Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 321–332
Expert Systems with Applications
www.elsevier.com/locate/eswa

Speech detection in noisy environments by wavelet energy-based recurrent neural fuzzy network

Chia-Feng Juang *, Chun-Nan Cheng, Tai-Mao Chen

Department of Electrical Engineering, National Chung-Hsing University, Taichung, 402 Taiwan, Taiwan, ROC

Abstract

This paper proposes a new speech detection method by recurrent neural fuzzy network in variable noise-level environments. The detection method uses wavelet energy (WE) and zero crossing rate (ZCR) as detection parameters. The WE is a new and robust parameter, and is derived using wavelet transformation. It can reduce the influences of different types of noise at different levels. With the inclusion of ZCR, we can robustly and effectively detect speech from noise with only two parameters. For detector design, a singleton-type recurrent fuzzy neural network (SRNFN) is proposed. The SRNFN is constructed by recurrent fuzzy if-then rules with fuzzy singletons in the consequences, and the recurrent property makes them suitable for processing speech patterns with temporal characteristics. The learning ability of SRNFN helps avoid the need of empirically determining a threshold in normal detection algorithms. Experiments with different types of noises and various signal-to noise ratios (SNRs) are performed. The results show that using the WE and ZCR parameters-based SRNFN, a pretty good performance is achieved. Comparisons with another robust detection method, the refined time–frequency-based method, and other detectors have also verified the performance of the proposed method.

Keywords: Fuzzy neural networks; Time–frequency parameter; Wavelet transform; Recurrent fuzzy rules

1. Introduction

The purpose of speech detection is to find the endpoints of speech so that it is distinguished from noise. The detection of speech is the first stage of speech processing, and the performance affects deeply the follow-up process, such as multimedia communication and speech recognition. In digital communication system, the bit rate used for silence is considerably lower than that used for active speech coding. The speech detection operation helps detect the pauses in a typical telephone conversation and reduce the bit rate. For speech recognition, the recognized speech should be detected beforehand. In practical environments, speeches are usually corrupted by noises due to interferences from telecommunication circuits, motor engines, and so on. Detection in the presence of variable-level noise is more changeling than in the presence of impulse noise or fixed-level noise. A robust speech detection method in the presence of different types of noises of various levels is necessary and is studied in this paper.

Depending on the characteristics of speech, a variety of parameters have been proposed for speech detection. They include the time energy (the magnitude in time domain), zero crossing rate (ZCR) (Lamel, Rabiner, Rosenberg, & Wilson, 1981; Savoji, 1989), cepstral coefficient (Haigh & Mason, 1993) and pitch information (Rouat, Liu, & Morissette, 1997). These parameters usually fail to detect speech 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 (Junqua, Mak, & Reaves, 1994) 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.

* Corresponding author. Fax: +886 4 22851410.
E-mail address: cfjuang@dragon.nchu.edu.tw (C.-F. Juang).

0957-4174/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2007.10.028

However, its detection performance degrades for background noise of various levels. For this problem, some modified TF parameters are proposed (Wu & Lin, 2000, 2001; Wu & Wang, 2005). In this paper, we present a wavelet energy (WE) 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 systems (Britelli, Casale, & Cavallaro, 1989) neural networks (Qi & Hunt, 1993), fuzzy neural networks (Juang & Lin, 1998; Wu & Lin, 2000, 2001), and support vector machines (SVMs) (Enqing, Guizhong, Yatong, & Xiaodi, 2002; Xianbo & Guangshu, 2005). Among these studies, a promising method for speech detection in variable noise-level environment is the refined TF-based recurrent self-organizing neural fuzzy inference network (RSONFIN) method (Wu & Lin, 2001). In fact, the refined TF-based RSONFIN method has been shown to have better performance than TF-based algorithm (Junqua et al., 1994) and adaptive TF-based neural fuzzy method (Wu & Lin, 2000). In this paper, a singleton-type recurrent neural fuzzy network (SRNFN) is proposed as a detector. The SRNFN modifies and simplifies the Takagi–Sugeno–Kang (TSK) type consequent of a TSK-type recurrent fuzzy network (TRFN) (Juang, 2002) to fuzzy singletons. This modification makes the SRNFN more robust to noisy speech signals than TRFN. This robustness is verified from simulation results in Section 4.

WE-based SRNFN for speech detection in variable noise-level environment is proposed in this paper. The main advantage of the proposed WE parameter is its ability to separate speech from noise in the wavelet-transformed domain. The WE together with ZCR are used as detection parameters, which are robust to noise at different levels. These two parameters are fed as inputs to a SRNFN detector to automatically learn the decision rules. Due to the recurrent structure of SRNFN, only the two detection parameters in current frame are fed to network inputs. No adjacent frame parameters are used, which reduces detector input dimension. Experiments are performed with different types of noises in variable noise-level environment.

The rest of the paper is organized as follows. Section 2 introduces the derivation and analysis of the WE and ZCR parameters. The SRNFN detector is introduced in Section 3. Experiments on speech detection are studied in Section 4. Finally, conclusions are drawn in the last section.

2. Robust detection parameters

The WE together with zero crossing rate (ZCR) are used as detection parameters. In Subsection 2.1, basic concept of wavelet transform used in WE derivation is described. Details of WE derivation are introduced in Subsection 2.2.

2.1. Basic concept of wavelet transform

Wavelet transform (WT) is a technique for analyzing the time–frequency domain that is most suited for a non-stationary signal (Chan, 1995). The WT uses the localized basis function to capture the localized features of a signal. Hence, it provides better signal approximation than other tools such as Fourier, sine or cosine transform. A continuous WT (CWT) maps a signal function in time domain into a two-dimensional function of a and τ. The parameter a is called the scale and corresponds to frequency in Fourier transform. It scales a function by compressing or stretching it, and τ is the translation of the wavelet function along the time axis. The CWT is defined by

$$\mathrm{CWT}(a,\tau)=\frac{1}{\sqrt{a}}\int s(t)\psi\left(\frac{t-\tau}{a}\right)\mathrm{d}t \tag{1}$$

where $s(t)$ is the signal, and $\psi(t)$ is the basic (or mother) wavelet and $\psi[(t-\tau)/a]/\sqrt{a}$ is the wavelet basis function, sometimes called baby wavelets.

For short-time analysis and discrete speech signal, instead of CWT, discrete-time WT (DTWT) should be used. Let the amplitude of the kth point in the ith 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:

$$\mathrm{DTWT}(m,n)=\frac{1}{\sqrt{a_0^m}}\sum_{k=1}^{N}s(i,k)\psi(a_0^{-m}k-n\tau_0), \tag{2}$$

where a_0^m is the scale and the translation parameter τ_0 is set to be 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. (2) can be written as

$$\mathrm{DTWT}(m,n)=\frac{1}{\sqrt{2^m}}\sum_{k=1}^{N}s(i,k)\psi[2^{-m}(k-n)], \tag{3}$$

2.2. Wavelet energy (WE) parameter

For basic wavelet selection, Modulated Gaussian (Morlet), second derivative of a Gaussian, Haar, and Shannon are the common ones. Among these wavelets, since the Haar wavelet achieves the best performance in our experiments, it is used in this paper. The Harr wavelet in Eq. (3) is described by

$$\psi[2^{-m}(k-n)]=\begin{cases}1, & 0\leqslant 2^{-m}(k-n)\leqslant\frac{1}{2}\\ -1, & \frac{1}{2}\leqslant 2^{-m}(k-n)\leqslant 1\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

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 as shown in Fig. 1, where different values of scales corresponds to different band-pass filters. The outputs of DTWT at differ-

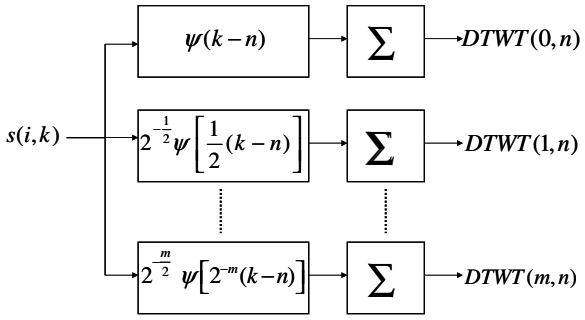

Fig. 1. Filter bank implementation of the wavelet transform.

ent scales contain different amounts of speech and noise information, and only the crucial scale(s) that contains maximum speech signal information and is robust to noise should be used. Loosely speaking, it has been found that the perception of a particular frequency by the auditory system is influenced by the energy in a critical band of frequencies around the frequency (Allen, 1985). Energy of the crucial scale is then adopted as detection parameter for distinction between speech and noise.

To find the crucial scale, some observations on the effect of additive noise are made on different scales of DTWT. Fig. 2a shows the clean speech. The amplitudes of the DTWT of the speech with SNR5 white noise at different scales are shown in Figs. 2b–d, which show the amplitudes at scales $a_o^m = 2^4, 2^6$, and 2^8 for $n = 0$–N, respectively. As shown in Fig. 2, it is found the speech section corresponds to large DTWT amplitude values. Thus, summation of the amplitudes could be used as a parameter to stand for the amount of speech signal information. It is also found from Fig. 2 and several other observations 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 wavelet energy (WE), is computed as follows,

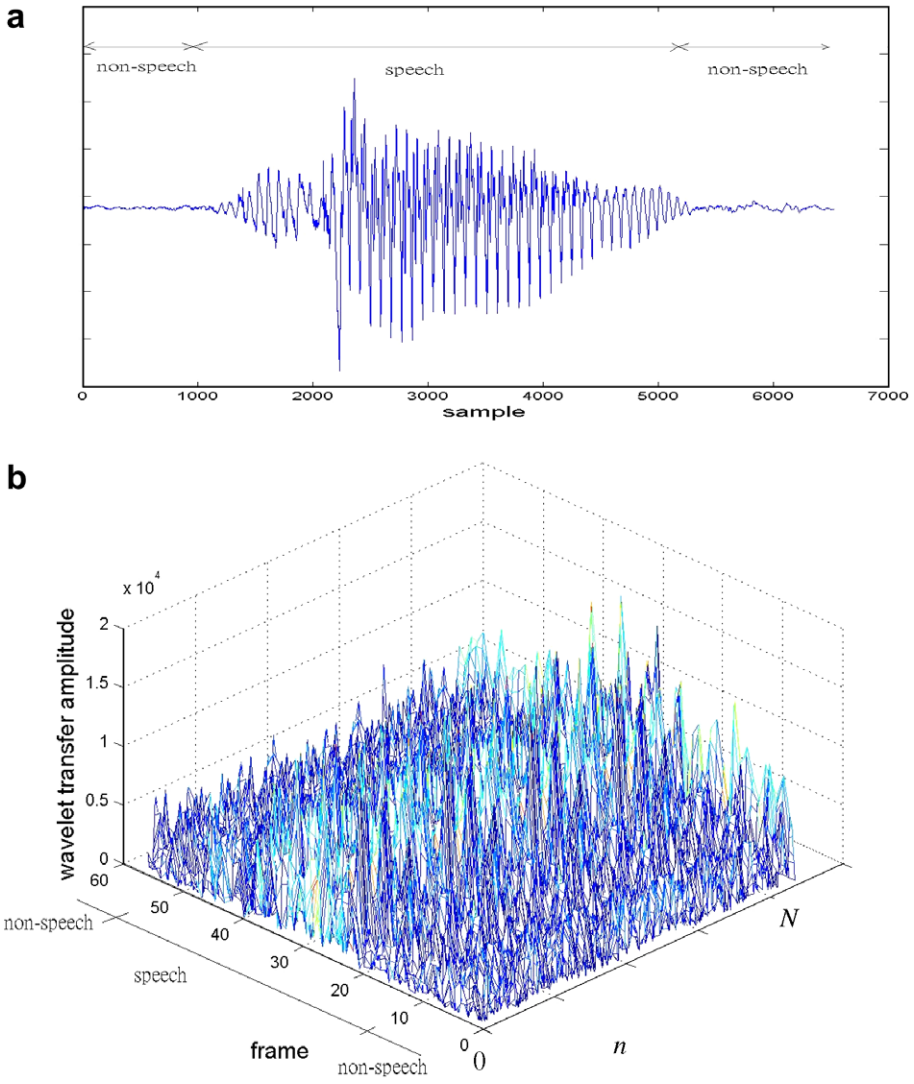

Fig. 2. (a) The clean speech. (b)–(d) The DTWT amplitudes of the speech with SNR5 white noise when $a_0^m =$ (b)2^4, (c)2^6, and (d) 2^8.

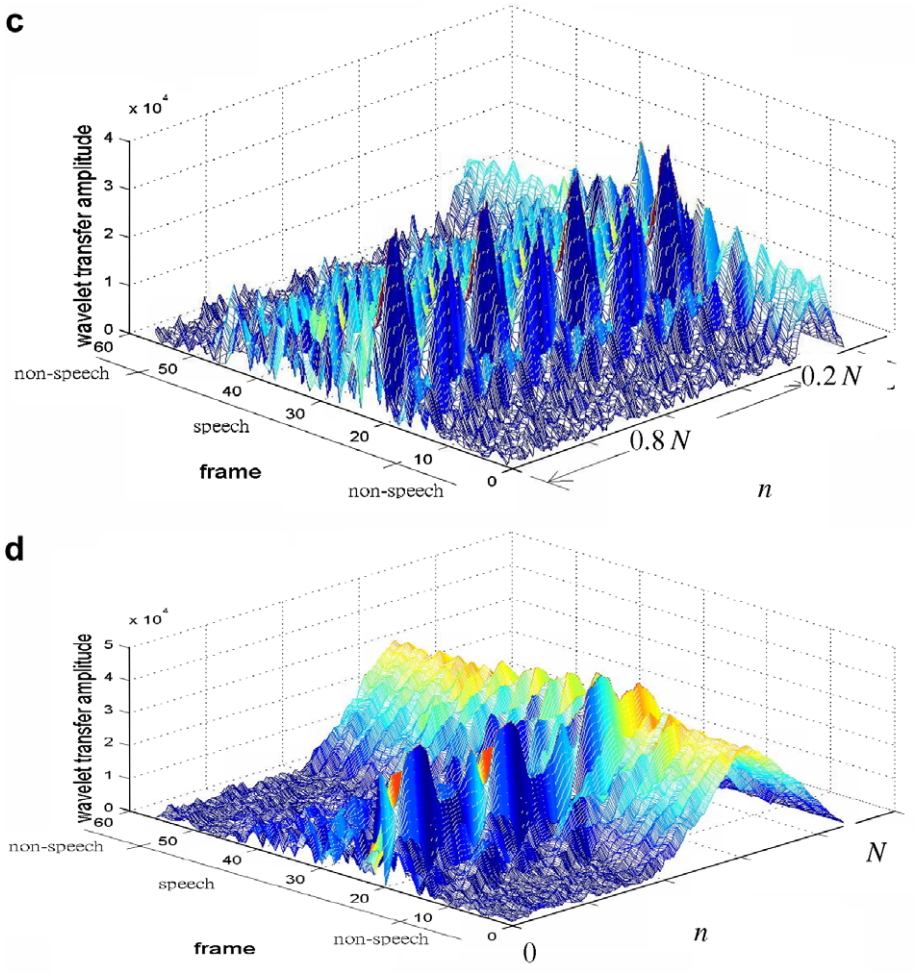

Fig. 2 (*continued*)

$$\mathrm{WE}(m) = \sum_{n=0}^{0.8N} \left| \mathrm{DTWT}(m,n) \right| = \sum_{n=0}^{0.8N} \left| \frac{1}{2^{-m/2}} \sum_{k=1}^{N} s(i,k)\psi(2^{-m}(k-n)) \right| \quad (5)$$

For each speech and non-speech frame in a recorded sequence, the wavelet energies with scales $m = 2$, 4, 6, and 8 are computed at SNR = 5, 10, 15, and 20. For each scale, variance of each frame at these SNR values is computed. The average result of the frames show that the WE at $m = 6$ achieves the smallest variance. That is, the WE computed at the scale $m = 6$ is more robust to noise than the other scales. So it is used as a detection feature. The robustness of WE at $m = 6$ can also be found from Fig. 2. Fig. 2b shows that the amplitudes are similar for speech and non-speech frames at the scale of $a_0^m = 2^4$. Fig. 2d shows that at larger values of n, the amplitudes does not match well with the speech interval. In Fig. 2c, it is found that at the scale of, $a_0^m = 2^6$, distribution of the amplitudes matches well with the speech interval. Here, only a constant scale in DTWT is used. This is in contrast to the adaptive or refined TF parameter (Wu & Lin, 2000, 2001), where different frequency bands by discrete Fourier transform are used for different detected frames. The use of a constant scale eases the computation and it is found in Section 4, the experiment section, that it is more robust than the refined TF parameter. Although the WE is proposed based on results in white noise, it is also suitable for detection for other types of noise as shown in Section 4.

Fig. 3a shows a clean speech and its corresponding WE parameters of each frame. The speech with white noise and its corresponding WE parameters at SNR20 and SNR5 are shown in Fig. 3b and c, respectively. This example shows that the WE parameter can robustly represent the energy of speech signal at different SNRs. This is in contrast to the refined or adaptive TF parameter (Wu & Lin, 2000, 2001), where the frequency energy should be summed with time energy to obtain the TF parameter, which obviously contains energy of speech and noise. To find the speech energy, an additional noise_time parameter should be used in Wu and Lin, 2000, 2001 to estimate the noise energy at each frame. At variable noise-level environment, estimating the noise is not easy, which degrades the performance the TF related parameters.

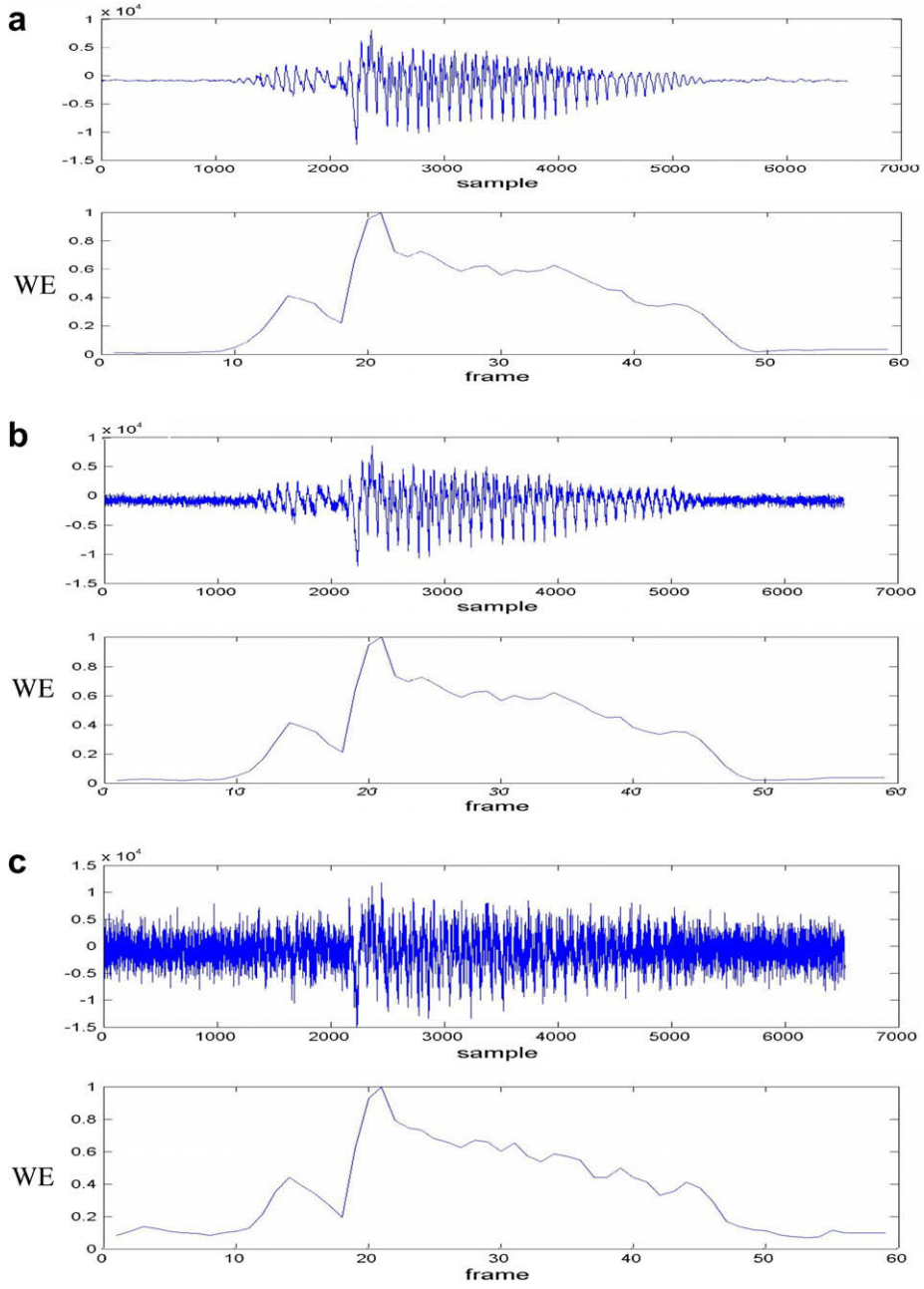

Fig. 3. (a) The WEs of clean speech (b) the WEs of speech with white noise added at SNR20 (c) the WEs of speech with white noise added at SNR5.

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 ZCR for the ith frame is computed by

$$\mathrm{ZCR}(i) = \frac{1}{N-1}\sum_{k=1}^{N-1}\frac{1}{2}|\mathrm{sgn}[s(i,k)] - \mathrm{sgn}[s(i,k-1)]| \qquad (6)$$

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. The major concern of using ZCR is that if it may degrade the detection performance at low SNR. Fortunately, the ZCR is not dependent on amplitude, and therefore, is less effected when noise-level changes.

To see the robustness of the used WE and ZCR parameters, distributions of speech and non-speech frames on the two-dimensional parameter plane rare plotted. Fig. 4 shows the distributions of speech and non-speech frames with white noise added at different SNRs. The results show the most speech frames concentrate on a certain region at different SNRs. It is also found from this figure that using only the WE or ZCR parameter cannot detect the speech well. Thus, these two parameters are both used.

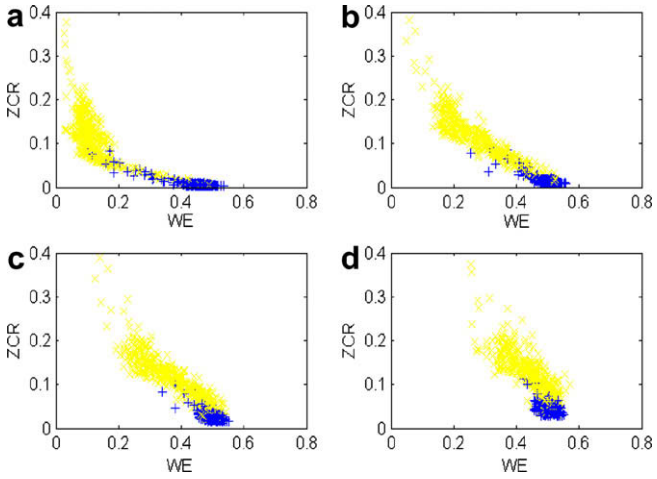

Fig. 4. Distributions of speech frame ("+") and non-speech frame ("×") on the WE and ZCR parameters plane. (a) SNR20 (b) SNR10 (c) SNR5 (d) SNR0.

3. Singleton-type recurrent neural fuzzy network (SRNFN) detector

To deal with problems with temporal characteristics, many recurrent fuzzy neural networks have been proposed. Structurally, one category of these networks focuses on the inclusion of external feedback as recurrence (Kumar, Kumar, Shankar, Tiwari, & Kumar, 2007; Mastorocostas & Theocharis, 2002; Theocharis & Vachtsevanos, 1997; Zhang & Morris, 1999). The disadvantage of this category is that the order of both control input and network output participating in the recurrent model should be known in advance. Another category of recurrent fuzzy network includes fuzzy models with internal recurrence (Juang, 2002; Lee & Teng, 2000). In (Juang, 2002), a TRFN was proposed, where the TSK-type consequence is a linear combination of input variables plus a constant. The performance of the TRFN was demonstrated to outperform the compared recurrent networks. In this paper, detector design using a singleton-type recurrent neural fuzzy network (SRNFN) is proposed. The SRNFN modifies the TSK-type consequence of TRFN to be fuzzy singletons, such that the consequent values are less sensitive to noises contained in input variables. In this section, the structure and learning of SRNFN are introduced below.

A. Structure of SRNFN

The structure of SRNFN is shown in Fig. 5. A network with two external inputs and a single output is considered here for convenience. In contrast to a six-layered TRFN, there are only five layers in SRNFN. This five-layered network realizes a recurrent fuzzy network of the following form:

Rule1: IF $x_1(t)$ is A_{11} and $x_2(t)$is A_{12} and $h_1(t)$ is G
THEN $y(t+1)$ is b_1 and $h_1(t+1)$ is w_{11}
and $h_2(t+1)$ is w_{21}

Rule2: IF $x_1(t)$ is A_{21} and $x_2(t)$ is A_{22} and $h_2(t)$ is G
THEN $y(t+1)$ is b_2 and $h_1(t+1)$ is w_{12}
and $h_2(t+1)$ is w_{22},

where A and G are fuzzy sets, w and b are fuzzy singleton values functioned as the consequent parameters for inference output h and y, respectively. In Fig. 5, a network constructed by the above two rules is shown. There are two external input variables x_1 and x_2, and single output y. Accordingly, SRNFN has two nodes in layer 1 and one node in layer 5, respectively. To give a clear understanding of the mathematical function of each node, functions of SRNFN are described layer by layer below.

Nodes in layer 1 are input nodes. The node only transmits external input values x_j and internal input values h_i to layer 2. Each node in layer 2 acts as a membership function. Two types of membership functions are used in this layer. For external input x_j, Gaussian membership functions which locally map the input spatial space to the output space are used, and the mathematical function is

$$\mu_i = \exp\left\{-\frac{(x_j - m_{ij})^2}{\sigma_{ij}^2}\right\}, \tag{7}$$

where m_{ij} and σ_{ij} are, respectively, the center and the width of the Gaussian membership function. For internal variable h_i, sigmoid membership function $1/(1+\exp\{-h_i\})$ is used. Each internal variable has a single corresponding fuzzy set. Unlike the case in the external input space domain where local Gaussian membership functions are used, a single sigmoid membership function instead of multiple Gaussian membership functions is adopted on the domain of the internal variable to simplify the feedback structure. Each node in layer 3 is called a rule node. The output of each node in this layer is determined by fuzzy AND operation. Here, the product operation is utilized

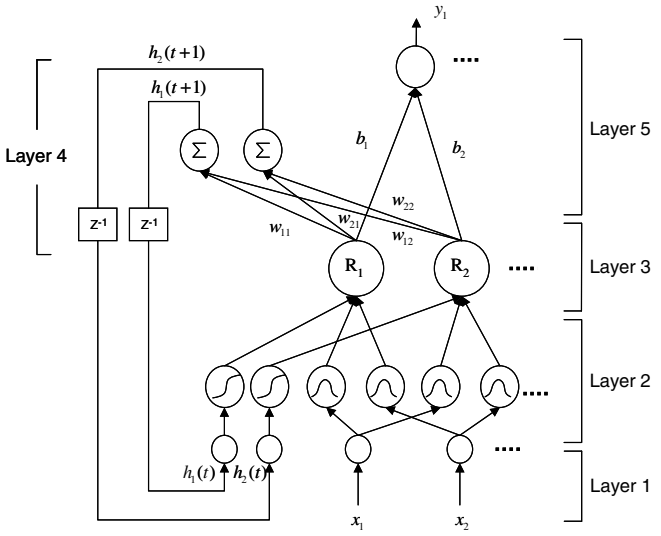

Fig. 5. Structure of the singleton-type recurrent neural fuzzy network (SRNFN).

to determine the firing strength of each rule. The function of each rule is

$$\Phi_i = \frac{1}{1+\exp\{-h_i\}} \cdot \prod_{j=1}^{n} \mu_j = \frac{1}{1+\exp\{-h_i\}} \cdot \exp\left\{ -\sum_{j=1}^{n} \frac{(x_j - m_{ij})^2}{\sigma_{ij}^2} \right\} \quad (8)$$

Each node in layer 4 is called a context node and performs defuzzification operation for internal variable h_i. The number of internal variables in this layer is equal to the number of rule nodes. The link weights represent the fuzzy singleton values in the consequent part of the internal rules. The simple weighted sum is calculated in each node,

$$h_i = \sum_{j=1}^{r} w_{ij}\mu_j \quad (9)$$

where r is the number of rules. As in Fig. 5, the delayed value of h_i is fed back to layer 1 and acts as an input variable to the precondition part of a rule. Each rule has a corresponding internal variable h_i and is used to decide the influence degree of temporal history to the current rule. Nodes in layer 5 are called defuzzification nodes. Each node performs weighted average operation for output y. The node in this layer computes the output signal y of the SRNFN. The output node together with links connected to it act as a defuzzifier. The mathematical function is

$$y = \frac{\sum_{j=1}^{r} b_j \Phi_j}{\sum_{j=1}^{r} \Phi_j} \quad (10)$$

B. Learning of SRNFN

The task of constructing the SRNFN is divided into two subtasks: structure learning and parameter learning. There are initially no rules in SRNFN, and the rules are constructed by on-line structure and parameter learning. The objective of the structure learning is to decide the number of fuzzy rules, initial location of membership functions, and initial consequent parameters. On the contrary, the objective of parameter learning is to tune the free parameters of the constructed network to an optimal extent.

The first task in structure learning is to decide when to generate a new rule. The criterion of generating a new rule is the same as that used in TRFN, where the spatial firing strength $F_i(\vec{x}) = \prod_{j=1}^{n} \mu_j(\vec{x})$ is used as the criterion to decide if a new fuzzy rule should be generated. For each incoming data $\vec{x}(t)$, find

$$I = \arg \max_{1 \leqslant i \leqslant r(t)} F_i(\vec{x}(t)) \quad (11)$$

where $r(t)$ is the number of existing rules at time t. If the rule with the maximum firing strength is smaller than a pre-defined threshold, i.e., $F_I \leqslant F_{\text{in}}(t)$, then a new rule is generated, where $F_{\text{in}}(t) \in (0,1)$ is a pre-specified threshold that decays during the learning process. Once a new rule is generated, the initial centers and widths of the corresponding membership functions are computed by

$$m_{(r(t)+1)i} = x_i(t)$$
$$\sigma_{(r(t)+1)i} = \beta \cdot \sum_{j=1}^{n} \frac{(x_j - m_{Ij})^2}{\sigma_{Ij}^2} \quad (12)$$

for $i = 1 \ldots n$, where $\beta > 0$ decides the overlap degree between two clusters. In this paper, β is set at 0.8. The number of fuzzy sets in each external input dimension is equal to the number of fuzzy rules. Generation of a context node in layer 4 accompanies the generation of a rule. The initial link weights w_{ij} in layer 4 are set as random values in $[-1,1]$. The above process is repeated for every incoming training data. During the learning process, a new recurrent rule is generated, one after another, and a whole SRNFN is constructed finally.

As to the parameter learning, the cost function used is

$$E(t+1) = \frac{1}{2}(y(t+1) - y^d(t+1))^2, \quad (13)$$

and the parameters are tuned by real-time recurrent learning algorithm (Juang, 2002).

When SRNFN is used as a detector, its input is a two-dimensional vector composed of WE and ZCR. The output of SRNFN indicates whether the corresponding frame is speech signal or noise. For this purpose, there are two outputs in SRNFN. When the input frame is speech signal, the desired output is $(1,0)$. On the contrary, when the input is noise, the desired output is $(0,1)$. During test, if the SRNFN outputs are (y_1, y_2) then the corresponding input frame is detected as speech signal if $y_1 > y_2 + \theta$, where θ is a threshold.

4. Experiments

The wave files of speech are recorded by 11 kHz sample rate, mono channel and 16-bit resolution. The length of each frame is 10 (ms). For detector training, the training sequence length is 13 s and is shown in Fig. 6. It consists of 20 Mandarin words and added white noise, whose energy level increases from the start to SNR = 0 and decreases till the end of the sequence. Training of SRNFN with inputs WE and ZCR is performed. Structure of SRNFN is influenced by the structure learning threshold F_{in}. Training performance of SRNFN with different thresholds is evaluated. The classification rate defined by

$$\text{Classification rate} = \frac{\text{number of correctly detected frames}}{\text{total number of frames in training sequence}} \quad (14)$$

and decision threshold $\theta = 0$ is employed to evaluate the training performance. After several trials, the threshold $F_{\text{in}} = 0.56$ that achieves the best performance is used. The final structure, total number of parameters, and training performance of SRNFN is shown in Table 1.

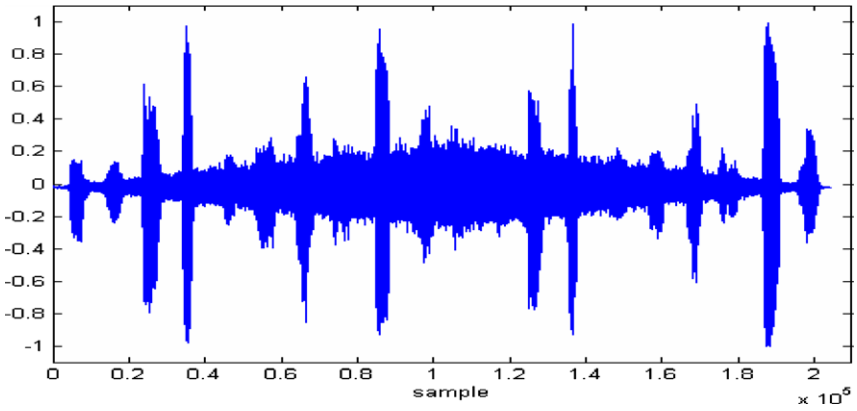

Fig. 6. The sequence of speech used for network training, where there are 20 Mandarin words with white noise added in the sequence.

Table 1
The size of detectors and training performance of different detection methods

Methods	WE-based SRNFN	WE-based TRFN	WE-based SVM	RTF-based RSONFIN(1)	RTF-based RSONFIN(2)
Training sequence	1	1	1	1	2
Structure	6 rules	6 rules	1284 SVs	7 rules	7 rules
Parameter number	66	84	3853	90	90
Training performance (%)	71.25	74.23	84.53	64.04	71.11

For comparison, training of WE-based TRFN, WE-based SVM and refined TF (RTF)-based RSONFIN method (Wu & Lin, 2001) are performed. The inputs and number of rules in TRFN are the same as those in SRNFN. The final structure, total number of parameters, and training performance of TRFN are also shown in Table 1. Since there are more free parameters in the consequence of TRFN than in SRNFN, training result of TRFN is a little better than SRNFN. The inputs of SVM are the same as SRNFN, and a Gaussian kernel-based SVM is used. Different cost parameters C are tried, and the best value is $C = 500$. Table 1 shows the training performance of SVM and the total number of support vectors and parameters. The SVM achieves a better training result than SRNFN with the costing of using a much larger set of parameters. For RTF-based RSONFIN method, three features are used for detection, namely RTF, ZCR and noise_time. During training, the noise_time parameter that measures the noise energy is assumed to be known in advance and is computed before training. The training sequence in Fig. 6 is also used for training RTF-based RSONFIN. Moreover, training of RTF-based RSONFIN using two sequences consisting of 40 words is performed. Two sequences of training data are utilized because the succeeding test results show that a bad detection performance is achieved if only one training sequence is used for RTF-based RSONFIN. Structure and training performance of RTF-based RSONFIN with one and two sequences of training are also shown in Table 1, where training result of RTF-based RSONFIN using two training sequences is similar to WE-based SRNFN.

For performance evaluation, the white noise and other types of noises used in following experiments are taken from the NBOISE-92 database (Varga & Steeneken, 1993) and are downsampled to 11 kHz. To get a quick view on the performance of WE-based SRNFN and RTF-based RSONFIN with continuously varying SNR levels, two illustrative examples are experimented. The results are shown in Fig. 7. In Fig. 7b, white noise with changing noise-levels is added to the clean speech in Fig. 7a. In Fig. 7c, white noise of increasing level is added to the clean speech in Fig. 7a. For RTF-based RSONFIN test, the parameter "noise_time" that estimates noise energy is computed on-line. The results in Fig. 7 show that when the noise-level changes successively or when the time between two words is too short, inaccurate estimation of "noise_time" occurs leading to a poor estimation result of RTF-based RSONFIN. For WE-based SRNFN, estimation of noise energy is unnecessary and much better detection results are achieved.

In the following experiments, the detection rate (DR) when false positive rate (FPR) is equal to 10% is used for performance evaluation of different detection methods in testing, where the detection rate (DR) is defined as

$$\mathrm{DR} = \frac{\text{number of correctly detected speech frames}}{\text{number of speech frames}} \tag{15}$$

and the false positive rate (FPR) is defined as

$$\mathrm{FPR} = \frac{\text{number of noise frames detected as speech frames}}{\text{number of noise frames}} \tag{16}$$

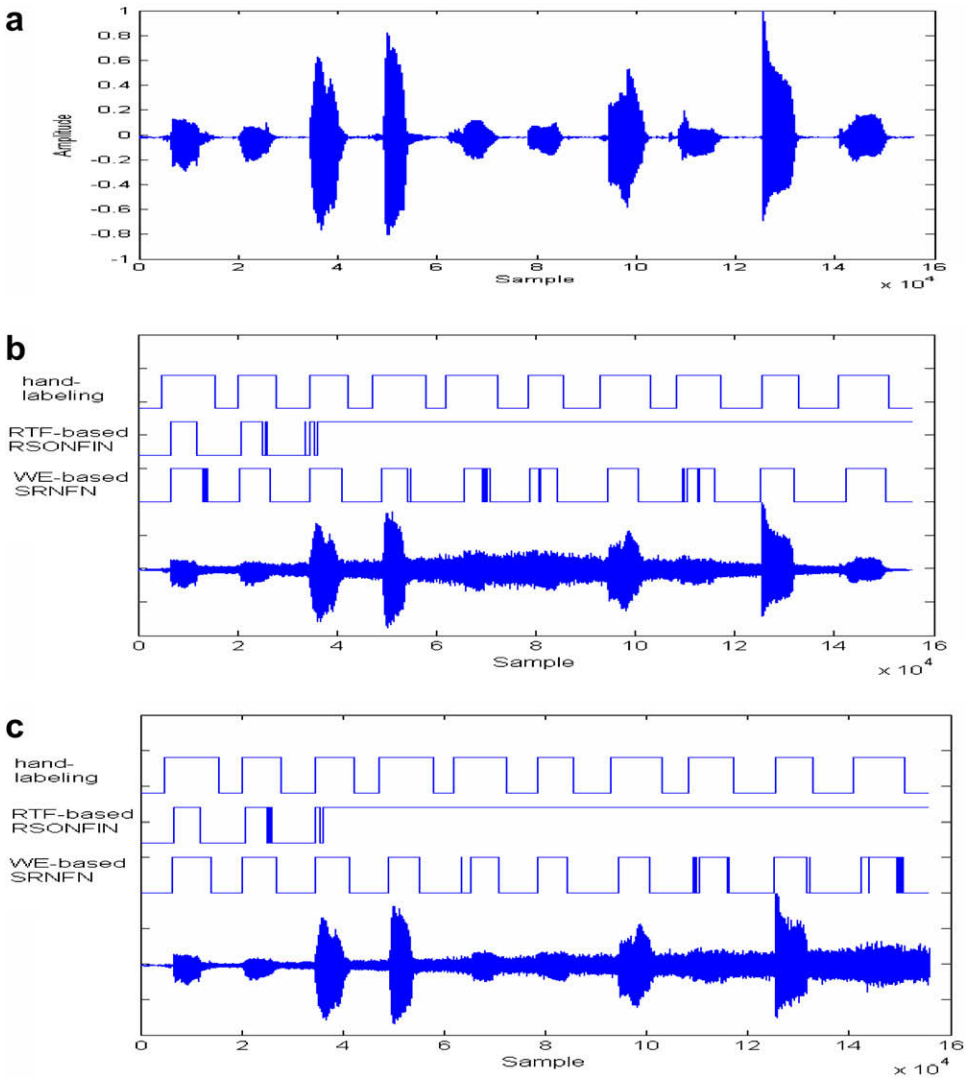

Fig. 7. (a) The clean speech. (b), (c) The speech with white noise added and detection results of RTF-based RSONFIN and WE-based SRNFN.

Experiment 1

For testing, the speech database is made up of sequences of words, where each sequence consists of ten isolated Mandarin words "0", "1",..., "9". Each sequence is uttered fifty times. Thus, there are fifty test sequences consisting of 500 words in all. The total number of test samples in these test sequences is about 7.5×10^7. The noise added to the speech sequence is of variable-levels. One of the test sequences consisting of 10 words with white noise of variable levels added during the sequence is shown Fig. 8. A detection result for one of the test sequence with white noise of various levels added at average SNR5 by WE-based SRNFN is shown in Fig. 9. The detection rates when FPR = 0.1 for the WE-based SRNFN, WE-based TRFN, WE-based SVM, and RTF-based RSONFIN methods when variable-level white noise is added to each sequence with average SNR 5–20 in the test database are shown in Table 2. The comparison results show that the WE-based SRNFN has a higher detection rate than WE-based TRFN and RTF-based RSONFIN at different SNRs. For TRFN, the consequence is a linear combination of input variables, which makes the output being sensitive to noises in input variables, especially at lower SNRs. This is the reason why training of WE-based TRFN is better than WE-based SRNFN but the test performance is poorer. Detection rate of WE-based SVM is only little higher than WE-based SRNFN though it can achieve a higher training rate. One reason is the SVM is a static classifier that can not capture the temporal relationship of speech sequence as SRNFN does. One major disadvantage of using SVM is that needs more memory for parameters storage.

Experiment 2

To see the robustness of a detection method, other types of unknown noises, including engine room, babble, and cockpit noises, are also tested. The source of the babble noise is 100 people speaking in a canteen. For these noises, they are added in the same manner. Detection rates

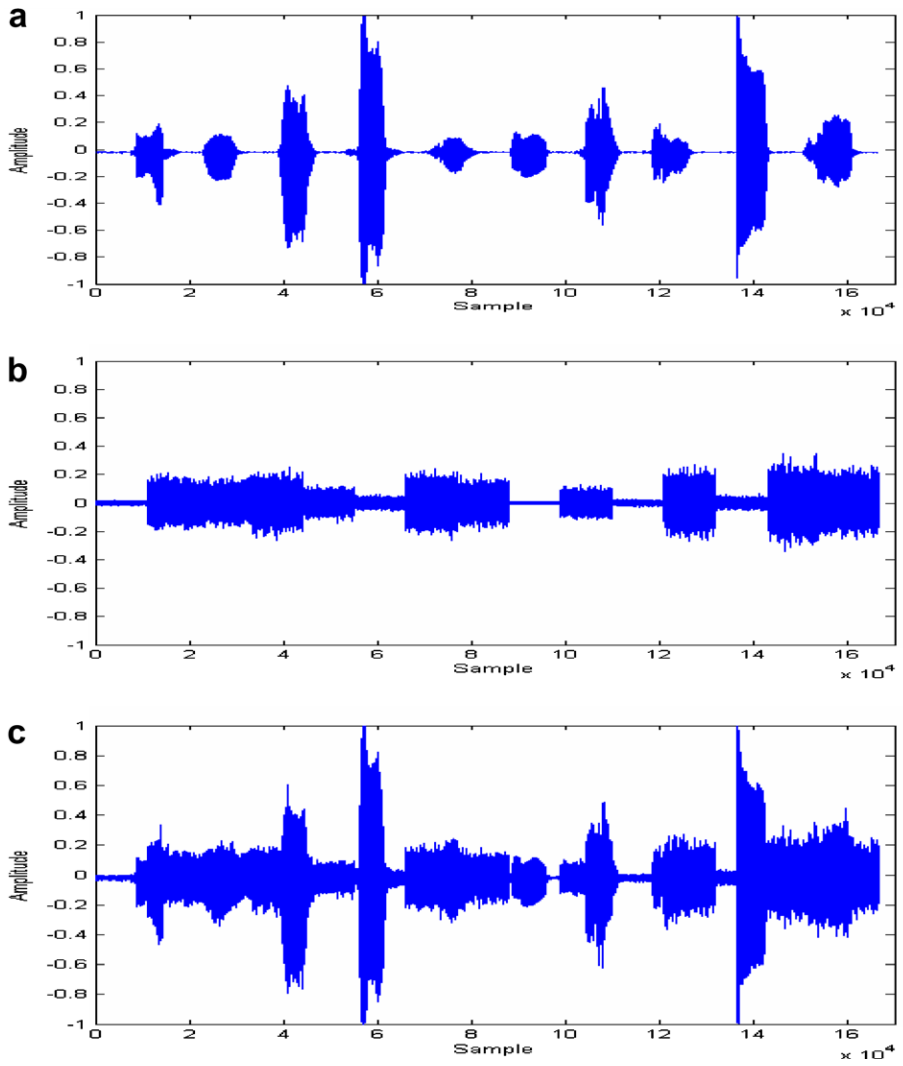

Fig. 8. (a) The clean speech. (c) The variable-levels white noise with average SNR5. (b) The noisy speech with white noise of various levels added at average SNR5.

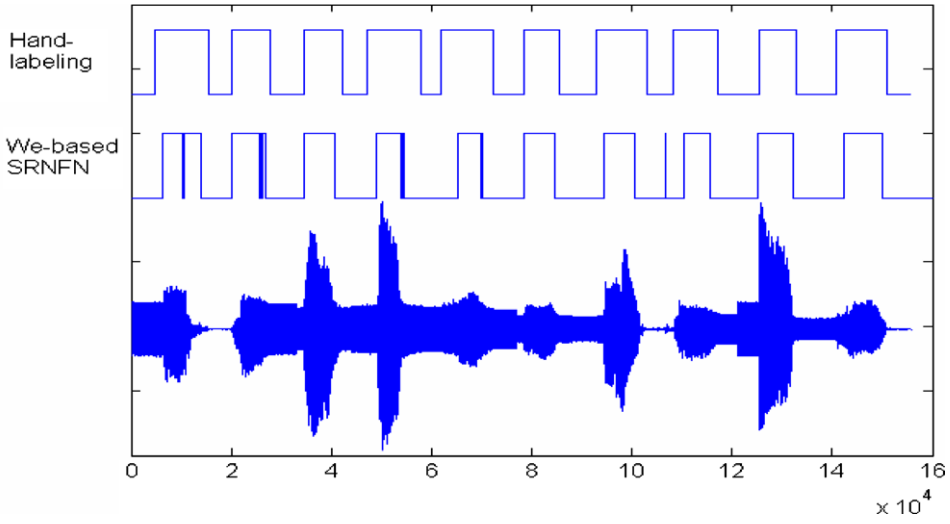

Fig. 9. A detection result for the noisy speech with white noise of various levels added at average SNR5 by WE-based SRNFN in experiment 1.

Table 2
Detection rates for different methods with white noise in experiment 1

SNR	WE-based SRNFN	WE-based TRFN	WE-based SVM	RTF-based RSONFIN(1)	RTF-based RSONFIN(2)
20	92.66	91.21	94.37	35.64	80.71
15	89.68	80.23	92.84	36.91	74.51
10	86.65	68.91	90.66	32.36	57.37
5	82.74	50.23	81.26	19.04	36.22

(FPR = 0.1) of WE-based SRNFN, WE-based SVM, and RTF-based RSONFIN for these noises are shown in Tables 3–5. For engine room and cockpit noises, the results show that detection rates of WE-based SRNFN is higher than RTF-based RSONFIN at different SNRs. For babble noise, WE-based SRNFN achieves higher rates than RTF-based RSONFIN at SNR = 20 and 15, and lower rates at SNR = 10 and 5. Since babble noise is also from human speaking, it is more difficult to distinguish the noise from detected words than other types of noises. This is the reason why the detection rates of WE-based SRNFN are slower than those of the other types of noises, especially at low SNRs. Performance of WE-based SRNFN and WE-based SVM is very similar.

Experiment 3

In this experiment, a larger speech database made up of more sequences of words is used for testing. In addition to the sequence that contains the ten Mandarin digits, another two sequences are used. Each of the two additional sequences contains another ten different Mandarin words. That is, there is a total of 30 Mandarin words in the three test sequences. Each of the three sequences is uttered fifty times by the same speaker, and there are a total of 150 test sequences consisting of 1500 words in all. The total number of test samples in these test sequences is about 2.2×10^8. The noise added to the speech sequence is of variable-levels. The testing process of this experiment is the same as experiment 1 except that a larger database is tested in this experiment.

Table 3
Detection rates for different method with engine room noise in experiment 2

SNR	WE-based SRNFN	WE-based SVM	RTF-based RSONFIN(1)	RTF-based RSONFIN(2)
20	94.46	95.05	37.44	84.87
15	93.13	92.93	37.49	82.16
10	89.68	88.7	38.37	76.26
5	81.26	79.75	30.46	54.3

Table 4
Detection rates for different methods with babble noise in experiment 2

SNR	WE-based SRNFN	WE-based SVM	RTF-based RSONFIN(1)	RTF-based RSONFIN(2)
20	90.17	90.64	41.13	85.06
15	84.57	85.5	40.07	81.73
10	71.89	73.91	38.17	73.54
5	45.09	46.01	30.87	56.44

Table 5
Detection rates for different methods with cockpit noise in experiment 2

SNR	WE-based SRNFN	WE-based SVM	RTF-based RSONFIN(1)	RTF-based RSONFIN(2)
20	91.14	92.76	38.76	84.17
15	87.38	88.08	37.77	81.05
10	81.55	80.97	36.96	74.02
5	67.41	67.12	29.56	56.06

Table 6
Detection rates for different methods with white noise in experiment 3

SNR	WE-based SRNFN	WE-based SVM	RTF-based RSONFIN(1)	RTF-based RSONFIN(2)
20	90.24	92.94	35.12	76.55
15	86.03	90.88	35.04	72.44
10	81.68	88.1	32.63	61.04
5	78.02	77.34	23.02	40.52

Detection rates of WE-based SRNFN and RTF-based RSONFIN at different SNR when FPR = 0.1 are shown in Table 6. The WE-based SRNFN also achieves better performance than RTF-based RSONFIN and is very similar to WE-based SVM. For WE-based SRNFN, the experiment shows similar results to experiment 1 though another unknown twenty words are included in the new test database.

5. Conclusion

This paper has proposed a new WE-based SRNFN detector for speech detection in variable noise-level environments. The WE parameter is used to extract speech energy in the wavelet-transformed domain. Only the crucial scale in wavelet transformation is used, which shows both computation efficiency and detection effectiveness. Experimental results have shown that the combination of WE and ZCR parameters is feasible and robust for speech detection over variable-level white noise and unknown noises. For detector design, the SRNFN detector has been proposed. The SRNFN is shown to be more robust than TRFN and requires a smaller number of parameters than SVM for similar detection performance. Experimental results show that the WE-based SRNFN shows better performance than the RTF-based RSONFIN for almost all types of noises at different SNRs. In the future, the proposed detection method will be combined with speech recognizers to obtain a robust speech recognition system.

References

Allen, J. B. (1985). Cochlear modeling. *IEEE Acoustic, Speech, and Signal Processing Magazine, 2*, 3–29.

Britelli, F., Casale, S., & Cavallaro, A. (1989). Robust voice activity detector for wireless communications using soft computing. *IEEE Selected Areas In Communication, 16*(9), 1818–1829.

Chan, Y. T. (1995). *Wavelet basics*. Kluwer Academic Publishers.

Enqing, D., Guizhong, L., Yatong, Z., & Xiaodi, Z. (2002). Applying support vector machines to voice activity detection. *Proceedings of International Conference on Signal Processing*, 1124–1127.

Haigh, J. A., & Mason, J. S., (1993). Robust voice activity detection using cepstral features. In *Proceedings of IEEE region 10 Conference on Computer, Communication, Control and Power Engineering 3* (pp. 321–324).

Juang, C. F. (2002). A TSK-type recurrent fuzzy network for dynamic systems processing by neural network and genetic algorithm. *IEEE Transactions on Fuzzy Systems, 10*(2), 155–170.

Juang, C. F., & Lin, C. T. (1998). An on-line self-constructing neural fuzzy inference network and its applications. *IEEE Transactions on Fuzzy Systems, 6*, 12–32.

Junqua, J. C., Mak, B., & Reaves, B. (1994). A robust algorithm for word boundary detection in the presence of noise. *IEEE Transactions on Speech and Audio Processing., 2*, 406–412.

Kumar, S., Kumar, S., Shankar, P. R., Tiwari, M. K., & Kumar, S. B. (2007). Prediction of flow stress for carbon steels using recurrent self-organizing neuro-fuzzy networks. *Expert Systems with Applications, 32*(3), 777–788.

Lamel, F., Rabiner, L. R., Rosenberg, A. E., & Wilson, J. G. (1981). An improved endpoint detector for isolated word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 29*(4), 777–785.

Lee, C. H., & Teng, C. C. (2000). Identification and control of dynamic systems using recurrent fuzzy neural networks. *IEEE Transactions on Fuzzy Systems, 8*(4), 349–366.

Mastorocostas, P. A., & Theocharis, J. B. (2002). A recurrent fuzzy-neural model for dynamic system identification. *IEEE Transactions on Systems Man and Cybernetics, Part B: Cybernetics, 32*(2), 176–190.

Qi, Y., & Hunt, B. R. (1993). Voiced-unvoiced-silence classification of speech using hybrid features and a network classifier. *IEEE Transactions on Speech and Audio Processing, 1*, 250–255.

Rouat, J., Liu, Y. C., & Morissette, D. (1997). Pitch determination and voiced/unvoiced decision algorithm for noisy speech. *Speech Communication, 21*(3), 191–207.

Savoji, M. H. (1989). A robust algorithm for accurate end-pointing of speech signals. *Speech Communication, 8*(1), 45–60.

Theocharis, J. B., & Vachtsevanos, G. (1997). Recursive learning algorithms for training fuzzy recurrent models. *International Journal of Intelligent Systems, 11*(12), 1059–1098.

Varga, A., & Steeneken, H. J. M. (1993). Assessment for automatic speech recognition: II. NOISE-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication, 12*, 247–251.

Wu, G. D., & Lin, C. T. (2000). Word boundary detection with Mel-scale frequency bank in noisy environment. *IEEE Transactions on Speech and Audio Processing, 8*(5), 541–553.

Wu, G. D., & Lin, C. T. (2001). A recurrent neural fuzzy network for word boundary detection in variable noise-level environments. *IEEE Transactions on Systems, Man, and Cyber Part B: Cybernetics, 31*(1), 84–97.

Wu, B. F., & Wang, K. C. (2005). Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments. *IEEE Transactions on Speech and Audio Processing, 13*(5), 762–775.

Xianbo, X., & Guangshu, H. (2005). An incremental support vector machine based speech activity detection algorithm. *Proceedings of The 27th Annual International Conference on Engineering in Medicine and Biology Society*, 4224–4226.

Zhang, J., & Morris, A. J. (1999). Recurrent neuro-fuzzy networks for nonlinear process modeling. *IEEE Transactions on Neural Networks, 10*(2), 313–326.