SIViP
DOI 10.1007/s11760-013-0479-5

ORIGINAL PAPER

Robust voice activity detection directed by noise classification

Jamal Saeedi · Seyed Mohammad Ahadi · Karim Faez

Received: 12 June 2012 / Revised: 14 April 2013 / Accepted: 14 April 2013

Abstract In this paper voice activity detection (VAD) is formulated as a two-class classification problem using support vector machines (SVM). The proposed method combines a noise robust speech processing feature extraction process together with SVM models trained in different background noises for speech/non-speech classification. A multi-class SVM is also used to classify background noises in order to select SVM model for VAD. The proposed VAD is tested with TIMIT data artificially distorted by different additive noise types and is compared with state-of-the-art VADs. Experimental results show that the proposed VAD can extract speech activity under poor SNR conditions, and it is also insensitive to variable levels of noise.

Keywords Voice activity detection · Perceptual wavelet packet transform · Noise classification · Support vector machine

1 Introduction

Voice activity detection (VAD) is a process, which can detect speech and non-speech segments from a speech signal. A typical conversational speech is characterized by a speech-to-non-speech ratio of forty to sixty [1]. Hence, the use of VAD could improve the channel capacity as well as the power consumption of voice communication systems. VAD can also help in many speech-related applications such as speech coding [2], automatic speech recognition [3], and speech enhancement systems [4].

The basic procedure of most VADs in use today consists of a feature extraction step followed by a decision part. The feature extraction step extracts acoustic parameters from the input speech signal for discrimination of speech and non-speech segments. The conventional acoustic parameters are the short-time energy levels, zero-crossing rates, pitch period, and spectral difference. Then, the decision part makes use of these acoustic parameters with some decision rules to determine the VAD result. The decision rules could be simple threshold values or complex statistical models. It is possible to use a trained classifier such as support vector machines (SVM) for the decision rule part. This paper shows an effective method employing SVM for VAD in noisy environments.

Regardless of the decision rules, using appropriate features is very important in the performance of VAD. Since speech signals are non-stationary and contain many transient components, it is not appropriate to use a fixed time–frequency resolution method for feature extraction in VAD, especially in noisy environments. Wavelet transform is based on time–frequency signal analysis. The wavelet analysis adopts a windowing technique with variable-sized regions. It allows the use of long time intervals, when we want more precise low-frequency (LF) information, and shorter regions, where we want high-frequency (HF) information. Here, perceptual wavelet packet transform (PWPT) is used as a tool for feature extraction. PWPT is utilized to adjust the decomposition tree structure of the conventional wavelet packet transform (WPT) in order to approximate the critical bands of the psychoacoustic model as close as possible. The primary reason for embedding the psychoacoustic model in the PWPT is that humans are capable of detecting the desired speech in a noisy environment without prior knowledge of

J. Saeedi (✉) · S. M. Ahadi · K. Faez
Electrical Engineering Department, Amirkabir University of Technology, 424 Hafez Ave., Tehran, Iran
e-mail: jamal.saeedi@yahoo.com

S. M. Ahadi
e-mail: sma@aut.ac.ir

K. Faez
e-mail: kfaez@aut.ac.ir

Published online: 30 April 2013

the noise [5]. Therefore, the human's auditorium system is capable of distinguishing between different acoustical noises.

In noisy environments, the performance of VADs is severely affected. Commonly, there are two main methodologies to deal with noise in VADs. In the first approach, a speech enhancement method is usually used for noise reducing [6], and in the second one, noise robust features are extracted from noisy speech for VADs [7]. There are many different acoustical noises in the environment (such as babble, street, car, etc.), which result in performance degradation of VADs. Usually, the effect of different noises is not considered in VADs. By modifying the processing according to the type of background noise, the performance of VAD can be enhanced. This requires noise classification, which has been used in many applications, such as robust speech recognition [8], and speech enhancement [9].

The remainder of this paper is organized as follows. The PWPT is briefly reviewed in Sect. 2. Section 3 gives the description of the noise classification algorithm and the SVM-based VAD directed by noise classification. Section 4 illustrates experimental results and compares to other methods. Finally, conclusions are given in Sect. 5.

2 Perceptual wavelet packet transform

The mathematical work of the WPT was first proposed by Coifman [10]. WPT is a wavelet transform where the discrete-time (sampled) signal is passed through more filters than the DWT, and therefore, there are more HF sub-bands to appropriately represent the signal.

The PWPT method is developed to adjust the decomposition tree structure of the conventional WPT in order to approximate the critical bands of the psychoacoustic model. In the psychoacoustic model, frequency components of sounds can be integrated into critical bands that refer to bandwidths at which subjective response becomes significantly different [11]. Critical bands are important in understanding many auditory phenomena, such as perception of loudness, pitch, and timbre. One class of critical band scales is called Bark scale. Based on the measurements by Zwicker et al. [12], the Bark scale z can be approximately expressed in terms of the linear frequency by:

$$z(f) = 13 \arctan\left(7.6 \times 10^{-4} f\right) + 3.5 \arctan\left(1.33 \times 10^{-4} f\right)^2 \text{ [Bark]} \quad (1)$$

where f is the linear frequency in Hertz. The corresponding critical bandwidth (CBW) of the center frequencies can be expressed by

$$\text{CBW}(f_c) = 25 + 75\left(1 + 1.4 \times 10^{-6} f_c^2\right)^{0.69} \text{ [Hz]} \quad (2)$$

where f_c is the center frequency. Theoretically, the range of human's auditorium frequency spreads from 20 to 20,000 Hz and covers approximately 25 Barks.

Since the Bark scale is a function of linear frequency, the first step of constructing the PWPT is to set the sampling rate of speech signals in order to determine the valid Bark numbers. In this paper, the underlying sampling rate was chosen to be 8 kHz, yielding a bandwidth of 4 kHz. Within this bandwidth, there are approximately 17 critical bands [11]. The tree structure of the PWPT can be constructed as shown in Fig. 1a. The corresponding frequency bandwidths of the PWPT tree are shown in Fig. 1b. It contains 16 decompo-

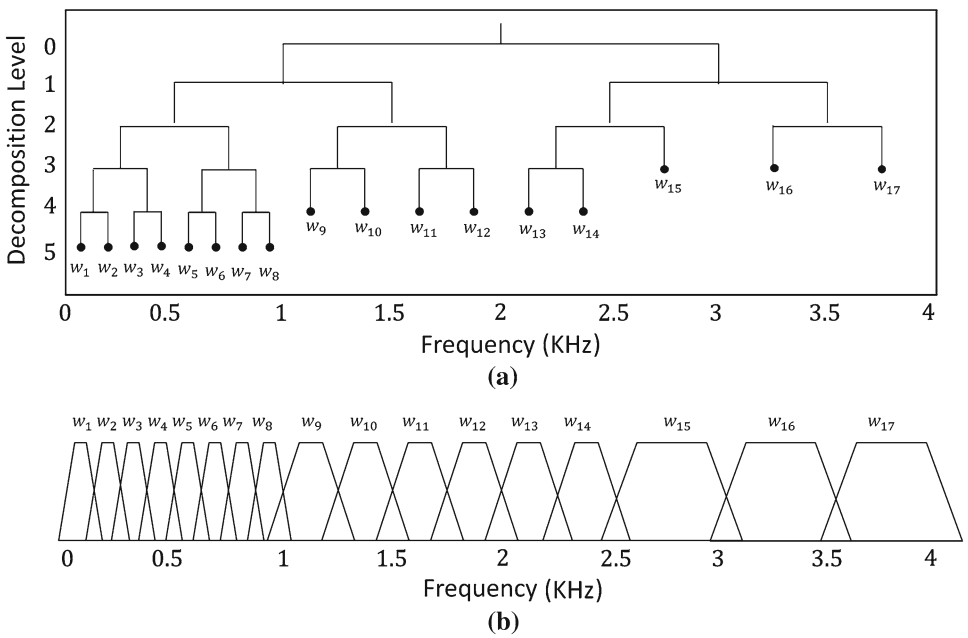

Fig. 1 **a** The tree structure of the PWPT and **b** the frequency bandwidths for the PWPT tree, where w_j defines the wavelet coefficient of the jth sub-band of PWPT, where $j = 1$–17. (The sampling rate is chosen to be 8 kHz, yielding a bandwidth of 4 kHz. Within this bandwidth, there are approximately 17 critical bands.)

sition cells with five decomposition stages to approximate these 17 critical bands.

3 Proposed method

As described in Sect. 1, the proposed robust VAD uses a classification-based technique, in which classification models are trained using noisy speech signals in specific environments. Given a speech signal, a set of features for noise classification is extracted from a short period of silence at the beginning of signal. Features extracted from the silence portion are then used to identify the type of environment. Once knowing the environment type, the recognizer selects a corresponding model for classifying the rest of signal as speech or non-speech.

First, the environment or noise classification module is constructed using PWPT and SVM. The computational overhead of the noise classification module should be kept as low as possible, so that the overall system can achieve an acceptable processing time. Then, a particular SVM is trained on noisy speech signals with various levels of SNR. Figure 2 shows block diagram of the proposed method, which consists of a number of essential stages:

1. A small frame of 1,024 samples at the beginning of the speech signal, which was expected to be silence, is used for noise classification.
2. Multi-class SVM classifier is used to identify the type of noise.
3. The input signal $x(n)$ sampled at 8 kHz is decomposed into 32-ms overlapped frames with a 15-ms window shift. Then, four types of features are extracted from each frame for the classification task.
4. The appropriate SVM model based on noise classification result is selected for classifying noisy speech frame as speech or non-speech.

In the following subsections, we have provided more detailed explications of the VAD process.

3.1 Noise classification

The goal of noise classification is to identify the type of speech environment. Here, a simple model based on multi-class SVM classifier is used to identify the type of noise.

3.1.1 Feature extraction

The choice of signal features is usually based on a priori knowledge of the nature of the signals to be classified. A variety of signal features have been used for this purpose, including low-level parameters such as the zero-crossing rate, signal bandwidth, spectral centroid, signal energy, and mel-frequency cepstral coefficients.

As discussed in Sect. 1, PWPT is selected as a tool for feature extraction. For better discrimination between different noises in the PWPT domain, three features including mean, standard deviation, and entropy are extracted from each sub-band as:

$$M_j = \frac{1}{N_j}\sum_{k=1}^{N_j}\left|w_j(k)\right| \tag{3}$$

$$\mathrm{Std}_j = \sqrt{\frac{1}{N_j}\sum_{k=1}^{N_j}\left(\left|w_j(k)\right| - \left|\overline{w_j}\right|\right)^2} \tag{4}$$

$$\mathrm{En}_j = -\sum_{l=0}^{L} h_j(l) \times \mathrm{Log}_2\left(h_j(l)\right) \tag{5}$$

where $w_j(k)$ defines the kth coefficient of the jth sub-band of PWPT, where j = 1–17, N_j is the number of coefficients in jth sub-band, and $k = 1, 2, \ldots, N_j$. h_j is normalized histogram of absolute values of wavelet coefficients at w_j sub-band, and L is the number of corresponding histogram levels.

At the end of feature extraction step, a stack of 51-dimensional feature vector is obtained. Now, PCA is used in order to extract the most significant features. PCA has been widely used for feature extraction in pattern recognition. The main concept of PCA is to project the original feature vector onto principal component axes. These axes are orthogonal

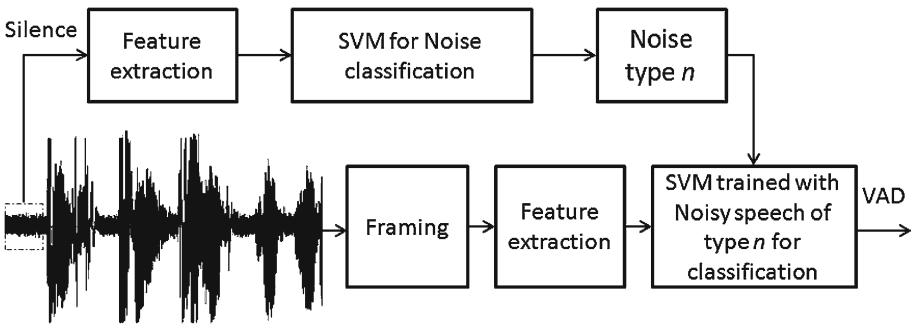

Fig. 2 Block diagram of the proposed VAD method

and correspond to the directions of greatest variance in the original feature space. Therefore, projecting input vectors onto this principal subspace allows reducing the redundancy in the original feature space as well as the dimension of input vectors.

3.1.2 Noise classification results

Five types of noise from NOISEX-92 [13] including factory, white, pink, babble, and car were preprocessed by reducing the sampling rate to 8 kHz. A total of 34,590 frames (17.75 min), equally distributed between the 5 classes, have been used for the classification. Figure 3 shows the projected feature vector (weight vectors) obtained from 34,590 feature vectors derived from all five types of environment (factory, white, pink, babble, and car) onto a three-dimensional space using PCA. The type of noise is easily identified, as shown in Fig. 3.

Noise classification problem is a multi-class classification. Therefore, we have used the "one-against-one" approach in which $k(k-1)/2$ classifiers are constructed and each one uses the training data from two different classes. The first use of this strategy on SVM was in [14]. In classification, we have used a voting strategy: Each binary classification is considered to be a voting where votes can be cast for all data points x. In the end, each point is designated to a class with maximum number of votes. The SVM model has been trained using LIBSVM software tool [15].

The performance of SVM can be controlled through the term C, which is the penalty parameter that controls the trade-off between the complexity of the decision function and the number of misclassified training examples, and the kernel parameters called hyper-parameters. These parameters influence the number of the support vectors and the maximization margin of the SVM. As mentioned, the performance of SVM-based classifier can be controlled through hyper-parameters of the SVM. Here, genetic algorithm (GA) implemented in Matlab optimization toolbox is used to find optimal parameters. Two parameters are considered to be optimized using GA as follows: C is the penalty parameter, and σ is the kernel parameter.

The whole dataset is divided into three parts: training set (30 %), evaluation set (30 %), and test set (40 %). The train and evaluation sets are used for finding optimal parameters by GA, and the fitness function is the classification error rate, which should be minimized. In addition, the population size is set to 15, and generation size is set to 10 for genetic optimization as a good compromise between accuracy and complexity.

The number of principal components of feature vector is another parameter, which should be selected for the classification. For this purpose, the minimum error rate using GA is obtained for different numbers of principal components in the classification algorithm. Figure 4 shows the classification error rate versus the number of principal components used for the classification. The classification error rate is almost fixed after 20 principal components; therefore, we have chosen 20 principal components for the classification task. In addition, the optimal parameters of SVM using GA are as follows: $C = 1.93$ and $\sigma = 0.53$.

After obtaining optimal parameters, SVM is retrained using train and evaluation sets (60 % of whole dataset), and then the trained SVM is tested using optimal parameters with test set (40 % of whole dataset), and the classification error rate is 1.62 %. A detailed presentation of the classification results for each class is given in the form of a classification matrix. Table 1 shows that the classification accuracies ranging from 96.46 to 100 % were obtained for different classes. Volvo and white noises are completely classified, and bab-

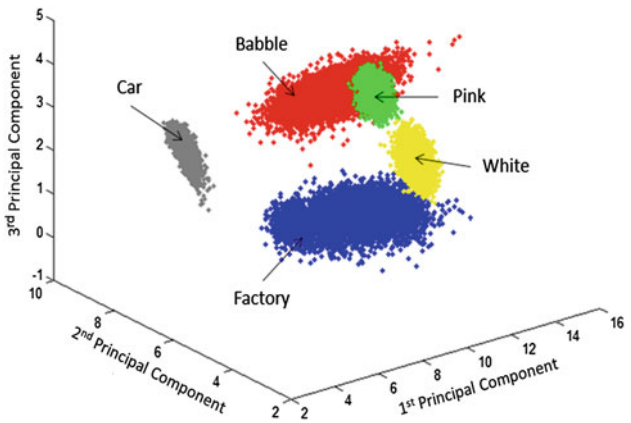

Fig. 3 Distributions of PCA-reduced features computed from five different noises

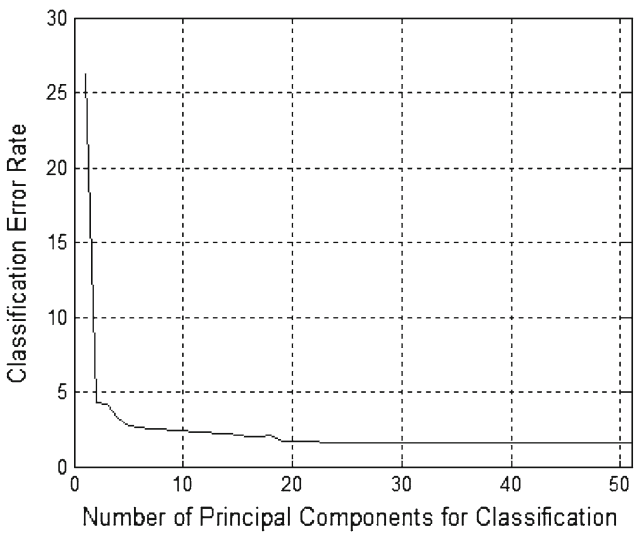

Fig. 4 The classification error rate versus the number of principal components used for noise classification

Table 1 The classification matrix

Noise type	Babble	Factory	Pink	Volvo	White
Babble	97.94	1.86	0	0	0
Factory	2.05	97.43	3.53	0	0
Pink	0	0.70	96.46	0	0
Volvo	0	0	0	100	0
White	0	0	0	0	100

ble, factory, and pink noises are misclassified with error rates ranging from 1 to 4 %.

It is also interesting to show level-by-level results showing the improvements caused by different feature groups in the noise classification algorithm. Figure 5 shows the classification accuracy results versus different feature groups including PCA feature vector used in the classification task. From the result, it is evident that each feature has important effect in the noise classification algorithm using SVM.

3.2 VAD directed by noise classification

After identifying the noise type using noise classification algorithm, a robust model based on noise type is constructed for a variety of signal-to-noise ratios (SNRs).

3.2.1 Feature extraction

The feature extraction step is used to increase discrimination between noise (non-speech) and speech for the classification task. The algorithm for feature extraction is stated as follows. The input signal $x(n)$ sampled at 8 kHz is decomposed into 32-ms overlapped frames with a 15-ms window shift. Then, four types of features are extracted from each frame for the classification task: (1) sum of autocorrelation (SAC) sequence, (2) entropy, (3) sum of local maxima (SLM) of power spectral density (PSD), and (4) mean of PWPT subbands.

3.2.1.1 Sum of autocorrelation sequence The periodic property is an inherent characteristic of speech signals and is commonly used to characterize speech. The periodic properties of speech signals are exploited to accurately extract speech activity. In fact, voiced or vowel speech sounds have a stronger periodic property than unvoiced sounds and noise signals. Consequently, the well-known autocorrelation function (ACF) is defined in the time domain to evaluate the periodic intensity of each frame. The biased estimate of the ACF is shown as:

$$R(k) = \frac{1}{N} \sum_{n=0}^{N-k-1} x(n)x(n+k), \quad k = 0, 1, \ldots, N-1 \quad (6)$$

The SAC sequence is used as the first feature:

$$\mathrm{SAC} = \sum_{k=0}^{N-1} R(k) \quad (7)$$

Figure 6c illustrates the first feature in different frames of speech signal when input speech is contaminated by white noise. From this figure, it is observed that the SAC of speech segments has more obvious peaks than that of non-speech and white noise.

3.2.1.2 Entropy Entropy is a statistical measure of randomness and measures information content in a signal. Because of periodic property of speech signal and random nature of noise, the entropy metric can effectively discriminate them.

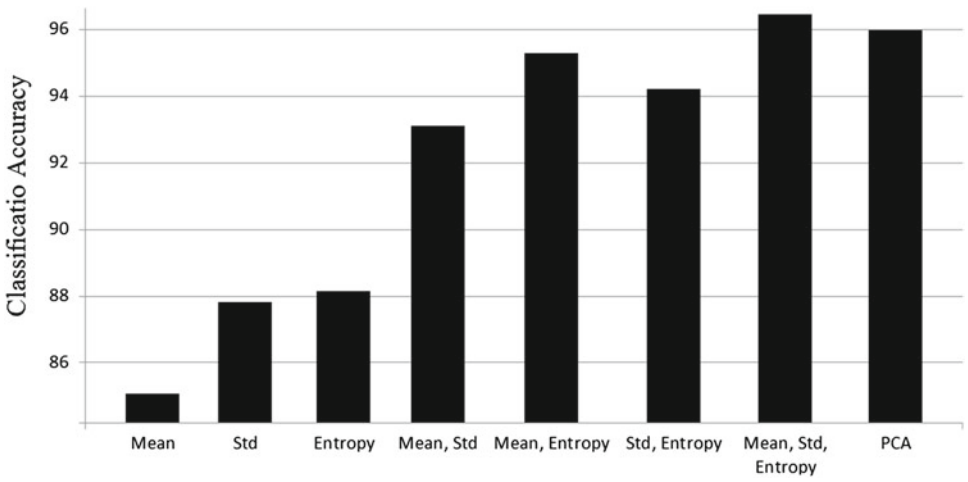

Fig. 5 The classification accuracy results versus different feature groups including PCA feature vector used in the noise classification algorithm

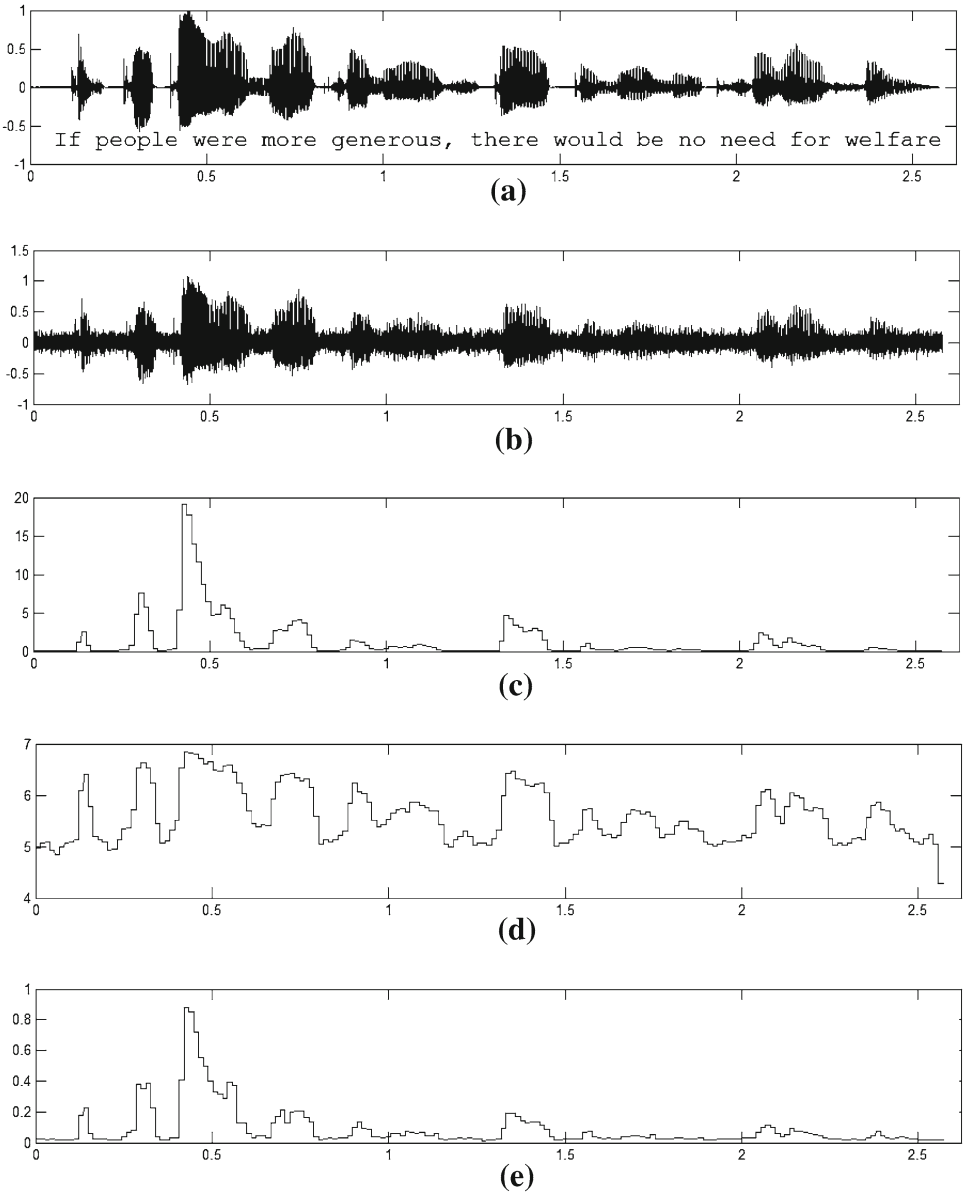

Fig. 6 **a** Clean speech signal, **b** noisy signal distorted by white noise (5 dB SNR), and different features extracted from noisy speech signal in consecutive frames including **c** sum of autocorrelation sequence (SAC), **d** entropy, and **e** sum of local maxima of power spectral density (SLM)

Therefore, we can use this measure to discriminate noise and speech in each frame (see Fig. 6d). The entropy index is defined as:

$$H = -\sum_{k=0}^{K} h(k) \times \mathrm{Log}_2(h(k)) \tag{8}$$

where h is the normalized histogram of the absolute value of the speech signal $x(n)$ in a frame with length $N, n = 0, 1, \ldots, N-1$, and K is the number of corresponding histogram levels.

3.2.1.3 Sum of local maxima of power spectral density The PSD of a stationary random process is mathematically related to the correlation sequence by the discrete-time Fourier transform. In general, the more correlated or predictable a signal, the more concentrated its power spectrum, and conversely, the more random or unpredictable a signal, the more spread its power spectrum. Therefore, the power spectrum of a signal can be used to deduce the existence of repetitive structures or correlated patterns in the signal process.

Welch's method is used for PSD estimation, which is a nonparametric algorithm [16]. Welch's method is attained by averaging modified periodograms from overlapped and windowed segments. After obtaining PSD for each frame, SLM of it is used as the third feature for the classification task. Figure 7 clearly illustrates that local maxima of PSD can effectively discriminate noise and speech signals (it is also observable from Fig. 6e).

3.2.1.4 Mean of PWPT sub-bands Using PWPT, the input speech signal can be decomposed into 17 sub-bands, which

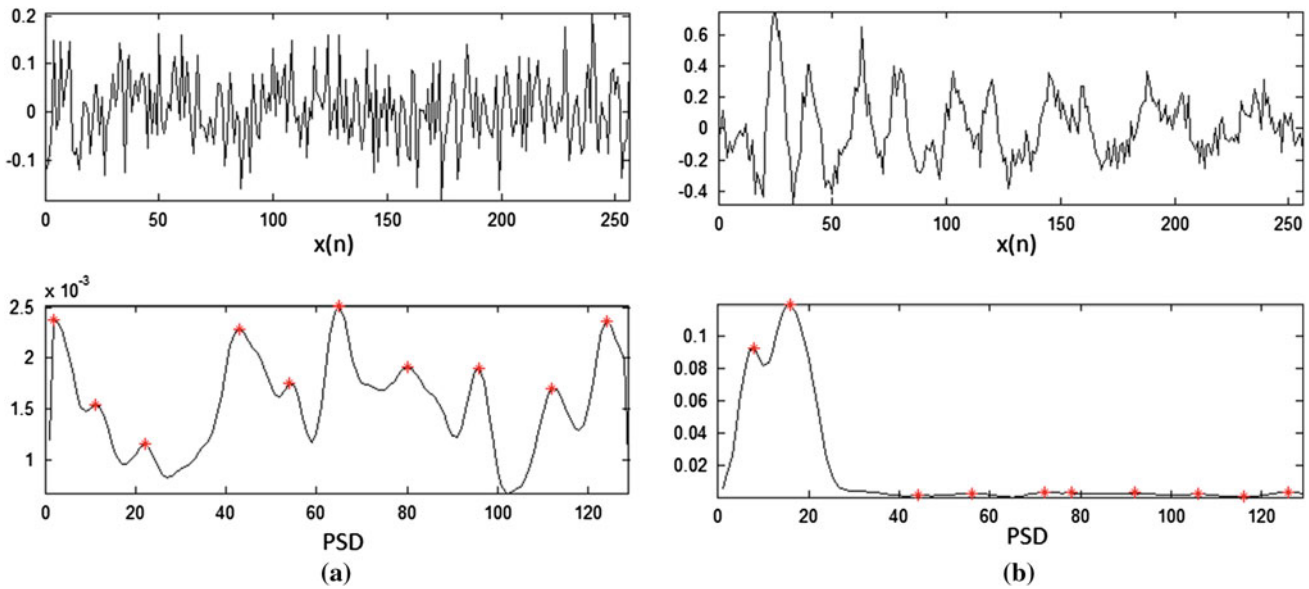

Fig. 7 Local maxima of power spectral density using Welch's method of two sample frames of **a** noise and **b** speech signals

are corresponding to wavelet coefficient sets. The white noise exists in all frequency sub-bands; however, this is not true for other noises. Therefore, for better discrimination between noise and speech, mean of noisy speech in each PWPT sub-band expressed in (3) is used as fourth feature.

In addition to the four features mentioned above, delta of each feature is used to exploit the correlation between neighboring frames in speech signal. The delta function for each feature is defined as follows:

$$\Delta F = 2F(n) - (F(n-1) + F(n+1)) \tag{9}$$

where n is the frame number.

At the end of feature extraction step, we have a stack of 40-dimensional feature vector (FV) for each frame to classify it as speech or non-speech:

$$\text{FV} = [\text{SAC}, H, \text{SLM}, M, \Delta\text{SAC}, \Delta H, \Delta\text{SLM}, \Delta M]$$

where each of the M and ΔM consists of 17 features corresponding to the 17 sub-bands of PWPT.

As a final stage, we also applied PCA to these features in order to extract the most significant ones to be used. The effect of different feature groups consisting of PCA feature vector in the proposed VAD algorithm is separately evaluated in terms of classification accuracy, and it will be discussed in the next subsection.

3.2.2 Construction of SVM model based on noise type

Having the feature vector, SVM is used for the classification problem. Using SVM, a robust model based on noise type is constructed for a variety of SNRs. In other words, a particular SVM model is trained on noisy speech with various levels of SNR. Clean speech, whose SNR exceeds 30 dB, is also combined in the training set of each noisy acoustic model.

In order to construct a particular SVM model for different noise types, 110 utterances of the TIMIT corpus [17] are used, in which each speech sample is artificially distorted by adding a particular noise type such as babble, white, factory, pink, and volvo provided in [13], at different SNR levels (clean, 30, 25, 20, 15, 10, 5, 0 dB). Therefore, a total of 880 speech files are considered for each noise to construct SVM model for it. It should be mentioned that the speech utterances are visually labeled into speech and non-speech classes. The SVM model has been trained using LIBSVM software tool [15].

The same experiments similar to the noise classification task are done for finding optimal parameters including C (the penalty parameter), σ (the kernel parameter), and the number of principal components. The whole dataset for each noise type is divided into two parts: training set (60 %) and evaluation set (40 %). The training and evaluation sets are used for finding optimal parameters using GA, and the fitness function is the classification error rate, which should be minimized. In order to obtain the number of principal components, the minimum error rate using GA is obtained for different numbers of principal components in the classification algorithm. Figure 8 shows weight vectors obtained from 85,685 feature vectors derived from 880 speech files distorted by white noise at different SNR levels onto a three-dimensional space using PCA.

Figure 9 shows the classification error rate versus the number of principal components used for the classification of noisy speech distorted by white noise. Based on the results,

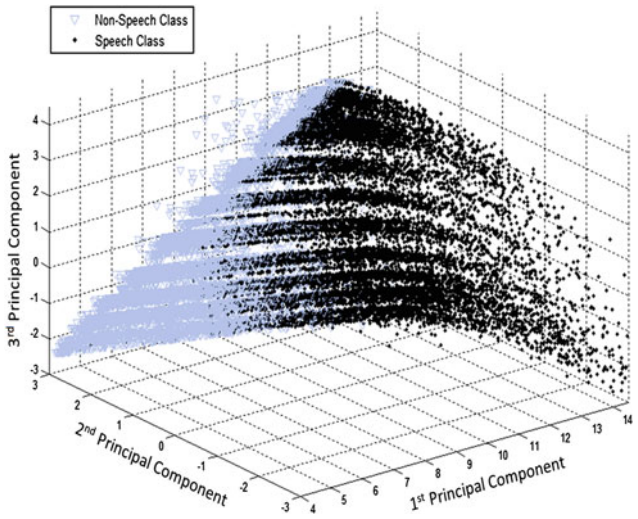

Fig. 8 Distributions of PCA-reduced features computed from 880 speech files distorted by white noise

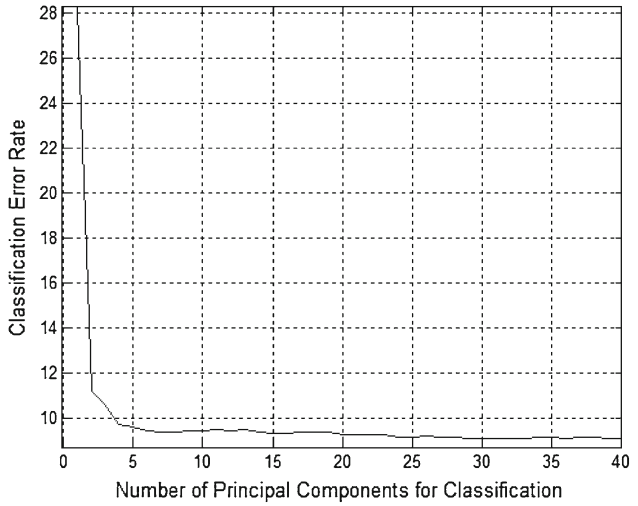

Fig. 9 The classification error rate versus the number of principal components used for speech/non-speech classification

we have chosen 25 principal components for the classification. In addition, Table 2 shows the optimal SVM parameters obtained by GA to generate SVM model for different noise types. After finding the optimal parameters for different noise types, the SVM classifier is retrained using the whole dataset.

Figure 10 shows the classification accuracy results versus different feature groups including PCA feature vector used in the classification task. From the result, it is evident that each feature is individually important in the speech/non-speech classification using SVM for different noise types. Most effective feature is mean of wavelet coefficients, which is obtained from 17 sub-bands of PWPT. The results are also shown that the use of delta feature has considerably improved the classification accuracy. It should be mentioned that different noise types result in different classification accuracies (white noise gives the best and babble noise gives the worst classification results). In the proposed algorithm, noise type information is used to improve VAD result, and different features are used to discriminate between specific noise type and speech signal. However, noises like babble (which looks like speech signal) and factory (which is highly non-stationary) are hard to model, and therefore, the classification accuracies are lower for them.

Table 2 The optimal SVM parameters obtained by genetic algorithm to generate SVM model for different noise types

Noise type	SVM parameters	
	C	σ
Babble	2.465	0.761
Factory	1.336	0.638
Pink	1.640	0.685
White	1.895	2.794
Volvo	0.616	0.289

Support vector machines (SVM) for VADs are examined in the literature [18,19]. The idea is very simple, using a feature extraction step and a SVM classifier. The main differences in our algorithm with SVM-based VADs are using of the noise classification and also a new robust feature vector. Here, the classification accuracy results for two feature vectors, which have been used for SVM-based VAD, are compared with the proposed feature vector. It can be seen from Table 3 that our proposed feature vector outperforms the other methods in different noise conditions.

4 Experimental results

The proposed VAD was evaluated in terms of the ability to discriminate speech from non-speech at different SNRs. By reducing the sampling rate to 8 kHz, 130 utterances of the TIMIT corpus were preprocessed and used for evaluating the proposed VAD algorithms [17]. Each speech sample was artificially distorted by adding five types of noise from NOISEX-92, i.e., factory, white, pink, babble, and volvo [13], at different SNR levels (20, 15, 10, 5, 0 dB).

The performances of the algorithms are evaluated by comparing the percentage of correct classifications (non-speech and speech) with manually marked decisions on all test utterances (see Fig. 11). The performance metrics are percentage of correct non-speech identification (Pcn) and percentage of correct speech identification (Pcs) described in [1].

Figure 12 shows speech/non-speech discrimination as a function of the SNR for different noise types. The best

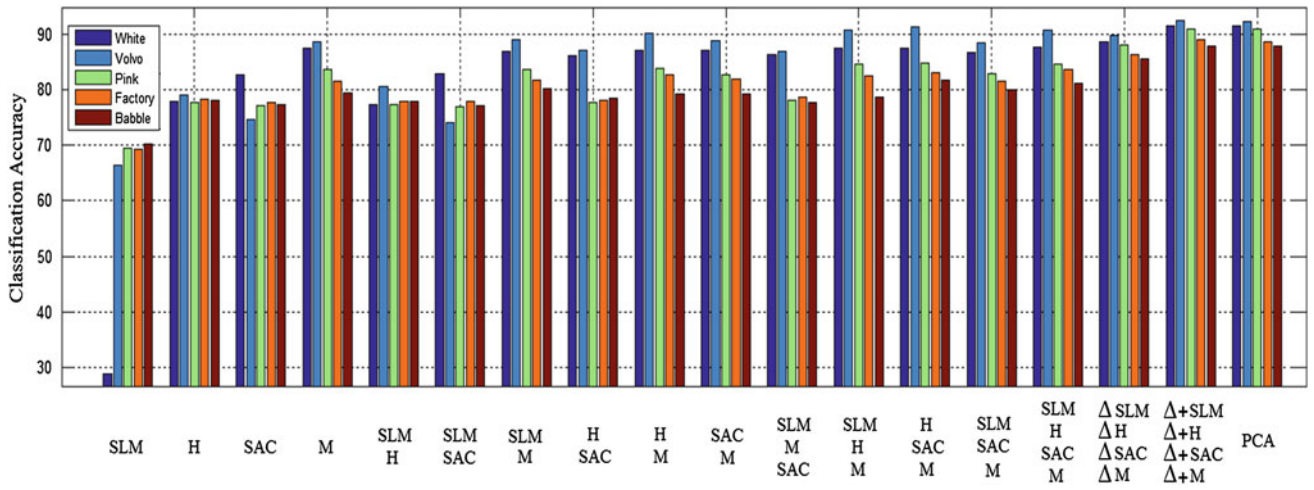

Fig. 10 Classification accuracy results versus different feature groups including PCA feature vector used in speech/non-speech classification for different environments. *H* is entropy feature, SLM is sum of local maxima of power spectral density, SAC is sum of autocorrelation sequence, and *M* is the mean of PWPT sub-bands

Table 3 The classification accuracy results for different features

Feature	Noise type				
	White	Volvo	Pink	Factory	Babble
Proposed	91.5	92.17	90.9	88.54	87.92
MFCC [18]	86.42	91.13	85.03	84.81	80.84
LTSD [19]	74.35	89.66	76.35	80.90	83.68

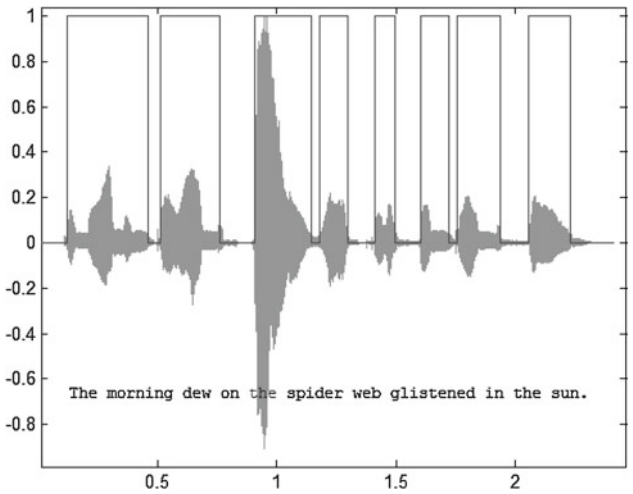

Fig. 11 An example of test set and its handmade label for evaluation of VAD result

results are related to the "volvo" and "white" noises. This is because these noises are rather stationary and uncorrelated. The results of "babble" and "factory" noises are weak at low SNRs. The "babble" noise is correlated and contains LF information, which makes it very hard to discriminate from speech signal at low SNRs. In addition, the non-stationary property of "factory" noise makes it hard to construct a good SVM model for it at low SNRs.

It is also interesting to discuss about the proposed VAD behavior in situations of noise types not observed in training set, which is a critical point in the algorithm performance. For better exploration of proposed VAD behavior in unseen noise type, Fig. 13 demonstrates the proposed VAD results for average SNRs (0–15 dB) in different train/test situations. It can be seen from the results that when SVM model is wrongly chosen for white and volvo noises, the performance of proposed VAD algorithm has highly affected. In other words, the performance of proposed algorithm in these cases is highly related to the noise type. However, it can be seen from Table 1 that the classification accuracy for noise classification algorithm is 100 % for both volvo and white noises.

The performance of the proposed VAD is compared with the state-of-the-art methods such as Sohn's, order-statistics filters (OSF), and long-term spectral divergence (LTSD) VADs. Sohn et al. [20] applied Markov model on a statistical likelihood ratio test to build a robust voice activity detector. The developed VAD employs the decision-directed parameter estimation method for the likelihood ratio test. OSF proposed in [6] is based on the determination of the speech/non-speech divergence by means of specialized OSFs working on the sub-band log-energies. The LTSD VAD measures the LTSD between speech and noise and formulates the speech/non-speech decision rule by comparing the long-term spectral envelope to the average noise spectrum [21]. The decision threshold is adapted to the measured noise energy while a controlled hangover is activated only when the observed SNR is low.

An additional test was conducted to compare speech detection performance by means of the receiver operating

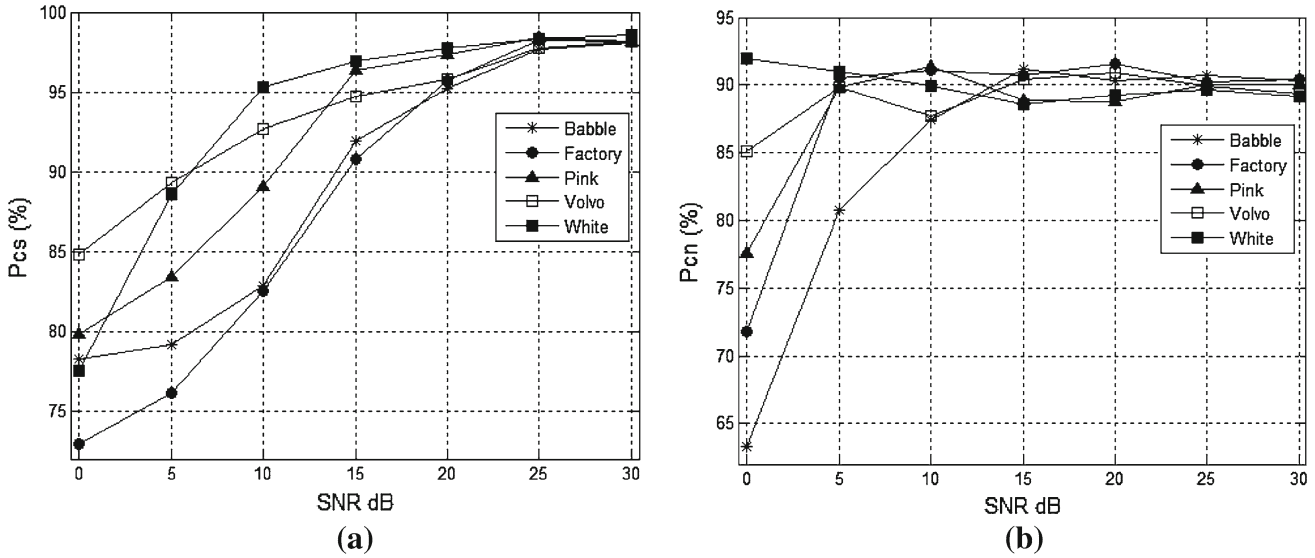

Fig. 12 Speech/non-speech discrimination analysis as a function of the SNR for different noise types. **a** Percentage of correct speech identification and **b** percentage of correct non-speech identification

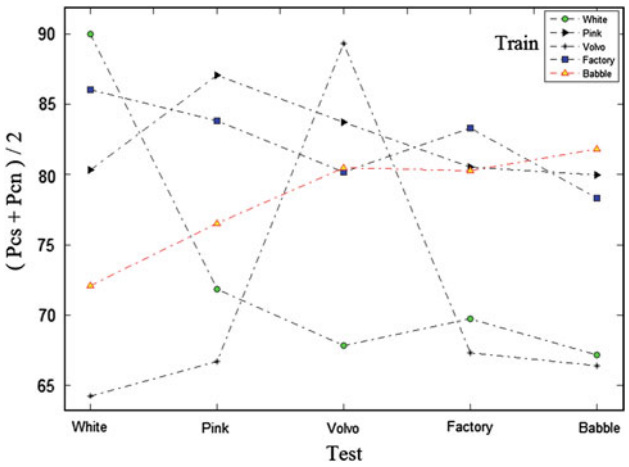

Fig. 13 Performance of the proposed VAD algorithm for average SNRs (0–15 dB) in different train/test situations

characteristic (ROC) curves [22], a frequently used methodology in communications based on the hit and error detection probabilities, which completely describes the VAD error rate. The Pcn or the pause hit rate and the false alarm rate (1-Pcs) were determined in each noise condition for different VADs in Fig. 14. From the results, it is clear that the proposed VAD outperforms the Sohn's, OSF, and LTSD methods in almost all cases. Among all the VADs examined, our VAD yields the lowest false alarm rate for a fixed non-speech hit rate and also the highest non-speech hit rate for a given false alarm rate.

Most speech enhancement algorithms make use of the VAD module in order to estimate the statistics of noise. Therefore, the effectiveness of the noise compensation algorithms is strongly affected by the accuracy of the VAD. The results show that the proposed VAD can be used to let us estimate statistics of noise for speech enhancement algorithms, since the Pcn is acceptable under different noise conditions. However, the Pcn results for the Sohn's, OSF, and LTSD VAD methods are low, especially at low SNRs, and therefore, these methods are not appropriate for noise estimation and speech enhancement. It should be mentioned that the implementation codes of the proposed algorithm are available at Matlab central file exchange [23].

5 Conclusions

In this paper, we have tried to provide a simple model for VAD based on a noise classification as the first step of the algorithm. We have also proposed a new robust feature vector based on the PWPT for both noise and speech/non-speech classification.

The experimental results for noise classification have been very promising. We have reached 98.4 % classification accuracy to classify five noise types extracted from NOISEX-92. Experimental results for VAD show that the performance of the proposed algorithm is superior to the Sohn's, OSF, and LTSD VADs, especially in low SNRs. The proposed algorithm has also reached to 86.14 % Pcs and 86.44 % Pcn in five noise environments and four SNR levels (0, 5, 10, and 15 dB) on average.

One aspect that we would like to explore in the future is to consider speech phase information in the feature extraction

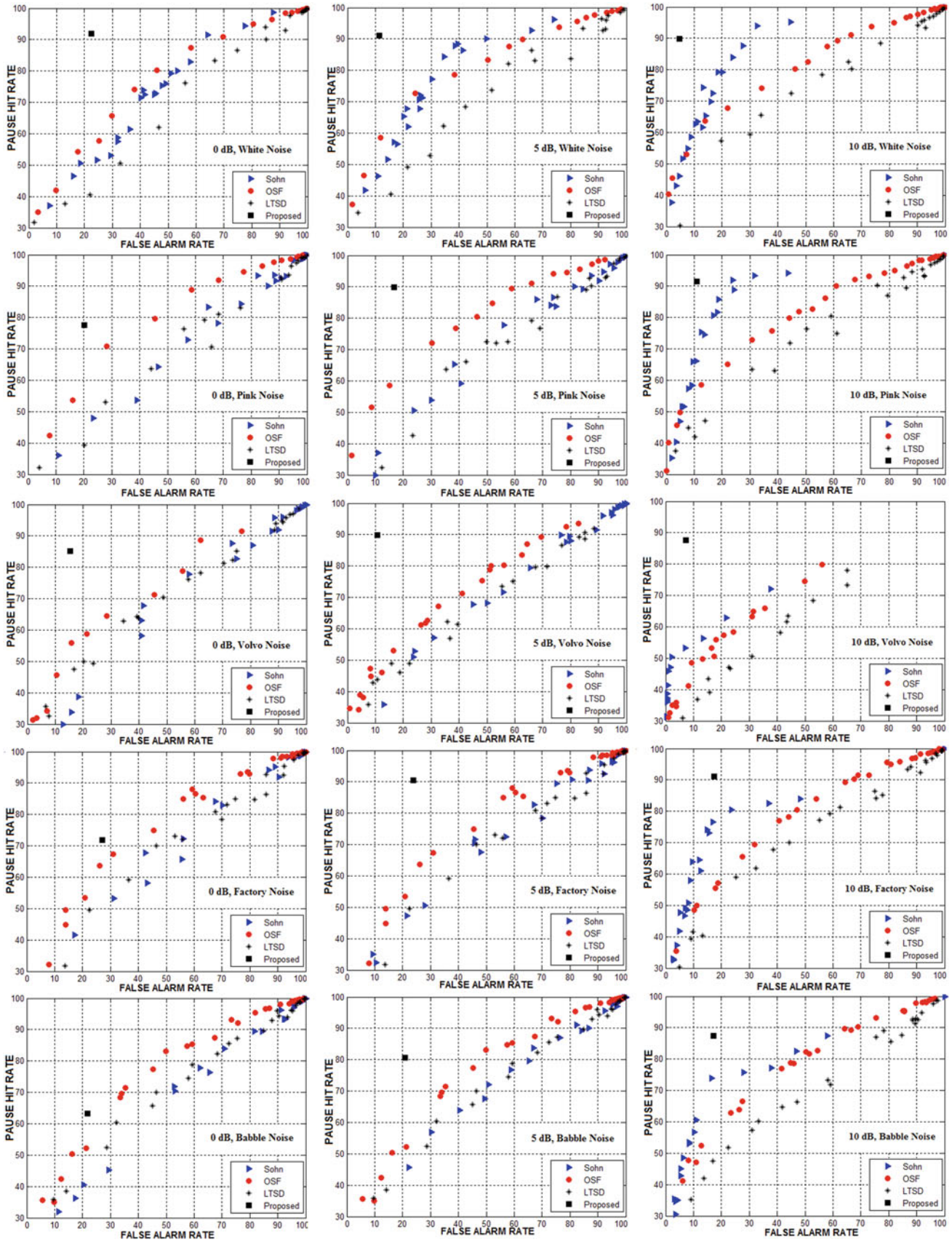

Fig. 14 ROC curves for different VADs and different noise types

process. We are also considering improving the VAD performance by using other classification algorithms. Taking into account more noise types in the proposed VAD can improve the performance in real-world applications. Future work should be done on these promising issues.

References

1. Beritelli, F., Casale, S., Ruggeri, G.: Performance evaluation and comparison of ITU-T/ETSI voice activity detectors. In: Proceedings ICASSP, pp. 1425–1428 (2001)
2. Srinivasant, K., Gersho, A.: Voice activity detection for cellular networks. In: Proceedings IEEE Speech Coding, Workshop, pp. 85–86 (1993)
3. Karray, L., Martin, A.: Towards improving speech detection robustness for speech recognition in adverse environment. Speech Commun. **40**, 261–276 (2003)
4. Woo, K.H., Yang, T.Y., Park, K.J., Lee, C.: Robust voice activity detection algorithm for estimating noise spectrum. IEE Electron. Lett. **36**(2), 180–181 (2000)
5. Chen, S.H., Wu, H.T., Chang, Y., Truong, T.K.: Robust voice activity detection using perceptual wavelet-packet transform and Teager energy operator. Pattern Recognit. Lett. **28**(11), 1327–1332 (2007)
6. Ramírez, J., Segura, J.C., Benítez, M.C., Torre, Á.D., Rubio, A.J.: An effective subband OSF-based VAD with noise reduction for robust speech recognition. IEEE Trans. Speech Audio Process. **13**(6), 1119–1129 (2005)
7. Wu, B.F., Wang, K.C.: Voice activity detection based on autocorrelation function using wavelet transform and Teager energy operator. Comput. Linguist. Chin. Lang. Process. **11**(1), 87–100 (2006)
8. Thatphithakkul, N., Kruatrachue, B., Wutiwiwatchai, C., Marukatat, S.: Robust speech recognition using PCA-Based noise classification. In: SPECCOM, pp. 45–53 (2005)
9. Mohammadi, M., Zamani, B., Nasersharif, B., Rahmani, M., Akbari, A.: A wavelet based speech enhancement method using noise classification and shaping. In: INTERSPEECH, pp. 561–564 (2008)
10. Coifman, R.R., Wickerhauser, M.V.: Entropy-based algorithms for best basis selection. IEEE Trans. Inf. Theory **38**(2), 713–718 (1992)
11. Rabiner, L., Juang, B.H.: Fundamental of Speech Recognition. Prentice-Hall, Upper Saddle River (1993)
12. Zwicker, E., Terhardt, E.: Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. J. Acoust. Soc. Am. **68**, 1523–1525 (1980)
13. Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D.: The NOISEX-92 study on the effect of additive noise on automatic speech recognition, http://spib.rice.edu/spib/select (1992)
14. Friedman, J.H.: Another Approach to Polychotomous Classification. Technical Report. Department of Statistics, Stanford University, pp. 1–14 (1996)
15. Chang, C., Lin, C.J.: LIBSVM: A Library for support Vector Machines, Technical Report. Department of Computer Science and Information Engineering, National Taiwan University (2001)
16. Welch, P.D.: The use of fast Fourier transform for the estimation of power spectra: a method based on time averaging over short, modified periodograms. IEEE Trans. Audio Electroacoust. **15**, 70–73 (1967)
17. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L., Zue, V.: TIMIT Acoustic- Phonetic—Continuous Speech Corpus. Technical Report. National Institute of Standards and Technology (1993)
18. Ramýrez, J., Yélamos, P., Górriz, J., Segura, J., Garcýa, L.: Speech/non-speech discrimination combining advanced feature extraction and SVM learning. In: Ninth International Conference on Spoken Language Processing, pp. 1662–1665 (2006)
19. Kinnunen, T., Chernenko, E., Tuononen, M., Frnti, P., Li, H.: Voice activity detection using MFCC features and support vector machine. Int. Conf. Speech Comput. **2**, 556–561 (2007)
20. Sohn, J., Kim, N.S., Sung, W.: A statistical model based voice activity detection. IEEE Signal Process. Lett. **6**(1), 1–3 (1999)
21. Ramirez, J., Segura, J., Benitez, C., De La Torre, A., Rubio, A.: Efficient voice activity detection algorithms using long-term speech information. Speech Commun. **42**(3–4), 271–287 (2004)
22. Madisetti, V., Williams, D.B.: Digital Signal Processing Handbook. CRC/IEEE Press, Boca Raton (1999)
23. http://www.mathworks.com/matlabcentral/fileexchange/39343