech recognition interface for realistic communication, spoken language translation system, speech input/output S/W for PC, and speech recognition S/W for low-cost multimedia communication terminal. Since 2000, he has been an assistant professor in Information and Communications University. His research interests include speech recognition, speaker recognition, audio classification, telecommunication network application of speech processing technology, and spoken language translation.

• Sang-Hun Kim

Sanghun Kim received the B. S. degree in Electrical Engineering from Yonsei University, Seoul, Korea in 1990 and the M. S. degree in Electrical and Electronic Engineering from KAIST, Daejon, Korea in 1992 and the Ph. D. degree in Dept of Electrical, Electronic, Information and Communication Engineering from Univ. of Tokyo, Japan in 2003. Since 1992, he has been with Research Department and Spoken Language Processing Section of ETRI, Daejon, Korea. Currently, he is a project leader in Speech Database Research Team, Speech and Language Information Research Center of ETRI. His interests include speech synthesis, speech recognition, and speech signal processing.