Wavelet Energy-Based Support Vector Machine for Noisy Word Boundary Detection With Speech Recognition Application

Chia-Feng Juang, Chun-Nan Cheng and Chiu-Chuan Tu

Department of Electrical Engineering
National Chung-Hsing University,
Taichung, 402 Taiwan, R.O.C.
e-mail: cfjuang@dragon.nchu.edu.tw

Abstract

Word boundary detection in variable noise-level environments by support vector machine (SVM) using Low-band Wavelet Energy (LWE) and Zero Crossing Rate (ZCR) features is proposed in this paper. The Wavelet Energy is derived based on Wavelet transformation; it can reduce the affection of noise in a speech signal. With the inclusion of ZCR, we can robustly and effectively detect word boundary from noise with only two features. For detector design, a Gaussian-kernel SVM is used. The proposed detection method is applied to detection word boundaries for an isolated word recognition system in variable noisy environments. Experiments with different types of noises and various signal-to-noise ratios are performed. The results show that using the LWE and ZCR parameters-based SVM, good performance is achieved. Comparison with another robust detection method has also verified the performance of the proposed method.

Keywords: Speech detection, word boundary detection, support vector machine, wavelet transform, noisy speech recognition.

1. INTRODUCTION

For speech recognition, the detection of speech affects recognition performance. A robust word boundary detection method in the presence of variable-label noises is necessary and is studied in this paper. Depending on the characteristics of speech, a variety of parameters have been proposed for boundary detection. They include the time energy (the magnitude in time domain), zero crossing rate (ZCR) [1] and pitch information [2]. These parameters usually fail to detect word boundary when signal-to-noise ratio (SNR) is low. Another parameter concerning frequency domain has also been recently proposed. According to the frequency energy, the time-frequency (TF) parameter [3] which sums the energy in time domain and the frequency energy was presented. The TF-based algorithm may work well for fixed-level background noise. However, its detection performance degrades for background noise of various levels. For this problem, some modified TF parameters are proposed [4]. In [5], the idea of using Wavelet transform features as speech detection features was proposed. In this paper, we present a new Low-band Wavelet Energy (LWE) parameter which separates the speech from noise in the domain of Wavelet transform. Computation of the WE parameter is easier than the modified TF parameters, and it is shown in the experiment section that a better detection performance is achieved.

After the features for detection have been extracted, the next step is to determine thresholds and decision rules. Many decision methods based on computational intelligence techniques have been proposed, such as fuzzy neural networks (FNNs) [4] and neural networks (NNs) [6]. Generalization performance may be poor when FNNs and NNs are over-trained. To cope with the low generalization ability problem, a new learning method, the Support Vector Machine (SVM), has been proposed [7, 8]. SVM is a new and useful learning method whose formulation is based on the principle of structural risk minimization. Instead of minimizing an objective function based on training, SVM attempts to minimize a bound on the generalization error. SVM has gained wide acceptance due to its high generalization abilities for a wide range of applications. For this reason, this paper used a SVM as a detector.

The rest of the paper is organized as follows. Section II introduces the derivation and analysis of the WE and ZCR parameters. Section III describes the SVM detector. Experiments on word boundary detection for noisy speech recognition are studied in Section IV. Finally, Section V draws conclusions.

2. ROBUST DETECTION PARAMETERS

Wavelet Transform (WT) is a technique for analyzing the time-frequency domain that is most suited for a non-stationary signal [9]. For short-time analysis and discrete speech signal, discrete-time WT (DTWT) is used. Let the amplitude of the k th point in the i th frame of a noisy speech signal be denoted by $s(i,k)$ and the frame length in sample number be represented by N. The DTWT of the i-th speech frame is as follows,

$$\mathrm{DTWT}(m,n)=\frac{1}{\sqrt{a_0^m}}\sum_{k=1}^{N}s(i,k)\psi(a_0^{-m}k-n\tau_0)\,, \tag{1}$$

where $\psi(\cdot)$ represents a wavelet basis function, a_0^m is the scale and τ_0 is a translation parameter which is set to a_0^{-m} in this paper. The commonly used value a_0=2 is used in this paper, resulting in a binary dilation. Thus, Eq. (1) can be written as

$$\mathrm{DTWT}(m,n)=\frac{1}{\sqrt{2^m}}\sum_{k=1}^{N}s(i,k)\psi[2^{-m}(k-n)] \tag{2}$$

.

In this paper, the Harr wavelet is used in Eq. (2), where

$$\psi[2^{-m}(k-n)]=\begin{cases}1\,, & 0\le 2^{-m}(k-n)\le\dfrac{1}{2}\\ -1\,, & \dfrac{1}{2}\le 2^{-m}(k-n)\le 1\\ 0\,, & \text{otherwise}\end{cases} \tag{3}$$

Generally, the DTWT is computed at scales a_0^m for, theoretically, all m. The output of DTWT can be regarded as finding the output of a bank of band-pass filters, where different values of scales corresponds to different band-pass filters. The outputs of DTWT at different

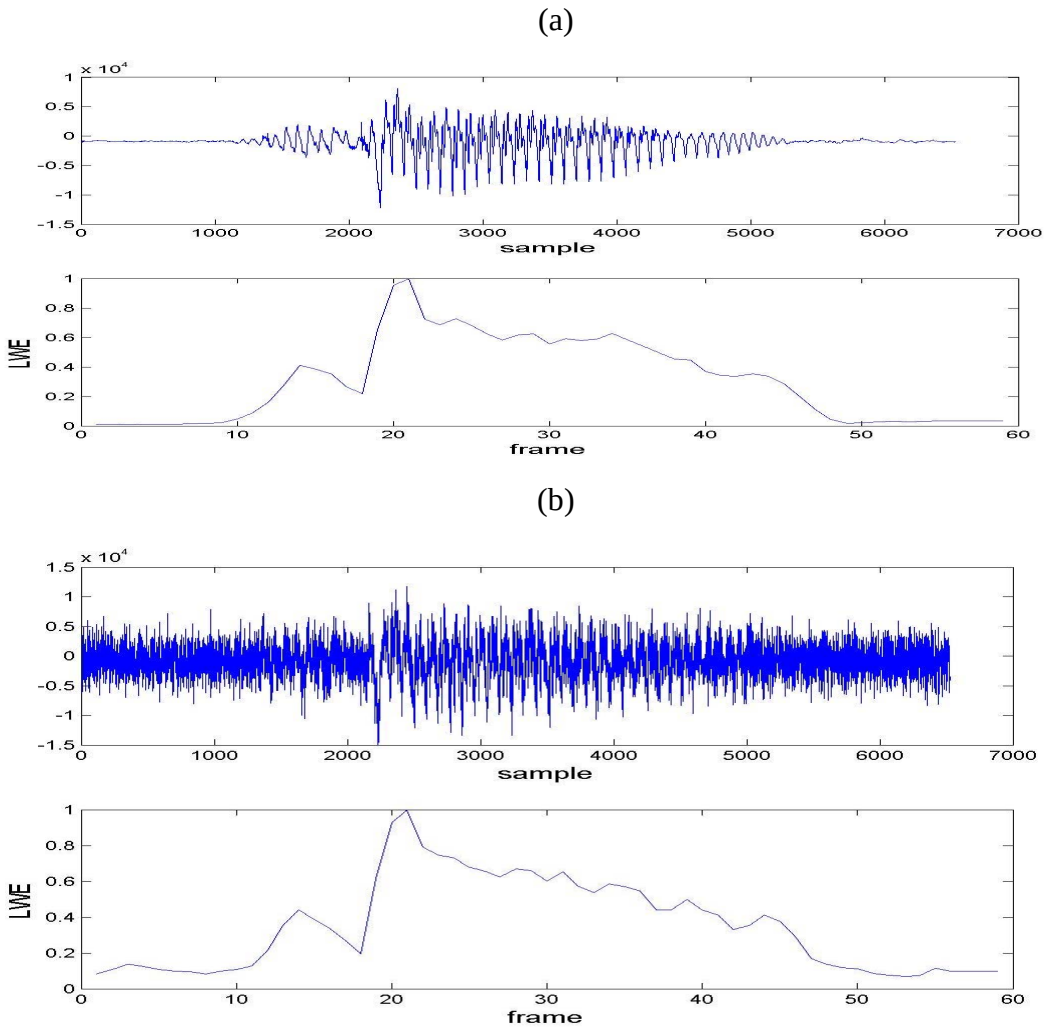

Fig. 1. (a) The LWEs of clean speech (b) The LWEs of speech with white noise added at SNR5.

scales contain different amounts of speech and noise information, and only the crucial scale(s) that contains maximum word signal information and is robust to noise should be used. Therefore, energy of the crucial scale is adopted as detection parameter for distinction between speech and noise in this paper.

To find the crucial scale, some observations on the effect of additive noise are made on different scales of DTWT. It is found that at the scale of $a_0^m = 2^6$, distribution of the STWT amplitudes matches well with the speech interval.

After computing DTWT for each time frame of a speech signal at the scale $a_0^m = 2^6$, the next step is to find an energy parameter to stand for the amount of word signal information at this scale. It is found the speech section corresponds to large DTWT amplitude values. Thus, summation of the amplitudes over n can be used as a parameter to stand for the amount of word signal information. It is also found that the amplitudes of noise tend to become larger when translation index n is larger than $0.8N$. Thus, summation is performed only from n=0 to $n = 0.8N$. This novel detection parameter, called low-band

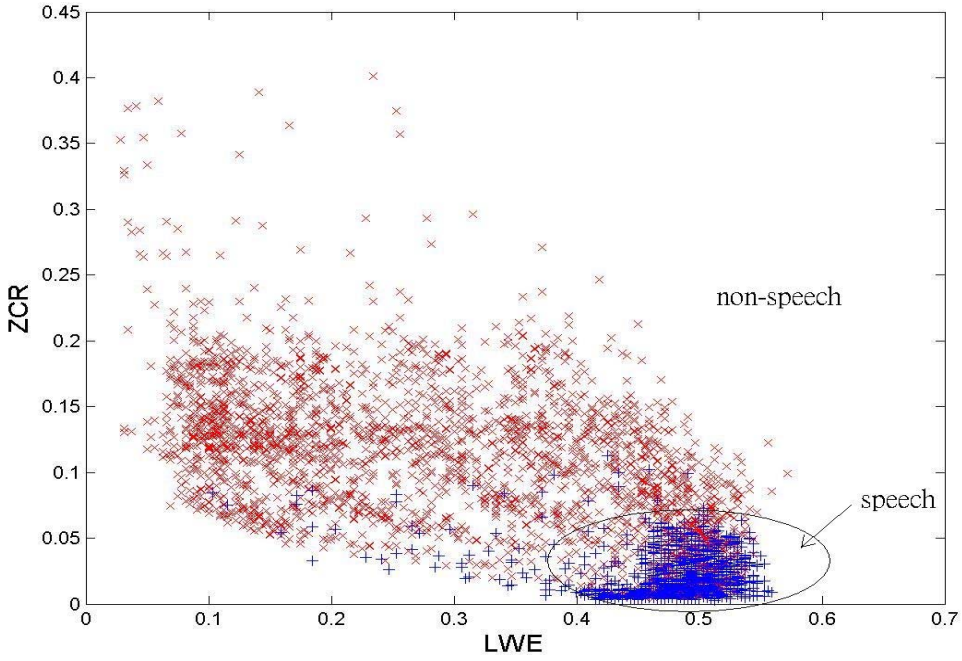

Fig. 2. Distributions of speech/non-speech frames in the LWE-ZCR plane with noise ranging from SNR20 to SNR0, where "×" and " +" denote non-speech and speech, respectively.

wavelet energy (LWE), is computed as follows,

$$\mathrm{LWE} = \sum_{n=0}^{0.8N} \left| \frac{1}{2^{-3}} \sum_{k=1}^{N} s(i,k)\psi(2^{-6}(k-n)) \right| \tag{4}$$

For illustration, a clean speech and its corresponding WE parameters of each frame are shown in Fig. 1(a). The speech with white noise and its corresponding WE parameters at SNR5 is shown in Fig. 1(b). This example shows that the WE parameter can robustly represent the energy of speech signal at different SNRs.

In addition to the WE parameter which is used to measure speech energy, the other parameter used for speech detection is the Zero Crossing Rate (ZCR). The reason for using the ZCR is that it is particularly suitable for un-voiced detection due to the high-frequency nature of the majority of fricatives.

Figure 2 shows distributions of speech/non-speech frames in the LWE-ZCR plane with noise levels SNR=20, 15, 10, and 5. The results show that the speech frames locate in a certain region of the two dimensional feature space.

3. SUPPORT VECTOR MACHINE DETECTOR

SVM is based on the statistical learning theory developed by Vapnik [7]. SVM first maps the input points into a high dimensional feature space and finds a separating hyperplane that maximizes the margin between two classes in this space. Suppose we are given a set S of labeled training set, $S = \{(\bar{x}_1, y_1), (\bar{x}_2, y_2), \cdots, (\bar{x}_N, y_N)\}$, where $\bar{x}_i \in \mathbb{R}^n$, and $y_i \in \{+1, -1\}$. Considering that the training data is linearly non-separable, the goal of SVM is to find an optimal hyperplane such that

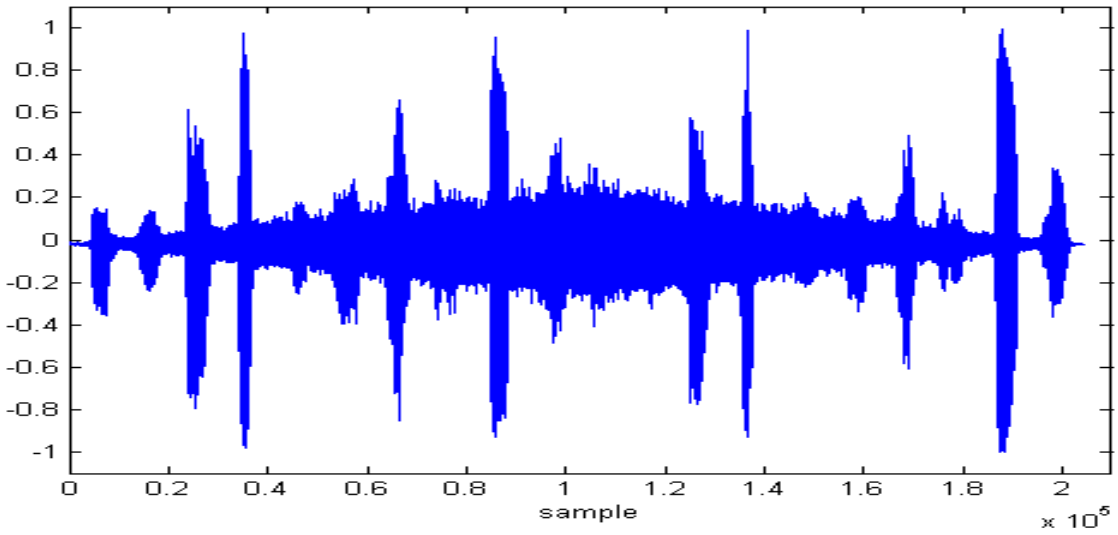

Fig. 3. The sequence of speech used for SVM training.

$$y_i\left(\vec{w}^T\vec{x}_i+b\right)\geq 1-\xi_i,\ i=1,\ldots,N \tag{5}$$

where $\vec{w}\in\mathbb{R}^n, b\in\mathbb{R}$, and $\xi_i\geq 0$ is a slack variable. For $\xi_i>1$, the data are misclassified. To find an optimal hyperplane is to solve the following constrained optimization problem:

$$\begin{aligned}&\text{Min}_{w,\xi}\quad \frac{1}{2}\vec{w}^T\vec{w}+C\sum_{i=1}^{N}\xi_i\\&\text{Subject to}\quad y_i(\vec{w}^T\vec{x}_i+b)\geq 1-\xi_i\end{aligned} \tag{6}$$

where C is a user defined positive cost parameter and $\sum\xi_i$ is an upper bound on the number of training errors. After solving Eq. (2), the final hyperplane decision function is achieved, and

$$f(\vec{x})=\text{sign}(\vec{w}^T\vec{x}+b)=\text{sign}(\sum_{i=1}^{N}y_i\alpha_i<\vec{x},\vec{x}_i>+b)=\text{sign}(\sum_{i\in SV}y_i\alpha_i<\vec{x},\vec{x}_i>+b) \tag{7}$$

where α_i is a Lagrange multiplier and the training samples for which $\alpha_i\neq 0$ are support vectors (SVs). A detailed derivation process can be found in [8].

The above linear SVM can be readily extended to a nonlinear classifier by first using a nonlinear operator Φ to map the input data into a higher dimensional feature space. In this way, it can solve nonlinear problems. By replacing $\vec{x}$ in Eqs. (1) and (2) with the feature space $\Phi(\vec{x})$ and solving the constrained optimization problem, the decision function

$$\begin{aligned}f(\vec{x})&=\text{sign}(\sum_{i=1}^{N}y_i\alpha_i<\Phi(\vec{x}),\Phi(\vec{x}_i)>+b)\\&=\text{sign}(\sum_{i=1}^{N}y_i\alpha_iK(\vec{x},\vec{x}_i)+b)\\&=\text{sign}(\sum_{i\in SV}y_i\alpha_iK(\vec{x},\vec{x}_i)+b)\end{aligned} \tag{8}$$

is achieved, where $K(\vec{x},\vec{x}_j)=\Phi(\vec{x})\cdot\Phi(\vec{x}_j)$ is called a kernel function. This paper uses a Gaussian-kernel SVM with $K(\vec{x},\vec{x}_j)=\exp(-\left\|\vec{x}-\vec{x}_j\right\|^2/\gamma)$, where γ is the width of a Gaussian-kernel. The two-dimensional inputs of the Gaussian-kernel SVM detector are ZER and LWE. For SVM, there is only one output and the desired output is "1" and "-1" if the input frame is speech and non-speech, respectively. During test, the SVM output indicates where or not the input frame is speech.

4. EXPERIMENTS

The wave files of speech are recorded by 11.025 kHz sample rate, mono channel and 16-bit resolution. For SVM training, the training sequence length is 13 seconds and is shown in Fig. 3. It consists of 20 words and is corrupted by white noise whose energy level increases from the start to SNR=0 and then decreases till the end of the sequence. For testing, the speech database is built of sequences of transcriptions from the same male speaker, where each sequence consists of ten isolated Mandarin words "0", "1",⋯, "9". There are a total of 50 test sequences used for playing the judicial role in performance comparison. The noise added to the speech sequence is of variable noise level during the sequence. Figure 4(a) shows the flowchart of training by

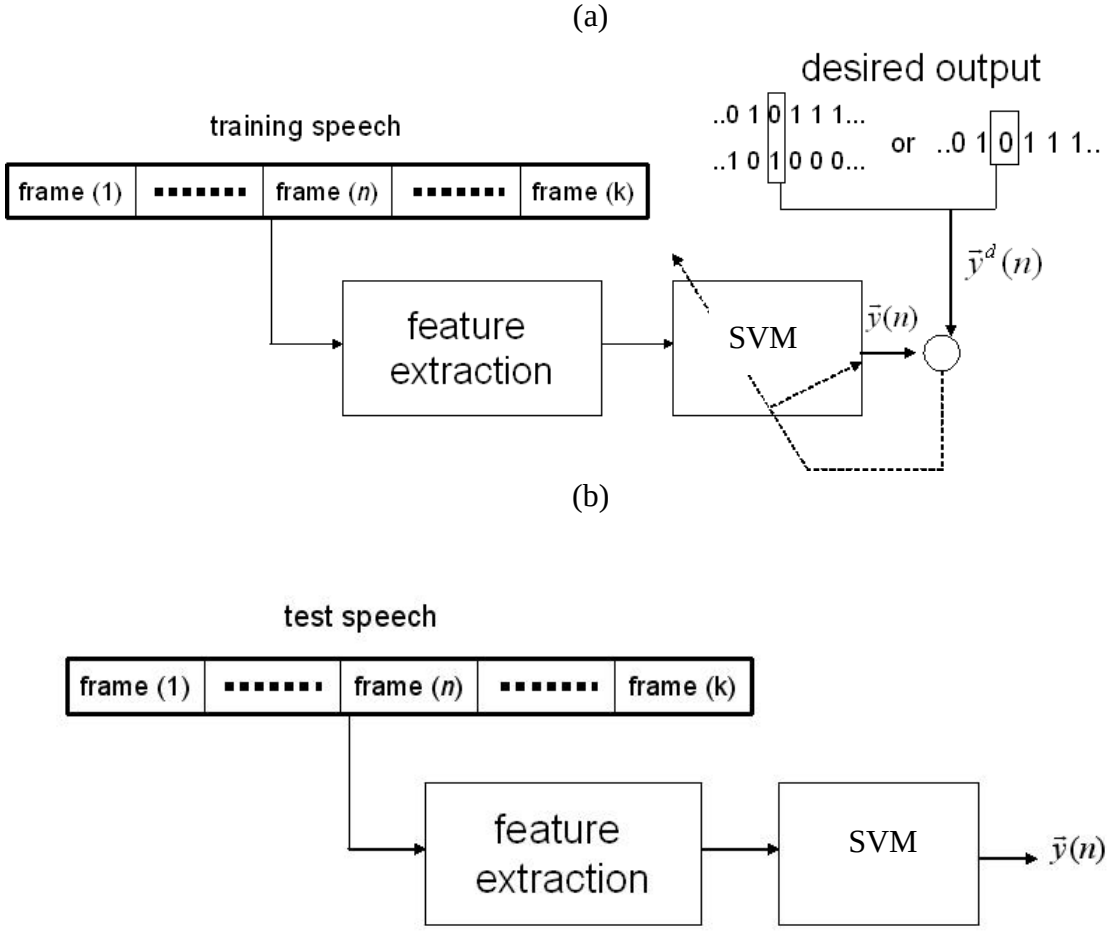

Fig. 4. (a) Flowchart of SVM training. (b) Flowchart of LWE-based SVM for test data.

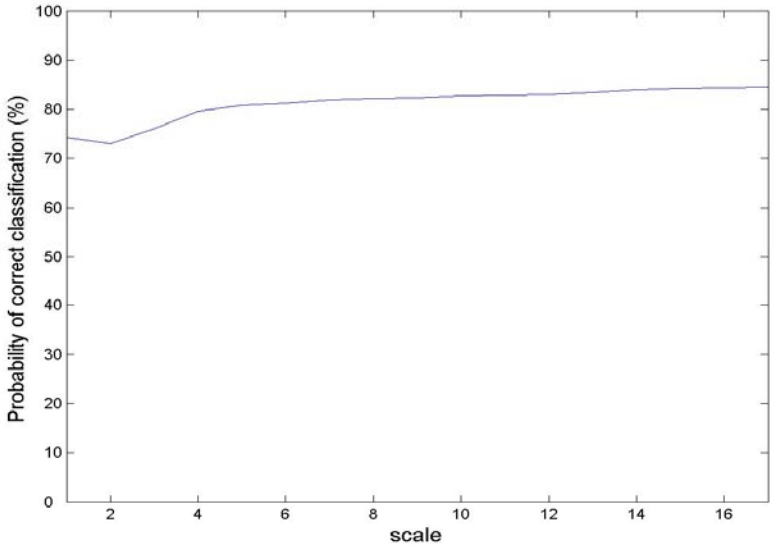

Fig. 5. Training performance of C in the range in the range [1, 85], where the range is spaced to 17 equal scales.

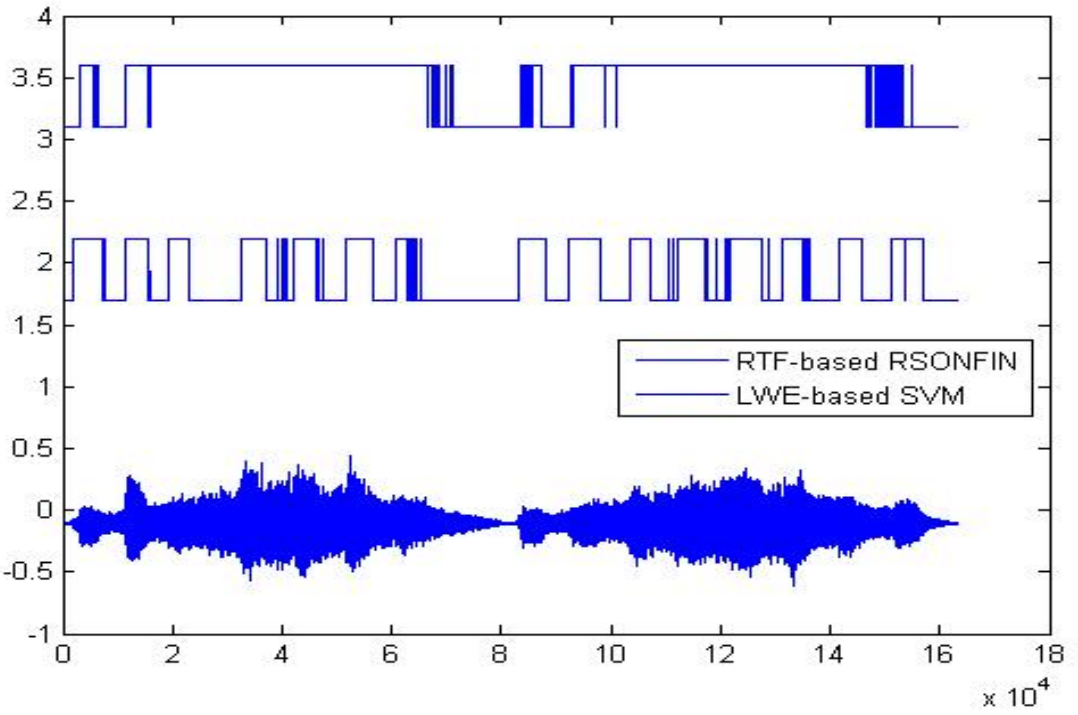

Fig. 6. Word boundary results by LWE-based SVM and RTF-based RSONFIN in variable noise level environment.

LWE-based SVM is shown, and Fig.4 (b) shows test of LWE-based SVM.

The classification rate defined in Eq. (9) is used as training performance index.

$$\text{Classification rate} = \frac{\text{Correctly detected frame number}}{\text{total frame number in training sequance}} \tag{9}$$

For SVM, the value of C influences the training performance. Fig. 5 shows the training performance of C in the range [1, 85], where the range is spaced to 17 equal scales. The cost value C is set to 40 in the following experiments, where there are a total of 1050

(a) (b)

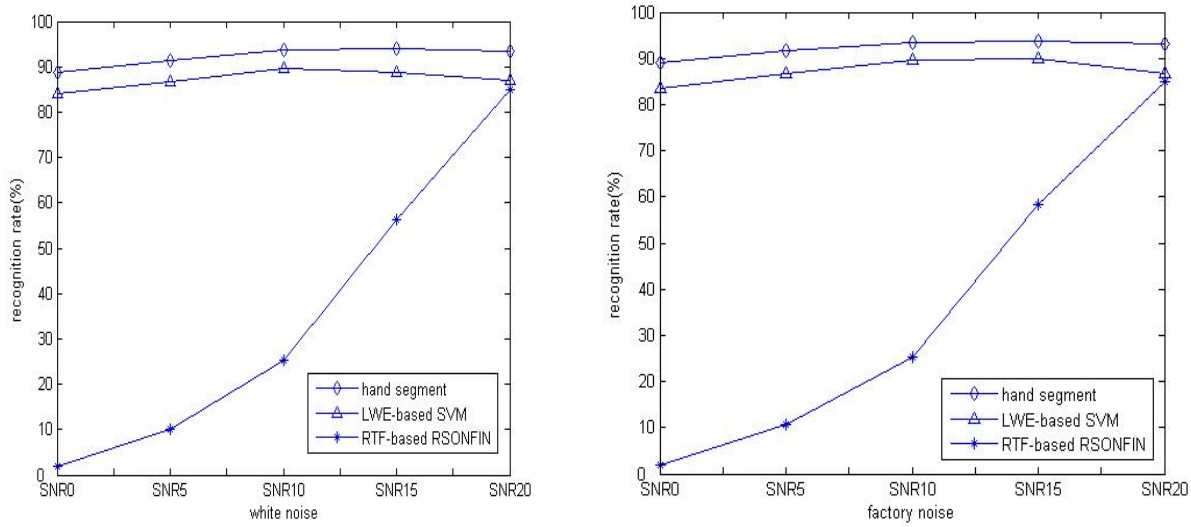

Fig. 7. Noisy speech recognition results by different word boundary detection methods. (a) white noise (b) factory noise.

SVs in the trained SVM.

To get a quick view on the test performance, some illustrative examples are experimented and shown in Fig. 6, where white noise with sharp variation in amplitude is added to the clean speech. Most word boundaries are correctly detected. For comparison, Fig. 6 also shows the performance of refined time-frequency (RTF) feature –based recurrent self-organizing neural fuzzy inference network (RTF-based RSONFIN) [4] detection method. The result shows that RTF-based RSONFIN almost fails to detect most of the words, and the performance of LWE-based SVM shows much better performance than RTF-based RSONFIN.

Next, the ten Mandarin digital words in each sequence of transcriptions in the test database are to be recognized. The words in each sequence are detected by the two methods respectively. When the number of successive frames being detected as speech is larger than 0.1 second, we regard it as word for recognition, otherwise these frames are discarded. So the number of words

detected in each sequence of transcription may be larger or smaller than exact ten words. Considering this phenomenon, we define the following recognition rate

$$\text{recognition rate} = \frac{\mathrm{T}-\mathrm{E}-\mathrm{U}-\mathrm{S}}{\mathrm{T}} \times 100\%, \tag{10}$$

where T is the total number of words in the reference transcriptions, E is the number of words recognized incorrectly, U is un-detect words of reference transcriptions, and S is surplus words of reference transcriptions.

For the recognizer, the hierarchical singleton-type recurrent neural fuzzy network (HSRNFN) [9] that put SNR20 white noise as training data is used. The reason we use HRNFN is that it achieves high recognition rate and is robust to different types of noise under different SNR. With HSRNFN recognizer, the recognition results by hand-segment,

LWE-based SVM, and RTF-based RSONFIN methods under white and factory noise are shown in Fig. 7. The results show that recognition rate of the LWE-based SVM method is slightly lower than that of hand segmentation, but is much larger than that of the RTF-RSONFIN method.

5. CONCLUSIONS

Two research results on robust speech detection in variable noise-level environment have been presented this paper, one is the robust LWE-based parameters, and the other is detector design by SVM. Variable noise-level instead of fixed noise-level is added to each sequence of transcript. Distributions of the LWE-based parameters in the 2-dimensional feature space for different SNRs have shown that the LWE-based parameters are feasible for speech detection over variable level noise. The LWE-based SVM can be applied to a speech recognition system as demonstrated in the experiments.

REFERENCES

[1] M. H. Savoji, A robust algorithm for accurate end-pointing of speech signals, *Speech Communication*, vol. 8. no. 1, 1989, pp. 45-60.

[2] J. Rouat, Y. C. Liu, and D. Morissette, Pitch determination and voiced/unvoiced decision algorithm for noisy speech, *Speech Communication*, vol. 21, no. 3, 1997, pp. 191-207.

[3] J. C. Junqua, B. Mak, and B. Reaves, A robust algorithm for word boundary detection in the presence of noise, *IEEE Trans. Speech and Audio Processing*, vol. 2, 1994, pp. 406-412.

[4] G. D. Wu and C. T. Lin A recurrent neural fuzzy network for word boundary detection in variable noise-level environments, *IEEE Transactions on systems, Man, and cybernetics*, vol. 31, no. 1, 2001, pp. 84-97.

[5] J. F. Wang and S. H. Chen, "A C/V segmentation algorithm for mandarin speech signal based on Wavelet transforms," *Proc. of ICASSP*, vol.1, pp.417-420, March 1999.

[6] Y. Qi and B. R. Hunt, Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier, *IEEE Trans. Speech and Audio Processing*, vol. 1, 1993, pp. 250-255.

[7] C. Cortes, and V. Vapnik, "Support vector networks," *International Journal on Machine Learning*, vol. 20, pp. 1-25, 1995.

[8] N. Cristianini and J. S.-Taylor, *An Introduction to Support Vector Machines And Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[9] Y. T. Chan, Wavelet Basics, 1995, Kluwer Academic Publishers.

[10] C. F. Juang, C. T. Chiou, and C. L. Lai, "Hierarchical singleton-type recurrent neural fuzzy networks for noisy speech recognition," *IEEE Trans. Neural Networks*, vol. 18, no. 3, pp. 833-843, May 2007.