Applied Acoustics 96 (2015) 11–19

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A speech enhancement approach based on noise classification

Wenhao Yuan [*], Bin Xia

College of Computer Science and Technology, Shandong University of Technology, Shandong 255049, China

ARTICLE INFO

Article history:
Received 25 August 2014
Received in revised form 16 December 2014
Accepted 11 March 2015
Available online 25 March 2015

Keywords:
Speech enhancement
Noise classification
Improved minima controlled recursive averaging

ABSTRACT

For speech enhancement, most existing approaches do not consider the differences, between various types of noise, which significantly affect the performance of speech enhancement. In this paper, we propose a novel speech enhancement approach by taking into account the different characteristic statistical properties of various noise on the basis of noise classification. To classify noise, an effective noise classification method is firstly developed by exploiting the features of noise energy distribution in the Bark domain. Then, based on the noise types, the speech enhancement approach is obtained by forming the optimal parameter combinations for the optimally modified log-spectral amplitude (OM-LSA) speech estimator with the improved minima controlled recursive averaging (IMCRA) noise estimator, where the parameter combinations consisting of the smoothing parameters for smoothing the noisy power spectrum and the recursive averaging in the noise spectrum estimation as well as the weighting factor for the *a priori* SNR estimation, are built through the enhancement of noisy speech samples. Finally, extensive experiments are carried out in terms of objective evaluation under various noise conditions, and the experimental results show that the proposed approach yields better performance compared with the conventional OM-LSA with IMCRA in speech enhancement.

1. Introduction

Single channel speech enhancement has been one of the most widely used approaches for the enhancement of noisy speech which is a crucial component of speech signal processing in noisy environments [1–6]. The spectral subtraction method proposed by Boll in [7] is a popular single channel speech enhancement technique, which substantially reduces the noise level in the noisy speech. According to the basic principle of spectral subtraction method, two major components generally should be considered in a practical speech enhancement system: the estimation of speech, and the estimation of noise power spectrum [8,9].

As for estimating the speech, a commonly used approach is the minimum mean-square error (MMSE) short-time spectral amplitude (STSA) estimator, which is derived by Ephraim and Malah in [10]. By utilizing decision directed approach to smooth the *a priori* SNR recursively, the MMSE estimator successfully conquers the main drawback of the conventional spectral subtraction method that it may introduce an annoying distortion called musical noise into the enhanced speech. Subsequently, Ephraim and Malah derived a MMSE log-spectral amplitude (LSA) estimator in literature [11] which minimizes the mean-square error of the log-spectra. Further, by modifying the gain function of the LSA estimator based on two hypotheses associated with the speech presence uncertainty, Cohen presented an optimally modified LSA (OM-LSA) speech estimator [8], which shows significant superiority in speech enhancement.

Considering the noise power spectrum estimation, Martin proposed a noise PSD estimation algorithm based on minimum statistics (MS) [12], which tracks the minima values of the smoothed spectrum of the noisy speech over a finite window, and then multiplies the result by a bias factor to achieve the unbiased estimate of noise spectrum. Another successful noise PSD estimation approach, known as the minima controlled recursive averaging (MCRA) algorithm [13], is to search the local minimum similarly to MS, and then compare the ratio of the noisy speech to the local minimum against a threshold to find the noise-only regions. The noise PSD estimate is updated by tracking the noise-only regions of the noisy speech spectrum. In [14], Cohen presented the improved MCRA (IMCRA), which uses a different method to track the noise-only regions based on the estimated speech presence probability. In [15], Rangachari and Loizou updated the noise PSD estimate continuously in every frame using the speech presence probability which was obtained by comparing the ratio of noisy speech power spectrum to its local minimum against a frequency-dependent threshold. The more recent work on noise PSD estimation is the MMSE-based algorithm with bias

[*] Corresponding author.
E-mail addresses: why_sdut@126.com (W. Yuan), xiabin@sdut.edu.cn (B. Xia).

http://dx.doi.org/10.1016/j.apacoust.2015.03.005
0003-682X/© 2015 Elsevier Ltd. All rights reserved.

compensation (MMSE-BC) proposed by Hendriks and Heusdens [16–18], which employs a limited maximum likelihood estimate of the *a priori* SNR to obtain an MMSE estimate of the noise periodogram. The MMSE-BC estimator reduces the computational complexity greatly without degrading the performance of noise tracking. In [19], Gerkmann and Hendriks further improved the MMSE-BC estimator by making use of a soft speech presence probability with fixed priors, and presented an unbiased MMSE-based noise PSD estimator, which is of an even lower complexity than the MMSE-BC.

To distinguish speech from noise, both the estimation of speech and the estimation of noise power spectrum have fully considered the differences between speech and noise. For example, the estimation of noise PSD is generally based on the assumption that the noise power is varying more slowly than the speech power. Besides, the differences are also made between different types of noise in the characteristic statistical properties. However, the design of the speech enhancement algorithms usually does not take the differences into consideration, which causes these algorithms to be not always optimal for various noise environments. As to improve the performance of speech enhancement, these algorithms should be adjusted to adapt to different types of noise and deal with them respectively. In other words, we can improve the performance of these speech enhancement algorithms by incorporating the noise classification into them.

For noise classification, a variety of features have been proposed, including time domain features [20], spectral domain features [21] and the features derived from linear predictive coding (LPC) and wavelet transforms [22,23], of which the mel-frequency cepstral coefficients (MFCC) features are most widely used. Much work has been done to recognize the nonspeech audio based on the MFCC features. Ma et al. described an acoustic environment classifier using a 39-dimensional MFCC feature vector [24]. In [25], to yield higher recognition accuracy for environmental sounds, the matching pursuit algorithm is used to obtain effective time–frequency features as the supplement of the MFCC features. Gopalakrishna et al. utilized the MFCC + Δ MFCC features to classify the background noise environment in real time for automatic tuning of noise suppression algorithms for cochlear implant applications [26]. The classification accuracy for the above studies varied from 80% to 95% under different databases, which implies that there is still work to do to extract more effective features for noise classification.

In this paper, on the basis of the speech enhancement scheme based on the IMCRA noise PSD estimator and the OM-LSA speech estimator, we propose a speech enhancement approach using noise classification of noisy speech. Firstly, we define a parameter combination related to the noise types, which includes some principal parameters in the OM-LSA with IMCRA, such as the smoothing parameters for the smoothing of the noisy power spectrum and the recursive averaging in the noise spectrum estimation, as well as the weighting factor for the *a priori* SNR estimation. Through the enhancement of noisy speech samples, by identifying the optimal parameter combinations for the speech enhancement scheme based on the IMCRA and the OM-LSA under specific noise environments, we obtain the optimal parametric OM-LSA with IMCRA. Secondly, to recognize the noise type of the noisy speech, we propose a support vector machine-based noise classification method, which exploits the features of noise energy distribution in the Bark domain. Thirdly, by choosing the the optimal parameter combination for the speech enhancement scheme based on the OM-LSA and the IMCRA according to the recognized noise type, we implement the noise PSD estimation and calculate the enhanced speech using the optimal parametric OM-LSA with IMCRA. Objective quality tests are performed to evaluate the proposed approach under various noise environments, which validate the superior performance of the proposed approach to the conventional speech enhancement scheme based on the OM-LSA and the IMCRA.

The rest of the paper is organized as follows. Section 2 briefly reviews the speech enhancement scheme based on the IMCRA and the OM-LSA, and Section 3 presents the optimal parametric OM-LSA with IMCRA for various noise. Section 4 introduces the support vector machine-based noise classification method. In Section 5, we describe the proposed noise classification-based speech enhancement approach. The performance of the proposed noise classification method and speech enhancement approach is evaluated in Section 6. Finally, conclusions are given in Section 7.

2. Review of OM-LSA with IMCRA

Let y denote an observed noisy signal in the time domain, which is the sum of a clean speech x and an uncorrelated additive noise d. By applying the short-time Fourier transform (STFT), we have

$$Y(k,l) = X(k,l) + D(k,l) \tag{1}$$

in the time–frequency domain, where k represents the frequency bin index, and l is the frame index.

In the IMCRA, the noise PSD is estimated by recursively averaging past spectral power values of the noisy measurement during periods of speech absence and holding the estimate during speech presence [14]. Under speech presence uncertainty, the conditional speech presence probability is employed, and the recursive averaging can be obtained by

$$\bar{\lambda}_d(k,l+1) = \tilde{\alpha}_d(k,l)\bar{\lambda}_d(k,l) + [1 - \tilde{\alpha}_d(k,l)]|Y(k,l)|^2 \tag{2}$$

where

$$\tilde{\alpha}_d(k,l) \triangleq \alpha_d + (1-\alpha_d)p(k,l) \tag{3}$$

is a time-varying frequency-dependent smoothing parameter. $\alpha_d(0 < \alpha_d < 1)$ denotes a smoothing parameter, and $p(k,l)$ is the conditional speech presence probability. Through introducing a bias compensation factor β, the noise PSD estimate is given by

$$\hat{\lambda}_d(k,l+1) = \beta \cdot \bar{\lambda}_d(k,l+1) \tag{4}$$

The estimation of the speech presence probability is based on a Gaussian statistical model in the IMCRA, and is obtained by

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}(1+\xi(k,l))\exp(-\upsilon(k,l))\right\}^{-1} \tag{5}$$

where $q(k,l)$ is the *a priori* probability for speech absence, γ and ξ represent the *a posteriori* and the *a priori* SNRs respectively, and $\upsilon \triangleq \gamma\xi/(1+\xi)$.

In order to calculate the *a priori* speech absence probability $q(k,l)$, two iterations of smoothing and minimum tracking are carried out. Let $S(k,l)$ denote the smoothed periodogram of the noisy measurement, then the time smoothing in the first iteration is performed by a first-order recursive averaging

$$S(k,l) = \alpha_s S(k,l-1) + (1-\alpha_s)S_f(k,l) \tag{6}$$

where $\alpha_s(0 < \alpha_s < 1)$ is a smoothing parameter, and $S_f(k,l)$ is obtained by the frequency smoothing of the noisy power spectrum

$$S_f(k,l) = \sum_{i=-\omega}^{\omega} b(i)|Y(k-i,l)|^2 \tag{7}$$

where b denotes a normalized window function of length $2\omega + 1$. The time smoothing in the second iteration is similar to that in the first iteration, and utilizes the same smoothing parameter.

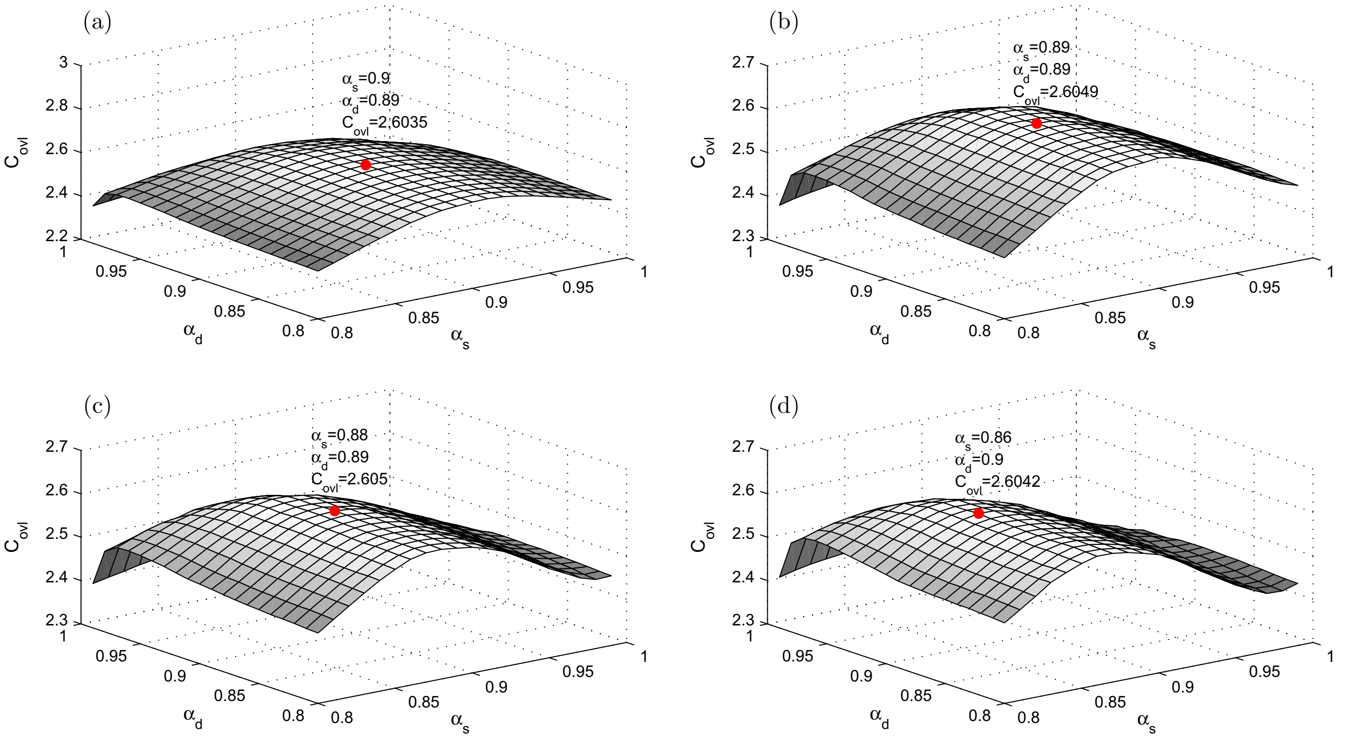

Fig. 1. The average C_{ovl} for various values of α_s and α_d according to different values of α: (a) $\alpha = 0.88$; (b) $\alpha = 0.89$; (c) $\alpha = 0.90$; (d) $\alpha = 0.91$.

The *a priori* speech absence probability is controlled by the minima values of $S(k,l)$, which is searched within a finite window of length D, for each frequency bin

$$S_{\min}(k,l) \triangleq \min\left\{S(k,l')\middle| l-D+1 \leqslant l' \leqslant l\right\} \qquad (8)$$

According to Eq. (5), the computation of the speech presence probability $p(k,l)$ also requires an estimate of the *a priori* SNR. In the IMCRA, the *a priori* SNR is commonly estimated by

$$\hat{\xi}(k,l) = \alpha G_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-\alpha)\max\{\gamma(k,l)-1,0\} \qquad (9)$$

where α is a weighting factor, G_{H_1} is the spectral gain function in the case that speech is present, and $\gamma(k,l) \triangleq |Y(k,l)|^2/\lambda_d(k,l)$.

Constrained to a lower bound threshold G_{min} when speech is absent, the spectral gain for the OM-LSA is given by

$$G(k,l) = \{G_{H_1}(k,l)\}^{p(k,l)} G_{\min}^{1-p(k,l)} \qquad (10)$$

3. Optimal parametric OM-LSA with IMCRA

The OM-LSA with IMCRA estimates the noise spectrum and speech frame by frame, and takes into account the strong correlation of speech activities in the adjacent frames by carrying out the smoothing of noisy power spectrum (Eq. (6)), the recursive averaging in the noise spectrum estimation (Eq. (2)) and the non-linear recursive procedure in the *a priori* SNR estimation (Eq. (9)). Three parameters α_s, α_d and α are set to control the tradeoff between the current frame and the previous frame in Eqs. (6), (2) and (9), respectively. The choices of these parameters deeply affect the estimation of the noise spectrum and the *a priori* SNR, and further have a large impact on the performance of the OM-LSA speech estimator, which means that these parameters are closely related to the speech enhancement performance of the OM-LSA with IMCRA. These parameters are set to fundamental values independently of the noise environments in the conventional OM-LSA with IMCRA. Actually, due to the different statistical properties of various noise, they have different interference with the speech signals in the noisy speech, therefore the correlation of speech presence between the current frame and the previous frame is different according to the noise types. Thus, these parameters should vary in terms of noise types to ensure the tracking of speech presence.

The optimal parameter combination is a combination of the aforementioned parameters, with which the OM-LSA with IMCRA can achieve the most accurate estimate of speech for specific noise. The accuracy of the speech estimate is reflected in the quality of the enhanced speech, which is measured in this paper using the well-known composite measure in [27]. We choose the composite measure for overall quality from the three composite measures, which is verified to have significant correlation with the subjective speech quality [27,28], and is given as

$$C_{ovl} = 1.594 + 0.8055S_{PESQ} - 0.512S_{LLR} - 0.007S_{WSS} \qquad (11)$$

where S_{PESQ}, S_{LLR} and S_{WSS} represent the measurements according to the perceptual evaluation of speech quality (PESQ), the log-likelihood ratio (LLR), and the weighted-slope spectral distance (WSS), respectively.

We seek the optimal parameter combinations through the enhancement of noisy speech samples. 20 clean speech segments chosen from IEEE sentence database [29] are utilized to create the samples. Half of them are from two male speakers, denoted by ‘sp01’, ‘sp02’, ..., ‘sp10’, while the others are from two female speakers, denoted by ‘sp11’, ‘sp12’, ..., ‘sp20’. The noise signals are taken from the Noisex92 database [30], including: N1, white noise; N2, F-16 cockpit noise; N3, HF channel noise; N4, factory floor noise 1; N5, speech babble; N6, Pink noise; N7, car interior noise; N8, destroyer operations room noise; N9, destroyer engine room noise; N10, jet cockpit noise 1; N11, tank noise; N12, military vehicle noise. Both the speech and the noise signals are sampled at 8 kHz. Applying all 12 types of noise to the 20 segments of clean speech with 0, 5 and 10 dB global SNRs, we obtain 20 × 12 × 3 segments of noisy speech.

The optimal parameters are searched within the range $[0.80, 0.99]$, and the searching step is 0.01. For each noise type, we first set α to be a fixed value, and let α_s and α_d vary within the range to produce different combinations, denoted as $\{(\alpha_s^i, \alpha_d^i)\}_{i=1}^{N_c}$. The 20×3 segments of noisy speech corrupted by the same noise with different SNRs are processed using the OM-LSA with IMCRA with these different combinations respectively. The average C_{ovl} of the 20×3 segments of enhanced speech is calculated for each combination, and the combination which achieves the maximal average is indicated. And then, with the variation of α, there will exist different combinations that maximize the average according to different values of α, where the one achieves the largest average is denoted as (α_s^*, α_d^*), and its corresponding α is denoted as α^*. Thus, $(\alpha_s^*, \alpha_d^*, \alpha^*)$ is the optimal parameter combination, with which the global maximal average C_{ovl} is obtained.

Fig. 1 is an example to show the procedure of deriving the optimal parameter combination for white noise. Fig. 1(a)–(d) show the average C_{ovl} for various values of α_s and α_d with α equaling 0.88, 0.89, 0.90 and 0.91, respectively. It can be observed that the maximal average for each value of α is obtained at $(0.90, 0.89)$ in Fig. 1(a), $(0.89, 0.89)$ in Fig. 1(b), $(0.88, 0.89)$ in Fig. 1(c) and $(0.86, 0.90)$ in Fig. 1(d). Among them, the global maximal average is obtained at $(0.88, 0.89)$ in Fig. 1(c) with α equaling 0.90. Therefore, the optimal parameter combination for white noise is $(0.88, 0.89, 0.90)$. Further, the optimal parameter combinations for all the given noise types are shown in Table 1. By choosing the optimal parameter combination for the OM-LSA with IMCRA according to the particular noise type, we achieve the optimal parametric OM-LSA with IMCRA.

4. Noise classification

Before applying the optimal parametric OM-LSA with IMCRA to speech enhancement, we should judge the noise type first. In order to classify the noise accurately, the features that effectively differentiate varieties of noise should be selected. In this paper, we exploit the features of noise energy distribution in the Bark domain and model them using the support vector machine (SVM) to achieve a successful noise classification method.

4.1. Feature extraction

Unlike the Fourier transformation, the Bark scale adopts the multi-resolution analysis and divides the time–frequency space non-equally. The Bark bands are narrower at the region of low frequencies than those at the region of high frequencies. By mapping the noise energy from the uniform time–frequency domain to the Bark domain, we acquire a feature vector that can effectively distinguish different types of noise. Let $D(k, l)$ denote the short-time Fourier transformation of a noise signal, and the smoothing of the noise power spectrum in time is performed as

$$P(k, l) = \alpha_p P(k, l-1) + (1 - \alpha_p)|D(k, l)|^2 \tag{12}$$

Table 1
Optimal parameter combinations for various noise.

Noise	α_s^*	α_d^*	α^*	Noise	α_s^*	α_d^*	α^*
white	0.88	0.89	0.90	car interior	0.93	0.89	0.93
F-16	0.93	0.86	0.88	destroyer operations	0.90	0.91	0.91
HF channel	0.91	0.88	0.89	destroyer engine	0.93	0.81	0.88
factory1	0.95	0.84	0.90	jet1	0.93	0.90	0.93
babble	0.97	0.82	0.95	tank	0.93	0.88	0.89
pink	0.89	0.90	0.90	military vehicle	0.94	0.81	0.93

Table 2
Mapping from 256-point STFT bins to Bark bands at a sampling frequency of 8 kHz.

Bark Band Number (j)	STFT Bins		Frequencies (Hz)
	Interval ($lb(j)$-$ub(j)$)	Number of bins	
1	1–3	3	0–94
2	4–6	3	94–187
3	7–10	4	187–312
4	11–13	3	312–406
5	14–16	3	406–500
6	17–20	4	500–625
7	21–25	5	625–781
8	26–29	4	781–906
9	30–35	6	906–1094
10	36–41	6	1094–1281
11	42–47	6	1281–1469
12	48–55	8	1469–1719
13	56–64	9	1719–2000
14	65–74	10	2000–2312
15	75–86	12	2312–2687
16	87–100	14	2687–3125
17	101–118	18	3125–3687
18	119–128	10	3687–4000

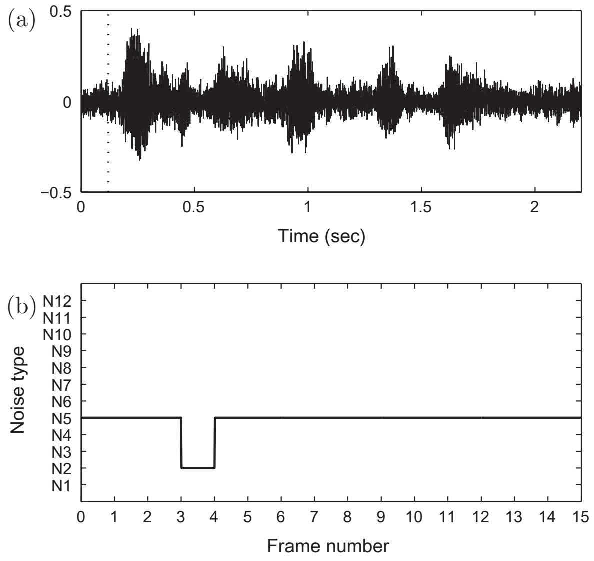

Fig. 2. Noise classification of a segment of speech corrupted by babble noise (N5) at 0 dB SNR. (a) Noisy speech waveform. (b) Classification results of the first 15 frames.

where $P(k, l)$ is the smoothed noise power spectrum, and $\alpha_p = 0.5$ is the smoothing parameter.

The mapping from 256-point STFT bins to Bark bands at a sampling frequency of 8 kHz is shown in Table 2, and it is the same as that in [31]. The noise energy in each Bark band is calculated according to its corresponding upper STFT bin and lower STFT bin in Table 2,

$$\mathbb{S}(j, l) = \sum_{k=lb(j)}^{ub(j)} P(k, l) \tag{13}$$

where j is the Bark band number, and $j = 1, 2, \cdots, 18$. The total noise energy of the lth frame is

$$\mathbb{S}_t(l) = \sum_{k=1}^{N} P(k, l) \tag{14}$$

where $N = 128$ is the total number of STFT bins. Then the ratio of the noise energy in the jth Bark band of the lth frame to the entire noise energy in the lth frame is defined by

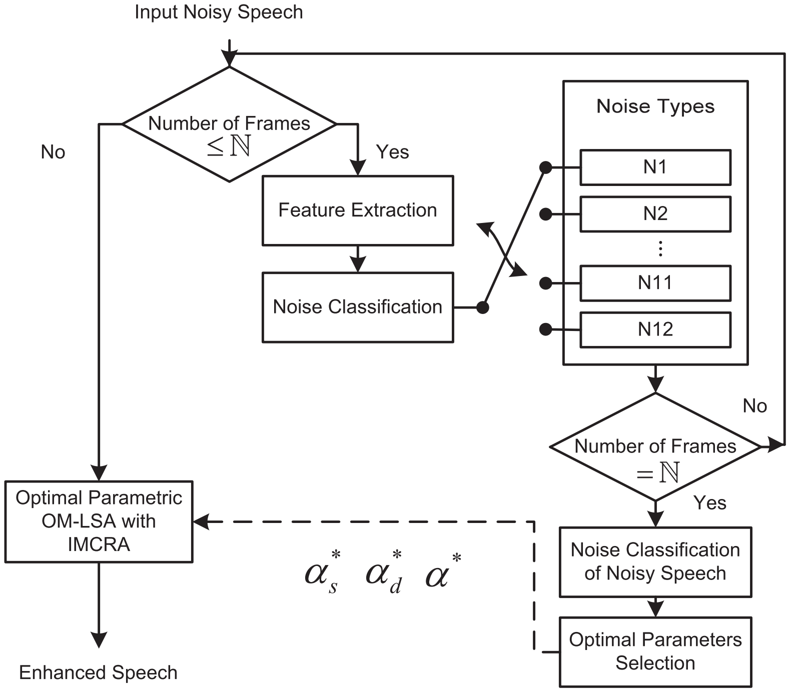

Fig. 3. Block diagram of the proposed speech enhancement approach.

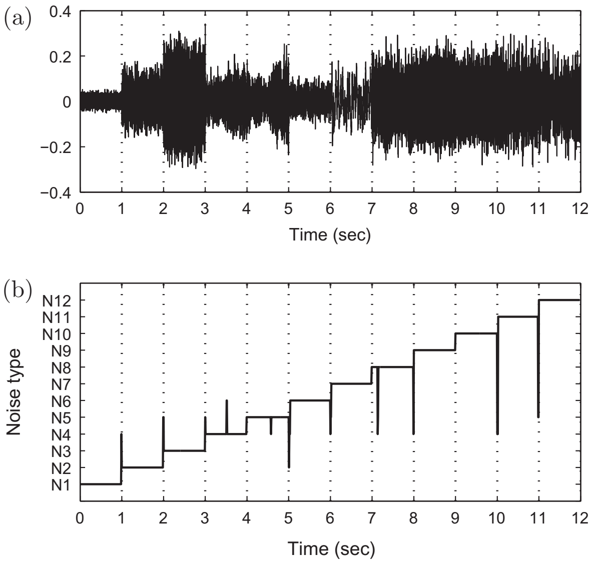

Fig. 4. Noise classification of a 12-s noise segment containing 12 types of noise. (a) Waveform of the noise segment. (b) Classification results of the noise segment.

$$R_{\mathbb{S}}(j,l) = \frac{\mathbb{S}(j,l)}{\mathbb{S}_t(l)} \tag{15}$$

Subsequently, to obtain a more effective feature vector for classification, we introduce a power η for the ratio $R_{\mathbb{S}}(j,l)$, and denote its exponential form as $R_{\mathbb{S}}^{\eta}(j,l)$. Then we acquire a 18-dimensional feature vector for the lth frame, which is given by

$$\mathbf{r} = (R_{\mathbb{S}}^{\eta}(1,l), R_{\mathbb{S}}^{\eta}(2,l), \ldots, R_{\mathbb{S}}^{\eta}(18,l)) \tag{16}$$

As shown in Eq. (16), different values of η will lead to different expressions of the feature vector, according to the performance of the feature vectors in the classification experiments, we decide the value of η to be $1/4$ in this paper.

4.2. Training of SVM

The SVM classifier is used to determine the noise type of every frame. SVM is a successful technique for data classification derived from statistical learning theory [32]. The main idea of SVM is to transform the input data set into a higher dimensional feature space by using a kernel function, where it is easier to classify with linear decision surfaces. Note that the SVM is originally designed for the binary classification problem, to solve the multi-class classification problem like that in this paper (12 classes), many approaches have been proposed to reduce the single multi-class problem into multiple binary classification problems.

In this paper, the SVM models are trained using LIBSVM software [33], in which the "one-against-one" approach [34] is implemented for multi-class classification. For a problem with c classes, the "one-against-one" approach constructs a total number of $c(c-1)/2$ classifiers, and each of them trains data from two classes.

To obtain the training and testing data, all 12 types of noise (N1-N12) in Section 3 is windowed into 256-point frames via a hamming window with 75% overlap. And then we choose a

Table 3
Classification accuracy for the pure noise signals.

Noise	Accuracy (%)			
	MFCC13	MFCC26	MFCC39	BARK18
N1	84.94	94.58	98.60	100.00
N2	74.48	88.14	95.61	99.86
N3	81.54	92.93	98.06	100.00
N4	61.49	70.22	79.66	92.54
N5	97.77	96.83	96.61	96.87
N6	78.83	86.39	91.95	97.40
N7	78.82	91.26	97.63	100.00
N8	87.42	95.56	98.25	99.58
N9	98.71	99.66	99.86	100.00
N10	98.54	99.15	99.60	99.94
N11	92.07	98.94	99.88	99.99
N12	87.94	93.95	97.21	99.98

Table 4
Classification accuracy for noisy speech for various noise types and levels.

Noise	SNR (dB)			Noise	SNR (dB)		
	0 (%)	5 (%)	10 (%)		0 (%)	5 (%)	10 (%)
N1	100.00	100.00	100.00	N7	100.00	100.00	100.00
N2	100.00	100.00	100.00	N8	100.00	100.00	100.00
N3	100.00	100.00	100.00	N9	100.00	100.00	100.00
N4	100.00	100.00	100.00	N10	100.00	100.00	100.00
N5	100.00	100.00	100.00	N11	100.00	100.00	100.00
N6	100.00	100.00	100.00	N12	100.00	100.00	100.00

M-frame stretch of noise from each type of noise and repeat the steps for feature extraction to calculate the feature vectors for the M frames ($M = 15000$). Let m and n denote two classes chosen out of the given noise types, then the training data for class pair mn consisting of 18-dimensional patterns and the corresponding class labels z can be expressed as

$$\mathcal{D}^{mn} = \{(\mathbf{r}_i, z_i)|\mathbf{r}_i \in \mathbb{R}^{18}, z_i \in \{-1, +1\}\}_{i=1}^{2M} \tag{17}$$

The decision function for noise class pair mn is defined by

$$f_{mn}(\mathbf{r}) = \sum_{\mathbf{r}_i \in sv} \alpha_i^{mn} z_i K(\mathbf{r}_i, \mathbf{r}) + b^{mn} \tag{18}$$

where α_i^{mn} is from the solution of the quadratic programming problem, b^{mn} represents the optimized bias, and K denotes the kernel function [33], which is chosen to be the radial basis function (RBF) in our approach. Since $f_{mn}(\mathbf{r}) = -f_{nm}(\mathbf{r})$, there are $12(12-1)/2 = 66$ different decision functions in this 12-class problem.

For the classification of the "one-against-one" approach, the most popular method is the voting strategy: each binary classifier casts one vote for its preferred class, and the feature vector $\mathbf{r}$ is designated to be in a class with the most votes. Thus the noise type of the lth frame corresponding to $\mathbf{r}$ is given by

$$C_{frame} = \arg\max_{m=1,2,\cdots,12} \sum_{n \neq m, n=1}^{12} \text{sgn}(f_{mn}(\mathbf{r})) \tag{19}$$

4.3. Noise classification of noisy speech

The noise type is judged during noise-only periods in the noisy speech. The first $\mathbb{N} = 15$ frames in a segment of noisy speech are assumed to be non-speech, and the noise classification is carried out in these frames. Within the $\mathbb{N}$ frames, the number of frames judged as type n is denoted by $L(n)$, and the noise type of the noisy speech corresponds to the n maximizing $L(n)$, as following:

$$C_{segment} = \arg\max_{n=1,2,\ldots,12} L(n) \tag{20}$$

Fig. 2 gives an example of noise classification of noisy speech. Fig. 2(a) illustrates the waveform of a segment of speech corrupted by babble noise (N5) at 0 dB SNR, and the noise types of its first 15 frames are judged frame by frame as shown in Fig. 2(b). It is obvious that only the noise in the 4th frame is judged to be N2 by mistake. As calculated using Eq. (20), the noise type of the noisy speech segment is judged to be N5 accurately.

Table 5
Results of composite measure for signal distortion (C_{sig}) obtained from the unprocessed noisy speech, the OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach.

Noise	SNR (dB)	Method			
		Unprocessed	OM-LSA	MMSE-BC	Proposed
white	0	1.57	2.05	2.07	2.45
	5	2.10	2.80	2.72	3.03
	10	2.65	3.35	3.21	3.54
F-16	0	2.20	2.91	2.66	3.09
	5	2.72	3.43	3.29	3.55
	10	3.29	4.04	3.82	4.11
HF channel	0	2.25	2.58	2.55	2.89
	5	2.79	3.25	3.14	3.44
	10	3.32	3.79	3.67	3.94
factory1	0	2.20	2.41	2.17	2.47
	5	2.77	3.09	2.87	3.13
	10	3.32	3.79	3.50	3.82
babble	0	2.51	2.46	1.98	2.52
	5	3.07	3.13	2.70	3.14
	10	3.63	3.82	3.48	3.83
pink	0	1.96	2.63	2.45	2.90
	5	2.54	3.22	3.08	3.45
	10	3.10	3.77	3.63	3.95
car interior	0	3.92	4.99	4.80	5.02
	5	4.31	5.22	5.08	5.23
	10	4.72	5.43	5.30	5.43
destroyer operations	0	2.45	2.82	2.51	2.85
	5	3.00	3.40	3.15	3.46
	10	3.50	3.92	3.71	3.98
destroyer engine	0	2.19	2.80	2.86	3.00
	5	2.76	3.41	3.45	3.56
	10	3.25	3.91	3.95	4.04
jet1	0	2.02	2.59	2.03	2.62
	5	2.56	3.21	2.70	3.21
	10	3.12	3.70	3.34	3.73
tank	0	2.62	3.25	3.05	3.37
	5	3.15	3.84	3.61	3.92
	10	3.65	4.43	4.22	4.46
military vehicle	0	3.32	3.92	3.72	3.95
	5	3.78	4.34	4.19	4.37
	10	4.20	4.80	4.65	4.83
factory2	0	2.65	3.23	3.04	3.34
	5	3.14	3.76	3.52	3.82
	10	3.62	4.27	4.07	4.31
jet2	0	1.75	2.18	1.96	2.25
	5	2.28	2.86	2.61	2.96
	10	2.83	3.43	3.28	3.52

5. Speech enhancement based on noise classification

After the noise type is determined, we apply the optimal parameter combination to the OM-LSA with IMCRA according to the noise type by replacing the parameters α_s, α_d and α with the optimal parameters α_s^*, α_d^* and α^*. Firstly, the smoothing parameter α_s in the two iterations of smoothing of noisy power spectrum (Eq. (6)) is substituted with α_s^*. Secondly, the smoothing parameter in the recursive averaging for noise PSD estimation (Eq. (3)) is calculated by using α_d^* instead of α_d. Finally, by replacing the weighting factor α with α^* in the estimation of the *a priori* SNR (Eq. (9)), we obtain the optimal parametric OM-LSA with IMCRA, using which the noise is estimated and subsequently the enhanced speech is calculated. Summing up the above steps, a speech enhancement approach based on noise classification is achieved, whose procedure is summarized in Fig. 3.

6. Performance evaluation

The performance evaluation of the proposed noise classification-based speech enhancement approach consists of two parts: (1) the accuracy of noise classification, and (2) the objective quality of the speech enhanced using the proposed approach. In order to evaluate the performance, 10 segments of clean speech which

Table 6
Results of composite measure for background intrusiveness (C_{bak}) obtained from the unprocessed noisy speech, the OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach.

Noise	SNR (dB)	Method			
		Unprocessed	OM-LSA	MMSE-BC	Proposed
white	0	1.66	2.16	2.09	2.33
	5	2.04	2.64	2.57	2.74
	10	2.46	3.06	2.98	3.15
F-16	0	1.56	2.34	2.16	2.43
	5	1.99	2.76	2.66	2.80
	10	2.47	3.28	3.16	3.30
HF channel	0	1.76	2.20	2.09	2.33
	5	2.14	2.67	2.56	2.75
	10	2.56	3.11	3.02	3.16
factory1	0	1.61	2.07	1.95	2.10
	5	2.06	2.57	2.45	2.58
	10	2.51	3.10	2.95	3.12
babble	0	1.68	1.96	1.69	2.00
	5	2.16	2.44	2.22	2.46
	10	2.65	3.00	2.83	3.01
pink	0	1.56	2.27	2.13	2.39
	5	2.01	2.72	2.62	2.83
	10	2.47	3.17	3.10	3.26
car interior	0	2.41	4.13	4.01	4.13
	5	2.82	4.46	4.38	4.45
	10	3.29	4.80	4.72	4.80
destroyer operations	0	1.71	2.33	2.18	2.34
	5	2.16	2.76	2.66	2.79
	10	2.61	3.16	3.08	3.19
destroyer engine	0	1.61	2.24	2.28	2.33
	5	2.05	2.68	2.75	2.76
	10	2.47	3.10	3.23	3.18
jet1	0	1.58	2.18	1.77	2.19
	5	2.01	2.63	2.25	2.62
	10	2.46	3.05	2.77	3.07
tank	0	1.76	2.59	2.48	2.65
	5	2.21	3.06	2.93	3.09
	10	2.66	3.57	3.48	3.57
military vehicle	0	2.08	2.94	2.88	2.96
	5	2.50	3.30	3.28	3.30
	10	2.93	3.75	3.71	3.77
factory2	0	1.73	2.52	2.42	2.57
	5	2.17	2.99	2.85	3.01
	10	2.61	3.44	3.36	3.46
jet2	0	1.58	2.12	1.95	2.15
	5	2.00	2.58	2.42	2.62
	10	2.46	3.02	2.94	3.08

are different from those mentioned above, are taken from the IEEE sentences database [29]. They are half from a male speaker and half from a female speaker, and denoted by ‘sp21’, ‘sp22’, . . ., ‘sp30’. The aforementioned various noise signals are added to them in the same way as that in the previous section, and then another 10 × 12 × 3 segments of noisy speech are obtained. In addition, to assess the robustness of the proposed approach in the case of noise types out of the training set, we also evaluate the performance of speech enhancement under open noise types such as the factory floor noise 2 and jet cockpit noise 2, which are also from the Noisex92 database but not part of the training set.

6.1. Performance of noise classification

To evaluate the performance of the proposed noise classification method, we first investigate the classification accuracy for the pure noise signals. With the exception of the training data, the rest of the noise signals from the Noisex92 database are used as the testing data. Fig. 4 illustrates an example of the classification of the pure noise signals. By choosing a 1-s segment from each type of noise in the testing data and concatenating them successively, we obtain a 12-s noise segment as shown in Fig. 4(a). The noise classification of the 12-s noise segment is illustrated in Fig. 4(b), from which we can see that the proposed method results in high accuracy of classification for all 12 types of noise, and only little noise is classified mistakenly. The classification accuracy for the whole testing data in percentage is shown in Table 3. For comparison, the classification accuracy of the methods using the MFCC features and the SVM classifier is also given in Table 3, in which the 13, 26 and 39-dimensional MFCC features are computed using the HTK [26]. It is obvious that the proposed noise classification method (represented by BARK18) outperforms the methods using the 13, 26 and 39-dimensional MFCC features (represented by MFCC13, MFCC26 and MFCC39 respectively) for all types of noise with the average accuracy of 98.85%.

Also, we investigate the classification accuracy for the noisy speech. A total number of 30 × 12 × 3 segments of noisy speech

Table 7
Results of composite measure for overall quality (C_{ovl}) obtained from the unprocessed noisy speech, the OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach.

Noise	SNR (dB)	Method			
		Unprocessed	OM-LSA	MMSE-BC	Proposed
white	0	1.52	1.87	1.92	2.21
	5	1.93	2.51	2.48	2.70
	10	2.38	3.00	2.93	3.16
F-16	0	1.86	2.47	2.30	2.64
	5	2.29	2.94	2.87	3.05
	10	2.79	3.50	3.36	3.55
HF channel	0	1.88	2.16	2.16	2.43
	5	2.31	2.76	2.69	2.92
	10	2.75	3.26	3.18	3.37
factory1	0	1.84	2.10	1.94	2.15
	5	2.33	2.68	2.54	2.71
	10	2.81	3.29	3.08	3.32
babble	0	2.04	2.08	1.72	2.13
	5	2.55	2.66	2.35	2.68
	10	3.04	3.26	3.03	3.27
pink	0	1.69	2.27	2.17	2.51
	5	2.18	2.80	2.73	2.99
	10	2.68	3.29	3.23	3.45
car interior	0	3.42	4.42	4.27	4.43
	5	3.82	4.64	4.54	4.65
	10	4.23	4.84	4.74	4.85
destroyer operations	0	2.04	2.45	2.25	2.48
	5	2.53	2.97	2.82	3.01
	10	2.98	3.42	3.30	3.46
destroyer engine	0	1.87	2.34	2.46	2.52
	5	2.33	2.87	2.99	3.00
	10	2.75	3.34	3.45	3.46
jet1	0	1.71	2.22	1.79	2.24
	5	2.17	2.76	2.37	2.77
	10	2.64	3.21	2.93	3.24
tank	0	2.23	2.80	2.70	2.91
	5	2.71	3.33	3.20	3.39
	10	3.15	3.87	3.75	3.89
military vehicle	0	2.75	3.35	3.26	3.39
	5	3.17	3.76	3.69	3.78
	10	3.55	4.17	4.11	4.20
factory2	0	2.20	2.74	2.63	2.84
	5	2.64	3.24	3.08	3.29
	10	3.07	3.70	3.59	3.73
jet2	0	1.56	1.96	1.83	2.03
	5	2.02	2.55	2.40	2.64
	10	2.49	3.05	2.98	3.13

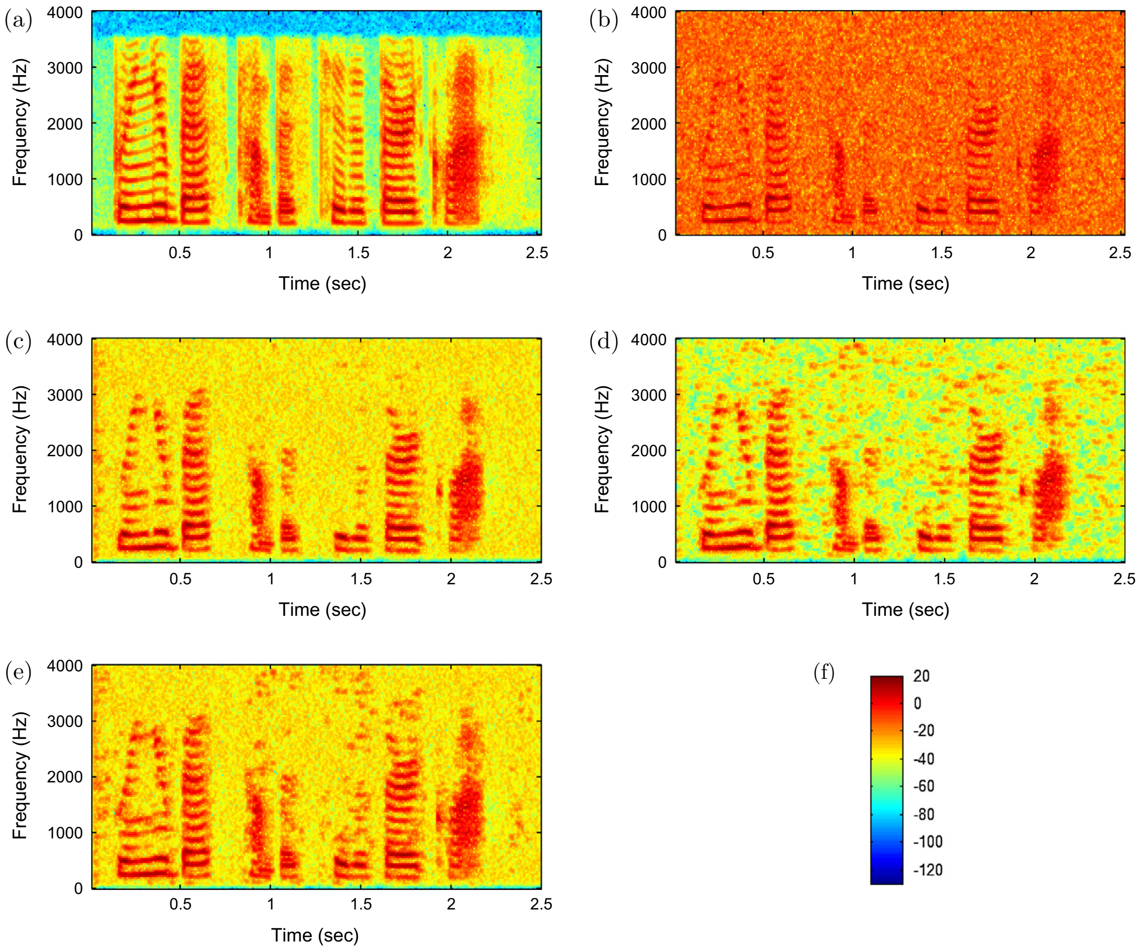

Fig. 5. Example of speech enhancement using the OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach: (a) Spectrogram of the clean speech. (b) Spectrogram of the unprocessed noisy speech (degraded by 5 dB white noise). (c) Spectrogram of the enhanced speech using the OM-LSA with IMCRA. (d) Spectrogram of the enhanced speech using the MMSE-BC with a super-Gaussian estimator. (e) Spectrogram of the enhanced speech using the proposed approach. (f) Colormap used in the above spectrograms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

are used to examine the accuracy of the noise classification, and the results in percentage for various noise types and levels are shown in Table 4. It can be seen that all the noisy speech segments are classified accurately for all the given noise conditions. Hence, the proposed method is suitable for the noise classification of noisy speech for speech enhancement.

6.2. Performance of speech enhancement

The speech enhancement performance of the proposed approach is compared with that of the conventional OM-LSA with IMCRA and the MMSE-BC with a super-Gaussian estimator from [35] using objective measures. Three composite measures given in [27] are used to evaluate the quality of the enhanced speech, including the signal distortion C_{sig}, background intrusiveness C_{bak} and the aforementioned overall quality C_{ovl}. The first two measures are defined as

$$C_{sig} = 3.093 - 1.029S_{LLR} + 0.603S_{PESQ} - 0.009S_{WSS} \tag{21}$$

$$C_{bak} = 1.634 + 0.4785S_{PESQ} - 0.007S_{WSS} + 0.063S_{segSNR} \tag{22}$$

where C_{sig} rates the enhanced speech on the speech signal alone using a five-point scale, C_{bak} rates the enhanced speech on the background noise alone using a five-point scale, and C_{ovl} rates the enhanced speech on the overall effect using the scale of the Mean Opinion Score (MOS) [27]. Also, S_{segSNR} represents the measurement according to the segmental SNR.

The noisy speech segments corresponding to the clean speech segments 'sp21' to 'sp30' are used as experimental data, and they are processed using the conventional OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach. The average C_{sig}, C_{bak} and C_{ovl} of the enhanced speech are calculated for each noise condition, the average measures of the unprocessed noisy speech are also calculated as comparison. Tables 5 and 6 present the results of the composite measures for signal distortion and background intrusiveness, respectively. It is obvious that the proposed approach outperforms the conventional OM-LSA with IMCRA and the MMSE-BC with a super-Gaussian estimator under all the tested conditions, which implies the proposed approach's superiority in preserving the speech component and suppressing the background noise. The results of the composite measure for overall quality are shown in Table 7, from which we can see that the proposed approach consistently yields a higher improvement in the speech quality than the conventional OM-LSA with IMCRA and the MMSE-BC with a super-Gaussian estimator. Also, it can be seen that the same conclusion can be made for two types of noise out of the training set, which implies the proposed speech enhancement approach is not dependent on the training set.

The comparison of speech enhancement performance among the conventional OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach for a segment of noisy speech is illustrated in Fig. 5. Fig. 5(a) shows the spectrogram of the clean speech, and the spectrogram of Fig. 5(b) presents the unprocessed noisy speech, which is degraded by 5 dB white noise. The spectrograms of the speech enhanced using the conventional OM-LSA with IMCRA, the MMSE-BC with a super-Gaussian estimator and the proposed approach are shown in Fig. 5(c), (d) and (e), respectively. As can be seen, the proposed approach has better noise suppression than the conventional OM-LSA with IMCRA, and preserves more speech spectra. The MMSE-BC with a super-Gaussian estimator obtains the best noise suppression, but it preserves less speech spectra than the proposed approach. Informal listening test confirms that the proposed approach achieves higher quality of enhanced speech.

7. Conclusions

In this paper, we have proposed a speech enhancement approach on the basis of noise classification, which comprises a SVM-based noise classification method and an optimal parametric OM-LSA speech estimator with IMCRA noise estimator. The noise classification method proposed in this paper exploits the features of noise energy distribution in the Bark domain and is implemented using the SVM classifiers. Through the enhancement of noisy speech samples, we obtain the optimal parameter combinations for the OM-LSA with IMCRA under various noise environments, using which we improve the conventional speech enhancement scheme based on the OM-LSA and the IMCRA to propose the optimal parametric OM-LSA with IMCRA for speech enhancement.

Performance evaluation is implemented to the proposed noise classification method and the noise classification-based speech enhancement approach. Experimental results of the noise classification show that the proposed method provides high classification accuracy to both pure noise and noisy speech. For speech enhancement, the results of the objective evaluation show that, compared to the conventional OM-LSA with IMCRA and the MMSE-BC with a super-Gaussian estimator, the proposed approach retains the speech component better, suppresses the background noise more effectively, and achieves the higher quality of enhanced speech.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No. 61271349).

References

[1] Ming J, Srinivasan R, Crookes D. A corpus-based approach to speech enhancement from nonstationary noise. IEEE Trans Audio Speech Lang Process 2011;19(4):822–36.
[2] Gunawan TS, Ambikairajah E, Epps J. Perceptual speech enhancement exploiting temporal masking properties of human auditory system. Speech Commun 2010;52(5):381–93.
[3] Paliwal K, Wojcicki K, Schwerin B. Single-channel speech enhancement using spectral subtraction in the short-time modulation domain. Speech Commun 2010;52(5):450–75.
[4] Lu C-T, Tseng K-F. A gain factor adapted by masking property and SNR variation for speech enhancement in colored-noise corruptions. Comput Speech Lang 2010;24(4):632–47.
[5] Lee W, Song J-H, Chang J-H. Minima-controlled speech presence uncertainty tracking method for speech enhancement. Signal Process 2011;91(1):155–61.
[6] Lu C-T. Enhancement of single channel speech using perceptual-decision-directed approach. Speech Commun 2011;53(4):495–507.
[7] Boll S. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans Acoust Speech Signal Process 1979;27(2):113–20.
[8] Cohen I, Berdugo B. Speech enhancement for non-stationary noise environments. Signal Process 2001;81(11):2403–18.
[9] Choi J-H, Chang J-H. On using acoustic environment classification for statistical model-based speech enhancement. Speech Commun 2012;54(3):477–90.
[10] Ephraim Y, Malah D. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1984;32(6):1109–21.
[11] Ephraim Y, Malah D. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process 1985;33(2):443–5.
[12] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Process 2001;9(5):504–12.
[13] Cohen I, Berdugo B. Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process Lett 2002;9(1):12–5.
[14] Cohen I. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans Speech Audio Process 2003;11(5):466–75.
[15] Rangachari S, Loizou PC. A noise-estimation algorithm for highly non-stationary environments. Speech Commun 2006;48(2):220–31.
[16] Hendriks R, Heusdens R, Jensen J. MMSE based noise psd tracking with low complexity. In: Proc. IEEE ICASSP; 2010. p. 4266–9.
[17] Taghia J, Taghia J, Mohammadiha N, Sang J, Bouse V, Martin R. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In: Proc. IEEE ICASSP; 2011. p. 4640–3.
[18] Naylor PA, Gaubitch ND. Acoustic signal processing in noise: it's not getting any quieter. In: Proc. IWAENC; 2012. p. 1–6.
[19] Gerkmann T, Hendriks R. Unbiased MMSE-based noise power estimation with low complexity and low tracking delay. IEEE Trans Audio Speech Lang Process 2012;20(4):1383–93.
[20] Kates JM. Classification of background noises for hearing-aid applications. J Acoust Soc Am 1995;97(1):461–70.
[21] Alexandre E, Cuadra L, Álvarez L, Rosa-Zurera M, López-Ferreras F. Automatic sound classification for improving speech intelligibility in hearing aids using a layered structure. In: Intelligent data engineering and automated learning–IDEAL 2006. Springer; 2006. p. 306–13.
[22] Büchler M, Allegro S, Launer S, Dillier N. Sound classification in hearing aids inspired by auditory scene analysis. EURASIP J Appl Signal Process 2005;2005:2991–3002.
[23] Xiang J-J, McKinney M, Fitz K, Zhang T. Evaluation of sound classification algorithms for hearing aid applications. In: 2010 IEEE international conference on acoustics speech and signal processing (ICASSP); 2010. p. 185–8.
[24] Ma L, Milner B, Smith D. Acoustic environment classification. ACM Trans Speech Lang Process 2006;3(2):1–22.
[25] Chu S, Narayanan S, Kuo C-C. Environmental sound recognition with time-frequency audio features. IEEE Trans Audio Speech Lang Process 2009;17(6):1142–58.
[26] Gopalakrishna V, Kehtarnavaz N, Mirzahasanloo TS, Loizou PC. Real-time automatic tuning of noise suppression algorithms for cochlear implant applications. IEEE Trans Biomed Eng 2012;59(6):1691–700.
[27] Hu Y, Loizou P. Evaluation of objective quality measures for speech enhancement. IEEE Trans Audio Speech Lang Process 2008;16(1):229–38.
[28] Hu Y, Loizou PC. Subjective comparison and evaluation of speech enhancement algorithms. Speech Commun 2007;49(7–8):588–601.
[29] IEEE Subcommittee. IEEE recommended practice for speech quality measurements. IEEE Trans Audio Electroacoust 1969;3(AU-17):225–46.
[30] Varga A, Steeneken HJ. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun 1993;12(3):247–51.
[31] Virag N. Single channel speech enhancement based on masking properties of the human auditory system. IEEE Trans Speech Audio Process 1999;7(2):126–37.
[32] Burges C. A tutorial on support vector machines for pattern recognition. Data Min Knowl Discovery 1998;2(2):121–67.
[33] Chang C, Lin C. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) 2011;2(3):27:1–27:27.
[34] Knerr S, Personnaz L, et al. Single-layer learning revisited: a stepwise procedure for building and training a neural network. Optim Methods Softw 1990;1:23–34.
[35] Erkelens J, Hendriks R, Heusdens R, Jensen J. Minimum mean-square error estimation of discrete fourier coefficients with generalized gamma priors. IEEE Trans Audio Speech Lang Process 2007;15(6):1741–52.