EURASIP Journal on Advances in Signal Processing

RESEARCH Open Access

Multi-resolution auditory cepstral coefficient and adaptive mask for speech enhancement with deep neural network

Ruwei Li[1*], Xiaoyue Sun[1], Yanan Liu[1], Dengcai Yang[1] and Liang Dong[2]

Abstract

The performance of the existing speech enhancement algorithms is not ideal in low signal-to-noise ratio (SNR) non-stationary noise environments. In order to resolve this problem, a novel speech enhancement algorithm based on multi-feature and adaptive mask with deep learning is presented in this paper. First, we construct a new feature called multi-resolution auditory cepstral coefficient (MRACC). This feature which is extracted from four cochleagrams of different resolutions can capture the local information and spectrotemporal context and reduce the algorithm complexity. Second, an adaptive mask (AM) which can track noise change for speech enhancement is put forward. The AM can flexibly combine the advantages of an ideal binary mask (IBM) and an ideal ratio mask (IRM) with the change of SNR. Third, a deep neural network (DNN) architecture is used as a nonlinear function to estimate adaptive mask. And the first and second derivatives of MRACC and MRACC are used as the input of the DNN. Finally, the estimated AM is used to weight the noisy speech to achieve enhanced speech. Experimental results show that the proposed algorithm not only further improves speech quality and intelligibility, but also suppresses more noise than the contrast algorithms. In addition, the proposed algorithm has a lower complexity than the contrast algorithms.

Keywords: Speech enhancement, Deep neural network, Multi-resolution auditory cepstral coefficient, Adaptive mask

1 Introduction

Over the past several decades, a large number of approaches were proposed to solve the problem of speech enhancement. The traditional methods, such as spectral subtraction [1], wiener filtering [2, 3], minimum mean square error (MMSE) [4], statistical model [5, 6], and wavelet transform [7, 8], make statistical assumptions about the background noise and do not handle properly non-stationary noises, which are very common in our daily life.

With the appearance of the computational auditory scene analysis (CASA), the method based on the auditory scene analysis was applied to the speech enhancement [9]. For example, Zhang et al. proposed a speech enhancement based in CASA [10], which extracted the features and estimated the spectrum in gammatone domain as well as filtered out the noise by IRM. This approach has no hypothesis about noise which makes it fit for handling non-stationary noises and has a better generalization capability to process in a complex noise environment. However, it is difficult to deal with unvoiced speech which will result in a poor perceptual quality.

As the development of the deep learning, DNN has become one of the most popular methods for speech enhancement. The speech enhancement algorithm based on DNN is to learn the complex nonlinear relationship between noisy speech and clean speech [11]. Its deep structures are good at learning the nonlinear relationship between noise and speech and performing better in non-stationary background noise. According to the training target, deep learning-based speech enhancement algorithms can be divided into mapping and masking [11]. Researchers have proposed many speech enhancement algorithms in mapping [12]. For example, in 2014, Weninger et al. proposed a single-channel speech separation with

* Correspondence: liruwei@bjut.edu.cn
[1] Beijing Key Lab of Computational Intelligence and Intelligent System, Faculty of Information Technology, School of Information and Communications Engineering, Beijing University of Technology, Beijing, China
Full list of author information is available at the end of the article

memory-enhanced recurrent neural networks [13]. In this algorithm, a long short-term memory recurrent neural network (LSTM-RNN) was employed as a nonlinear regression function to predict clean speech as well as noise features from noisy speech features, and then a magnitude domain soft mask was constructed from these features. In 2015, Xu et al. extended the DNN-based speech enhancement framework to handle adverse conditions and non-stationary noise types in real situations [14]. In the same time, Huang et al. put forward a joint optimization of masks and deep recurrent neural networks (DRNN) for Monaural Source Separation algorithm [15]. In 2016, Vu et al. also presented a speech enhancement algorithm combining non-negative matrix factorization and deep neural networks [16]. These algorithms mentioned above are all to estimate the amplitude spectrum of the target speech. However, it is very difficult to estimate the amplitude spectrum of the target speech accurately, and these algorithms all have so high algorithm complexity and long-time delay that they cost much on calculation and are not suitable for real-time system. Besides, Li et al. proposed an improved least mean square adaptive filtering (ILMSAF)-based speech enhancement algorithm with DNN and noise classification [17], which introduces an adaptive coefficient of filter's parameters based on ILMSAF. This algorithm has good performance, but is too complex to be used in practice.

In addition, many researchers regarded the time-frequency masking as the target of deep learning for speech enhancement and proposed some speech enhancement algorithms. For example, Wang et al. presented a speech enhancement system based on deep neural network-support vector machine (DNN-SVM) [18]. In this system, the IBM was the target of the DNN-SVM model. Arun et al. proposed an IRM estimator using deep neural networks for robust speech recognition [19]. In this algorithm, the estimated IRM in the Mel-frequency domain is used to filter out noise from noisy Mel spectrogram. In 2014, Wang et al. used a fixed set of complementary features which include amplitude modulation spectrogram, relative spectral transformed perceptual linear prediction coefficient, Mel-frequency cepstral coefficient (MFCC), and 64-channel gammatone feature [20]. In addition, Chen et al. proposed a new feature called multi-resolution cochleagram (MRCG) [21]. However, the MRCG dimension is so large that the algorithm complexity is very high. In 2015, Tseng et al. took a classification-based approach, where the goal is to estimate an IBM and the sparse non-negative matrix factorization (SNMF) is used to extract features from the noisy speech [22]. In 2016, Yi Jiang et al. developed a DNN parameter mask for binaural reverberant speech segregation [23]. In 2017, Li et al. presented an IRM estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions [24], Zhang et al. presented a multi-target ensemble learning for monaural speech separation [25], and Sun et al. also proposed a multiple-target deep learning for LSTM-RNN-based speech enhancement [26]. In this algorithm, an IRM and a log-power spectral are regarded as the goal of training DNN. But the performance of the above speech enhancement algorithms based on DNN is non-ideal in low SNR environments, and the complexity of these algorithms is very high.

Through the above analysis, a speech enhancement algorithm based on MRACC and DNN is proposed. Firstly, a new feature parameter called MRACC is presented on the basis of the MRCG feature adopted [21]. Secondly, in order to remove the noise, an adaptive mask is constructed. Thirdly, we adopt a DNN model with four hidden layers to estimate an adaptive mask. Finally, the enhanced speech is synthesized by using the estimated adaptive mask and noisy speech. The experimental results show that the proposed speech enhancement algorithm has stronger robustness, better denoising performance, and lower complexity than the contrast algorithm [20, 21].

This paper is organized as follows. In Section 2, the proposed speech enhancement based on MRACC and AM with deep neural network is presented. Simulation experiments are given in Section 3 to illustrate the proposed algorithm performance. Finally, we summarize our work in Section 4.

2 Speech enhancement algorithm with deep neural network

2.1 Time-frequency decomposition

The speech signal is a typical time-varying signal. Its time-frequency decomposition focused on the time-varying spectral features of speech signal components, which decomposes a one-dimensional speech signal into a two-dimensional signal in order to reveal the relationship between these frequency components obtained and time [27]. The gammatone filter is an excellent tool for time-frequency decomposition. It can well simulate the sharp filtering characteristics of the basilar membrane, and it is in accordance with the auditory perception of the human ear [28]. Besides, it is easy to be achieved. So in this paper, the gammatone filter is used to decompose the noisy speech into several sub-band signals (see our previous work [29]). The impulse response of the gammatone filter is as follows:

$$g(t, f_{\mathrm{c}}) = t^{l-1}\, e^{-2\pi B f_{\mathrm{c}} t} \cos(2\pi f_{\mathrm{c}} t + \varnothing)\ t \geq 0 \tag{1}$$

where t represents the sample index; f_{c} is the center frequency for c^{th} channel, which varies from 50 to 8000 Hz;

and ϕ is the initial phase of the gammatone filter. In order to simplify the model, ϕ is set to 0. l is the filter order. A large number of experiments show that, when $l = 4$, the filter can well simulate the cochlear filter characteristics. So l equals 4 in this paper. The sampling rate of experimental data is set to16 kHz. In order to better reflect the harmonic characteristics of speech signal in each sub-band signal, the number of filters is determined to be 64 in the proposed algorithm. $B(f_c)$ is the bandwidth of each frequency channel, which is defined as:

$$B(f_c) = b \cdot \mathrm{ERB}(f_c) \tag{2}$$

where b is an attenuation factor, the best filter performance can be obtained when b equals 1.019, so b is set to 1.019 in this paper. ERB(f_c) represents the equivalent rectangular bandwidth (equivalent rectangle bandwidth (ERB)), and the relationship between the equivalent rectangular bandwidth and the central frequency f_c can be described by:

$$\mathrm{ERB}(f_c) = 24.7(4.37 f_c/1000 + 1) \tag{3}$$

where the coefficients *24.7* and *4.37* are the empirical values obtained in the experiment [21].

The expression of the input signal can be expressed as:

$$x(t) = s(t) + n(t) \tag{4}$$

where $x(t)$ represents noisy speech signal, $s(t)$ represents clean signal, and $n(t)$ represents noise signal.

$x(t)$ is decomposed into 64 sub-band signals $G(t, f_c)$ by 64-channel gammatone filters, as shown in formula (5):

$$G(t, f_c) = g(t, f_c) \cdot U(t) \cdot x(t) \tag{5}$$

where $U(t)$ is the unit step function.

Then, each sub-band signal is divided into time-frequency (T-F) units with a 20-ms frame with a 10-ms frame shift. A T-F unit corresponds to a small auditory unit of the noisy speech. It is defined as:

$$y_i(t, f_c) = w(t) \cdot G(((i\text{–}1) \cdot inc + t), f_c) \tag{6}$$

where $w(t)$ is a window function. Compared with the rectangular window, Hamming window can better reflect the frequency characteristic of speech signal, so Hamming window is chosen in this paper. $y_i(t, f_c)$ is the sub-band T-F unit for c^{th} channel at time frame i; *inc* is a frame shift.

The power of the auditory filter (cochleagram) of each T-F unit CG(i, f_c) is calculated by:

$$\mathrm{CG}(i, f_c) = \sum_{t=0}^{L-1} {y_i}^2(t, f_c) \tag{7}$$

2.2 Feature extraction

Good features are crucial to the performance of speech enhancement. In 2014, a feature called MRCG [21] is proposed by Chen et al., which is extracted from four cochleagrams of different resolutions to capture both local information and spectrotemporal context. As we all know, human auditory nonlinearity expands small sounds and compresses large sounds. The MRCG simulates auditory nonlinearity by a log function. Log function can well compress high power intensity but overexpands very small signals to infinity. Considering that noise often occurs in very small energy, log function often emphasizes noise and results in poor noise robustness; in order to avoid overexpansion to very small noise, we use a power function in four cochleagrams, which can better simulate human auditory nonlinearity. In addition, the dimension of MRCG is so large that the computational complexity is very high. In order to reduce the computational complexity, we employ a discrete cosine transform (DCT) operation on the basis of the power compression. The modified MRCG is called MRACC.

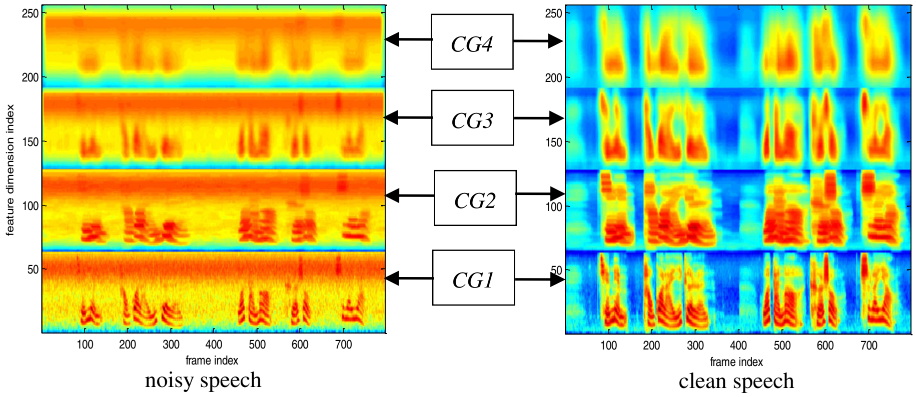

Fig. 1 MRCG feature (log compression)

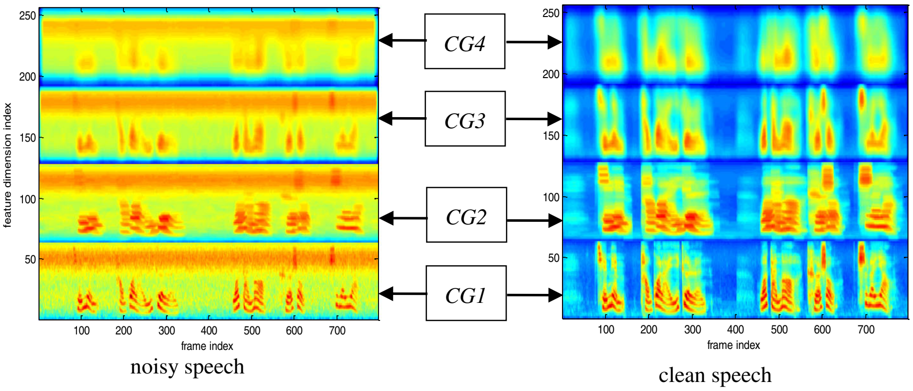

Fig. 2 IMRCG (power compression) feature examples

2.3 Extraction of MRACC feature

Firstly, the noisy speech $x(n)$ is decomposed by the gammatone filter bank into 64 sub-band signals, and the first 64-channel cochleagram (CG) is calculated with the frame length of 20 ms and frame shift of 10 ms. A power function is applied to the CG of each T-F unit. The mathematical expression of CG1 is:

$$CG1(i,f_c) = g[\mathrm{CG}(i,f_c)] \tag{8}$$

where $g()$ is a power function, $g = x^a$, and in this paper, $a = 1/15$ as Kim suggested [30].

Similarly, the second 64-channel cochleagram (CG2) is computed with the frame length of 200 ms and frame shift of 10 ms.

The third 64-channel cochleagram (CG3) is derived by averaging CG1 across a square window of 11 frequency channels and 11 time frames centered at a given T-F unit. It can be expressed as:

$$\mathrm{CG3}(i,f_c) = \sum_{k=c\text{-}5}^{c+5}\sum_{j=i\text{-}5}^{i+5}(\mathrm{CG1}(i,f_c))/(11*11) \tag{9}$$

The fourth 4-channel cochleagram (CG4) is calculated in a similar way to CG3, except that a square window of 23 frequency channels and 23 time frames is used. It can be shown as:

$$\mathrm{CG4}(i,f_c) = \sum_{d=c\text{-}11}^{c+11}\sum_{j=i\text{-}11}^{i+11}(\mathrm{CG1}(i,f_c))/(23^*23) \tag{10}$$

The CG1, CG2, CG3, and CG4 are connected to obtain an improved MRCG (IMRCG) feature, which has 64 × 4 dimensions for each time frame. The IMRCG feature is denoted as:

$$\mathrm{IMRCG}(i,f_c) = [\mathrm{CG1}(i,f_c);\mathrm{CG2}(i,f_c);\mathrm{CG3}(i,f_c);\mathrm{CG4}(i,f_c)] \tag{11}$$

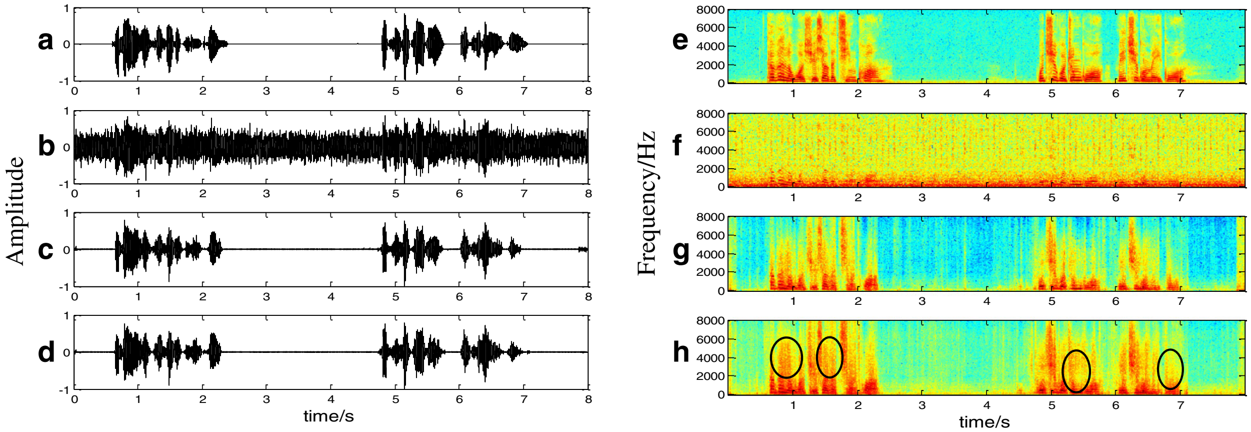

Fig. 3 Results sample of speech enhancement based on MRCG and MRACC features. **a** Waveform of clean speech. **b** Waveform of noisy speech. **c** Waveform of enhanced speech with MRCG. **d** Waveform of enhanced speech with MRACC. **e** Spectrogram of clean speech. **f** Spectrogram of noisy speech. **g** Spectrogram of enhanced speech with MRCG. **h** Spectrogram of enhanced speech with MRACC

The visualization of MRCG feature and the proposed IMRCG feature is given in Figs. 1 and 2, respectively. The left plots features extracted from a white noise mixture at − 5 dB SNR, and the right from the corresponding clean speech.

As shown in Figs. 1 and 2, both the MRCG feature and the IMRCG feature can partially retain spectrotemporal information of speech in noise environment. However, compared with the MRCG feature, the IMRCG feature has a clearer banded structure of speech than the MRCG. Therefore, the IMRCG is more capable of characterizing the difference between speech and noise.

In order to reduce the complexity of the algorithm, we reduce the dimension of the extracted features by a discrete cosine transform (DCT), because the DCT has the ability to aggregate the energy to the low frequency. Therefore, MRACC is obtained by a DCT operation to the IMRCG, which can be defined as follows:

$$\mathrm{MRACC}(i,m)=\left(\frac{2}{M}\right)^{0.5}\sum_{c=1}^{M}\mathrm{IMRCG}(i,f_c)\cos\left(\frac{\pi m(2c-1)}{2M}\right) \tag{12}$$

where MRACC(i, m) denotes the multi-resolution auditory cepstral coefficient of the i^{th} frame of the c^{th} sub-band, M is the number of channels, and M equals 64. m is the feature dimension index. When $m > 36$, the value of MRACC(i, m) is relatively small, so we retain the coefficient of the first 36 of MRACC(i, m).

2.4 Extraction of dynamic feature

In order to improve the accuracy of the target estimate, dynamic features are extracted from the MRACC, because delta features contain some temporal context. Therefore, the combination of the original and dynamic features can improve the accuracy of the target estimation. This method avoids having to rely on recurrent neural network to get temporal dynamics and reduce algorithm complexity.

The dynamic features (ΔMRACC and ΔΔMRACC) are obtained from formulas (13) and (14):

$$\Delta\mathrm{MRACC}\ (i,m)=\frac{\sum_{k=1}^{K}k(\mathrm{MRACC}\ (i+k,m)-\mathrm{MRACC}\ (i-k,m))}{\sqrt{2\sum_{k=1}^{K}k^2}} \tag{13}$$

$$\Delta\Delta\mathrm{MRACC}(i,m)=\frac{\sum_{k=1}^{K}k(\Delta\mathrm{MRACC}(i+k,m)-\Delta\mathrm{MRACC}(i-k,m))}{\sqrt{2\sum_{k=1}^{K}k^2}} \tag{14}$$

where k is a constant and it is set to 2, which represents the first two frames and the last two frames of the current frame. So, in this paper, the proposed feature v can be defined as:

$$v(i,m)=[\mathrm{MRACC}(i,m);\mathit{\Delta}\mathrm{MRACC}(i,m);\mathit{\Delta\Delta}\mathrm{MRACC}(i,m)] \tag{15}$$

Figure 3 shows the waveform and spectrogram of an utterance tested based on the proposed MRACC feature and the MRCG feature.

It can be seen from Fig. 3 that the residual noise in enhanced speech based on the MRACC feature is almost as much as the noise of the enhanced speech based on the MRCG feature. But compared with the enhanced speech based on MRCG feature, the enhanced speech based on MRACC feature retains more speech information and is closer to the clean speech. Therefore, the MRACC feature is better than the MRCG feature.

2.5 Deep neural network model

Due to the strong nonlinear mapping ability of DNN, we proposed a DNN-based adaptive mask estimator to calculate the adaptive mask for each T-F unit of the noisy speech. In the training phase, the adaptive mask for each T-F unit of noisy speech in the training data is calculated (see in Section 2.4) and used as the training target to train the DNN. Then, in the test phase, the adaptive mask is estimated by the trained DNN with MRACC inputs, which is used to synthesize the enhanced speech with noisy speech. The DNN is usually made up of three parts: the input layer, the hidden layer, and output layer. The input layer is the feature vector of the noisy speech, the hidden layer is stacked by the multiple hidden layers, and the output layer is an adaptive

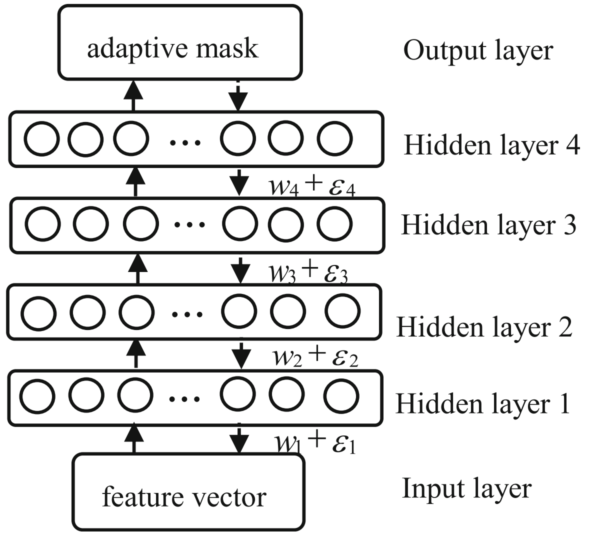

Fig. 4 The architecture of DNN

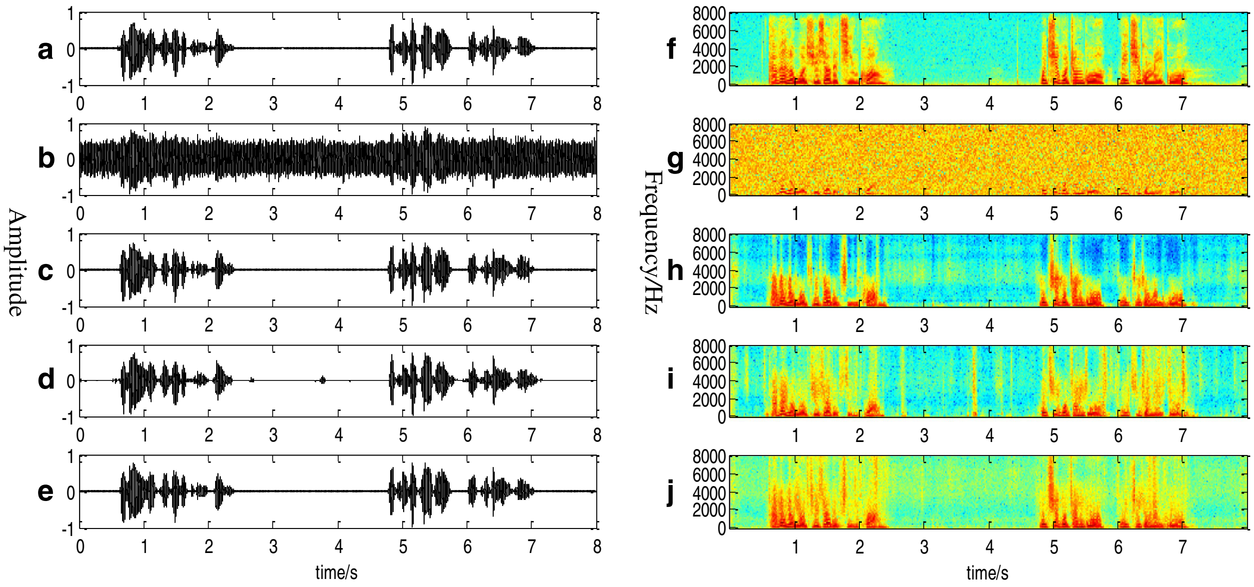

Fig. 5 Speech enhancement effect samples with − 5 dB white noise. **a** Waveform of clean speech. **b** Waveform of noisy speech. **c** Waveform of enhanced with IBM speech. **d** Waveform of enhanced speech with IRM. **e** Waveform of enhanced speech with the proposed adaptive mask. **f** Spectrogram of clean speech. **g** Spectrogram of noisy speech. **h** Spectrogram of enhanced speech with IBM. **i** Spectrogram of enhanced speech with IRM. **j** Spectrogram of enhanced speech with the proposed adaptive mask

mask. The structure of the DNN in this paper is shown in Fig. 4.

The structure of the DNN model constructed in this paper is composed of one input layer, four hidden layers, and one output layer. The proposed MRACC feature is a 432-dimensional vector, so the number of input layer's neuron is 432. The experimental results show that DNN has the best performance when the hidden layer units are 1024. Therefore, each hidden layer has 1024 rectified linear units (relu), which can improve generalization and avoid gradient disappearance problem. One frame adaptive masking threshold is a 64-dimensional vector, so the number of output layer's unit is 64. The activation function of the output layer is a sigmoid function. Consequently, the structure of DNN is 432-1024-1024-1024-1024-64.

The training of DNN employs the standard backpropagation (BP) algorithm which couples with dropout regularization. Dropout regularization can overcome the overfitting in DNN training, which discards a certain percentage of the neural units randomly to prevent complex co-adaptation among hidden units, forcing each hidden unit not to rely on each other. In this paper, the dropout rate is 0.2. Besides, no unsupervised pre-training is used. For a large training set, the effect of pre-training will be weakened. The mean squared error (MSE) is used as the loss function in the standard backpropagation algorithm. To improve the MSE function, we use an adaptive gradient descent algorithm along with a momentum term. In the training processing, the number of epochs is 25. For the first five epochs,

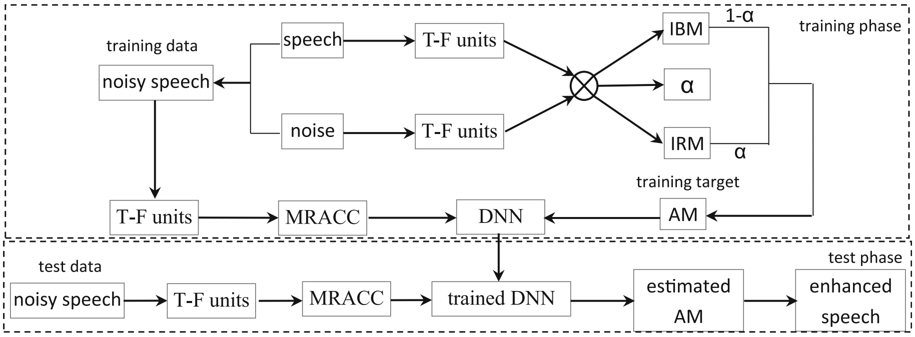

Fig. 6 The principle block diagram of speech enhancement algorithm

Table 1 The SegSNR of the proposed algorithm and the contrast algorithm

Noise type	SNR (dB)	SegSNR (dB)				Noise type	SNR (dB)	SegSNR (dB)			
		Noisy speech	Contrast1	Contrast2	Proposed algorithm			Noisy speech	Contrast1	Contrast2	Proposed algorithm
Buccaneer1	10	− 16.39	1.99	1.52	3.13	hfchannel	10	− 16.44	4.58	4.77	5.26
	5	− 21.18	0.03	− 0.09	0.92		5	− 21.13	2.45	2.16	2.78
	0	− 25.68	− 1.82	− 1.82	− 1.16		0	− 25.70	− 0.04	0.02	0.51
	− 5	− 29.59	− 4.72	− 4.91	− 4.69		− 5	− 29.65	− 2.38	− 2.74	− 2.23
Buccaneer2	10	− 16.40	2.36	2.33	3.75	Leopard	10	− 15.66	2.15	2.07	3.34
	5	− 21.15	0.43	0.56	1.57		5	− 20.36	0.96	0.73	1.51
	0	− 25.69	− 1.71	− 1.56	− 1.01		0	− 24.86	− 0.98	− 1.48	− 1.03
	− 5	− 29.52	− 4.92	− 4.51	− 4.41		− 5	− 28.92	− 6.25	− 5.50	− 5.90
Babble	10	− 15.91	0.86	1.55	2.45	m109	10	− 15.78	3.57	3.14	4.58
	5	− 20.59	− 1.95	− 1.45	0.24		5	− 20.55	1.18	0.75	2.06
	0	− 25.05	− 6.40	− 5.40	− 4.50		0	− 25.10	− 0.72	− 0.61	0.11
	− 5	− 28.82	− 14.84	− 11.66	− 12.15		− 5	− 28.98	− 3.15	− 2.53	− 3.09
Destroyerengine	10	− 16.26	2.87	2.67	3.80	Machinegun	10	− 2.16	10.76	10.16	10.82
	5	− 21.03	0.96	0.68	1.73		5	− 7.53	8.29	8.08	8.57
	0	− 25.56	− 0.75	− 0.92	− 0.38		0	− 12.82	6.13	6.14	6.16
	− 5	− 29.46	− 3.15	− 2.86	− 2.81		− 5	− 15.84	2.96	3.16	3.44
Destroyerops	10	− 16.09	1.66	1.63	3.30	Pink	10	− 16.24	2.33	2.48	3.99
	5	− 20.99	− 0.55	− 0.28	0.90		5	− 20.97	0.59	0.49	1.66
	0	− 25.29	− 2.66	− 2.13	− 1.68		0	− 25.45	− 1.65	− 1.07	− 0.09
	− 5	− 29.21	− 5.67	− 6.24	− 5.24		− 5	− 29.33	− 5.18	− 4.20	− 3.34
f16	10	− 16.20	2.46	2.58	3.90	Volvo	10	− 13.99	8.57	8.20	9.80
	5	− 20.98	0.48	0.54	1.64		5	− 18.77	6.85	6.23	7.86
	0	− 25.50	− 1.53	− 0.83	− 0.26		0	− 23.20	5.01	5.44	6.12
	− 5	− 29.33	− 4.66	− 2.42	− 2.22		− 5	− 27.14	3.73	3.12	3.85
Factory1	10	− 15.88	1.74	2.18	3.48	White	10	− 16.41	3.50	3.45	5.00
	5	− 20.80	− 0.67	0.17	1.18		5	− 21.20	1.16	1.08	2.45
	0	− 25.11	− 4.09	− 2.25	− 2.16		0	− 25.73	− 0.87	− 1.44	− 0.42
	− 5	− 29.00	− 11.71	− 19.01	− 8.52		− 5	− 29.30	− 3.23	− 4.95	− 2.90
Factory2	10	− 15.76	2.63	2.70	4.14	Street	10	− 16.04	− 3.37	− 3.25	− 2.92
	5	− 20.55	0.43	0.63	1.73		5	− 19.81	− 6.89	− 5.94	− 4.88
	0	− 25.09	− 1.87	− 1.04	− 0.49		0	− 24.39	− 9.93	− 9.60	− 7.02
	− 5	− 29.02	− 6.17	− 3.98	− 3.83		− 5	− 27.23	− 14.62	− 13.61	− 11.60
Office	10	− 16.52	0.23	0.95	1.15	Average	10	− 15.18	2.87	2.89	4.06
	5	− 21.56	− 3.63	− 2.97	− 1.57		5	− 19.95	0.59	0.66	1.79
	0	− 25.24	− 9.84	− 8.38	− 6.70		0	− 24.44	− 1.98	− 1.58	− 0.82
	− 5	− 28.93	− 18.61	− 16.34	− 15.98		− 5	− 28.19	− 5.30	− 5.83	− 4.80

a momentum rate is set to 0.5, after which the rate increases to 0.9.

IBM is the main computing objective of computational auditory scene analysis. It has been proved to be able to greatly improve speech intelligibility [11], but it seriously damages the quality of speech. Compared with IBM, IRM has a better speech quality, but it has a worse speech intelligibility [18]. Therefore, in order to balance speech quality and intelligibility, we propose an AM as the training target which is adaptively obtained by IBM and IRM according to the noise change. We calculate the energy of each time-frequency unit of speech and

Table 2 The LSD of the proposed algorithm and the contrast algorithm

Noise type	SNR (dB)	LSD (dB)				Noise type	SNR (dB)	LSD (dB)			
		noisy speech	Contrast1	Contrast2	Proposed algorithm			noisy speech	Contrast1	Contrast2	Proposed algorithm
Buccaneer1	10	14.81	6.98	4.47	4.12	hfchannel	10	13.35	6.29	3.97	3.97
	5	16.73	7.80	4.72	4.62		5	15.11	7.06	4.50	4.49
	0	18.63	8.71	5.29	5.21		0	16.92	8.07	5.05	4.98
	− 5	20.35	10.37	7.19	7.12		− 5	18.54	8.99	5.95	5.94
Buccaneer2	10	16.43	7.77	4.95	4.43	Leopard	10	9.82	5.26	3.74	3.54
	5	18.43	8.49	4.95	4.62		5	11.35	5.69	3.75	3.74
	0	20.41	10.00	6.27	6.27		0	12.95	6.38	4.36	4.29
	− 5	22.13	11.63	8.18	8.08		− 5	14.54	8.06	4.98	5.85
Babble	10	11.55	6.31	4.33	4.11	m109	10	11.76	5.71	5.80	3.61
	5	13.27	7.31	5.04	4.67		5	13.43	6.57	6.57	4.03
	0	14.95	8.95	6.41	6.14		0	15.16	7.28	7.22	4.43
	− 5	16.43	11.63	8.95	9.36		− 5	16.78	8.20	8.04	5.24
Destroyerengine	10	13.06	6.30	4.03	4.03	Machinegun	10	6.17	2.89	3.65	2.80
	5	14.87	7.12	4.41	4.24		5	7.42	3.04	4.22	3.01
	0	16.67	7.90	4.71	4.68		0	8.61	4.63	4.67	3.39
	− 5	18.27	9.00	5.42	5.41		− 5	9.35	5.32	5.32	3.96
Destroyerops	10	13.47	6.69	4.32	3.97	Pink	10	15.13	7.10	7.04	4.02
	5	15.39	7.41	4.68	4.48		5	17.05	7.89	7.76	4.46
	0	17.08	8.39	5.12	5.08		0	18.96	8.97	8.64	5.06
	− 5	18.74	9.76	7.11	6.66		− 5	20.68	10.82	10.46	6.54
f16	10	13.57	6.42	3.94	3.88	Volvo	10	9.31	4.45	4.34	2.78
	5	15.41	7.22	4.48	4.32		5	10.92	5.09	4.98	2.98
	0	17.20	8.21	4.65	4.64		0	12.47	5.55	5.60	3.19
	− 5	18.85	9.78	5.62	5.55		− 5	13.92	6.12	6.19	3.57
Factory1	10	13.72	6.94	4.51	4.14	White	10	16.95	7.82	8.07	4.52
	5	15.56	7.67	4.73	4.68		5	18.98	8.93	9.10	5.29
	0	17.37	9.09	5.95	5.94		0	20.99	10.27	6.75	6.67
	− 5	18.85	11.91	14.12	8.93		− 5	22.80	11.70	8.60	8.23
Factory2	10	11.37	5.92	4.10	3.81	Street	10	9.92	7.19	6.75	5.85
	5	13.01	6.59	4.55	4.42		5	11.46	9.93	9.23	8.27
	0	14.72	7.72	5.07	4.99		0	13.31	11.78	10.41	9.69
	− 5	16.32	9.25	6.07	6.07		− 5	15.67	12.50	11.48	10.53
Office	10	12.78	10.91	9.63	9.05	Average	10	12.54	6.53	5.16	4.27
	5	14.85	13.03	12.73	11.59		5	14.30	7.40	5.79	4.82
	0	16.41	15.14	14.67	13.48		0	16.05	8.51	6.40	5.67
	− 5	17.25	16.03	15.89	14.43		− 5	17.55	9.95	8.09	7.08

noise and obtain the IBM, IRM, and signal-to-noise ratio corresponding to the noisy speech according to Eqs. (17~20) and (22~25). The adaptive masking coefficient (a) is derived by the signal-to-noise ratio, which is used to weight IBM and IRM to get the adaptive mask as the training target for DNN through Eq. (16).

The formula of adaptive mask proposed in this paper is as follows:

$$\begin{aligned}\mathrm{AM}(i,f_{\mathrm{c}}) = (1{-}\alpha(i,f_{\mathrm{c}})) * \mathrm{IBM}(i,f_{\mathrm{c}}) \\ + \alpha(i,f_{\mathrm{c}}) * \mathrm{IRM}(i,f_{\mathrm{c}})\end{aligned} \tag{16}$$

Table 3 The PESQ of the proposed algorithm and the contrast algorithm

Noise type	SNR (dB)	PESQ (score)				Noise type	SNR (dB)	PESQ (score)			
		Noisy speech	Contrast1	Contrast2	Proposed algorithm			Noisy speech	Contrast1	Contrast2	Proposed algorithm
Buccaneer1	10	2.40	2.79	2.84	2.92	hfchannel	10	1.81	2.95	2.81	3.02
	5	1.95	2.45	2.44	2.52		5	1.52	2.65	2.38	2.68
	0	1.61	2.03	2.03	2.10		0	1.33	2.31	2.21	2.35
	− 5	1.34	1.51	1.49	1.54		− 5	1.21	1.97	1.74	1.98
Buccaneer2	10	2.04	2.93	2.90	3.20	Leopard	10	2.12	2.54	2.41	2.75
	5	1.62	2.66	2.63	2.87		5	1.86	2.33	2.44	2.69
	0	1.34	2.34	2.32	2.56		0	1.80	2.13	2.11	2.34
	− 5	1.16	1.91	1.92	2.17		− 5	1.73	2.05	2.01	2.28
Babble	10	2.23	3.02	3.13	3.16	m109	10	2.70	2.90	2.81	3.05
	5	1.78	2.76	1.84	2.81		5	2.62	2.36	2.40	2.79
	0	1.42	2.42	1.75	2.45		0	1.97	2.19	2.25	2.54
	− 5	1.14	1.99	1.66	2.04		− 5	1.55	1.71	1.82	1.91
Destroyerengine	10	2.18	2.97	2.83	3.06	Machinegun	10	2.93	3.43	3.01	3.53
	5	1.77	2.73	2.41	2.73		5	2.62	3.18	2.99	3.30
	0	1.54	2.30	2.01	2.39		0	2.36	2.97	2.96	3.04
	− 5	1.39	2.03	1.42	2.04		− 5	1.25	2.51	2.43	2.79
Destroyerops	10	2.50	2.96	2.82	3.12	Pink	10	2.52	2.69	2.50	2.89
	5	2.02	2.70	2.16	2.72		5	2.11	2.38	2.43	2.58
	0	1.60	2.35	1.82	2.38		0	1.97	2.09	2.15	2.50
	− 5	1.24	1.92	1.57	1.97		− 5	1.65	1.81	1.92	3.02
f16	10	2.57	2.72	2.94	3.10	Volvo	10	2.30	2.64	2.53	2.86
	5	2..01	2.22	2.48	2.82		5	1.96	2.34	2.33	2.78
	0	1.80	2.01	2.15	2.45		0	1.65	1.79	1.97	2.00
	− 5	1.30	1.95	1.62	1.98		− 5	1.22	1.63	1.51	1.75
Factory1	10	2.21	2.86	2.79	2.99	White	10	2.47	2.57	2.83	3.10
	5	1.79	2.53	2.33	2.59		5	2.17	2.49	2.65	2.80
	0	1.44	2.08	1.95	2.10		0	1.90	2.11	2.31	2.56
	− 5	1.21	1.58	1.32	1.68		− 5	1.62	1.84	1.91	2.90
Factory2	10	2.53	3.04	2.91	3.19	Street	10	2.05	2.31	2.25	2.56
	5	2.16	2.84	2.72	2.85		5	1.80	2.04	2.14	2.32
	0	1.72	2.40	2.35	2.47		0	1.59	1.85	1.92	2.08
	− 5	1.39	2.06	1.92	2.07		− 5	1.36	1.60	1.65	1.90
Office	10	2.01	2.56	2.54	2.91	Average	10	2.33	2.81	2.75	3.02
	5	1.75	2.03	2.15	2.55		5	1.97	2.51	2.40	2.73
	0	1.45	1.96	1.98	2.17		0	1.75	2.19	2.13	2.38
	− 5	1.18	1.59	1.73	1.94		− 5	1.35	1.86	1.74	2.12

where IBM(i, f_c) denotes the ideal binary mask (IBM) [18]; it can be defined as follows:

$$\mathrm{IBM}(i,f_{\mathrm{c}})=\begin{cases}1 & E_{\mathrm{s}}(i,f_{\mathrm{c}})\geq E_{\mathrm{n}}(i,f_{\mathrm{c}})\cdot 10^{\frac{lc}{10}}\\ 0 & \text{else}\end{cases} \tag{17}$$

E_s(i, f_c) and E_n(i, f_c) represent the energy of clean speech and noise, respectively. They are calculated by formulas (18) and (19). lc is a threshold and is usually set to 1.

$$E_{\mathrm{s}}(i,f_{\mathrm{c}})=\sum_{t=0}^{L-1}{s_i}^2(t,f_{\mathrm{c}}) \tag{18}$$

Table 4 The STOI of the proposed algorithm and the contrast algorithm

Noise type	SNR (dB)	STOI (score)				Noise type	SNR (dB)	STOI (score)			
		Noisy speech	Contrast1	Contrast2	Proposed algorithm			Noisy speech	Contrast1	Contrast2	Proposed algorithm
Buccaneer1	10	0.905	0.945	0.947	0.948	hfchannel	10	0.908	0.950	0.949	0.950
	5	0.832	0.902	0.900	0.903		5	0.848	0.916	0.914	0.917
	0	0.738	0.834	0.835	0.836		0	0.776	0.869	0.863	0.869
	− 5	0.630	0.749	0.743	0.750		− 5	0.694	0.804	0.799	0.805
Buccaneer2	10	0.931	0.955	0.955	0.956	Leopard	10	0.953	0.968	0.966	0.967
	5	0.871	0.913	0.913	0.914		5	0.923	0.943	0.938	0.939
	0	0.824	0.843	0.845	0.847		0	0.877	0.907	0.894	0.903
	− 5	0.675	0.754	0.750	0.758		− 5	0.823	0.848	0.843	0.842
Babble	10	0.929	0.944	0.945	0.941	m109	10	0.969	0.978	0.978	0.978
	5	0.858	0.891	0.883	0.880		5	0.934	0.955	0.953	0.954
	0	0.757	0.806	0.782	0.777		0	0.871	0.912	0.907	0.907
	− 5	0.633	0.655	0.624	0.650		− 5	0.775	0.834	0.836	0.830
Destroyerengine	10	0.923	0.956	0.959	0.959	Machinegun	10	0.968	0.974	0.975	0.975
	5	0.867	0.926	0.927	0.929		5	0.941	0.960	0.960	0.961
	0	0.788	0.867	0.875	0.875		0	0.901	0.939	0.930	0.940
	− 5	0.688	0.796	0.796	0.798		− 5	0.839	0.901	0.901	0.902
Destroyerops	10	0.936	0.955	0.952	0.956	Pink	10	0.930	0.957	0.956	0.957
	5	0.873	0.907	0.910	0.910		5	0.868	0.915	0.916	0.916
	0	0.784	0.840	0.839	0.840		0	0.776	0.834	0.849	0.850
	− 5	0.667	0.730	0.732	0.733		− 5	0.663	0.753	0.748	0.753
f16	10	0.930	0.959	0.979	0.979	Volvo	10	0.988	0.990	0.994	0.995
	5	0.869	0.920	0.922	0.922		5	0.980	0.983	0.989	0.990
	0	0.782	0.864	0.863	0.864		0	0.967	0.974	0.980	0.981
	− 5	0.673	0.778	0.772	0.779		− 5	0.945	0.957	0.963	0.965
Factory1	10	0.924	0.940	0.942	0.942	White	10	0.930	0.956	0.957	0.957
	5	0.857	0.880	0.885	0.888		5	0.874	0.922	0.921	0.922
	0	0.759	0.805	0.796	0.805		0	0.802	0.847	0.862	0.863
	− 5	0.643	0.660	0.658	0.661		− 5	0.720	0.772	0.788	0.788
Factory2	10	0.956	0.968	0.967	0.968	Street	10	0.925	0.948	0.950	0.953
	5	0.913	0.937	0.936	0.937		5	0.897	0.916	0.920	0.925
	0	0.840	0.876	0.874	0.876		0	0.809	0.835	0.833	0.837
	− 5	0.736	0.792	0.789	0.795		− 5	0.756	0.795	0.792	0.800
Office	10	0.915	0.925	0.923	0.927	Average	10	0.937	0.956	0.958	0.960
	5	0.837	0.842	0.840	0.845		5	0.885	0.919	0.919	0.920
	0	0.740	0.765	0.763	0.767		0	0.811	0.859	0.862	0.862
	− 5	0.605	0.612	0.615	0.620		− 5	0.715	0.775	0.773	0.778

$$E_{\mathrm{n}}(i,f_{\mathrm{c}})=\sum_{t=0}^{L-1}{n_i}^2(t,f_{\mathrm{c}}) \tag{19}$$

IRM(i, f_c) is an ideal ratio mask (IRM) [23], which is defined as:

$$\mathrm{IRM}_{\mathrm{gamm}}(i,f_{\mathrm{c}})=\left(\frac{E_{\mathrm{s}}(i,f_{\mathrm{c}})}{E_{\mathrm{s}}(i,f_{\mathrm{c}})+E_{\mathrm{n}}(i,f_{\mathrm{c}})}\right)^{\beta} \tag{20}$$

β is an adjustable scale factor, and a large number of experiments show that when $\beta = 0.5$, the IRM has the best performance. Therefore, β is set to 0.5.

Table 5 The A/B test of the proposed algorithm and the contrast algorithm with 15 noise types

Algorithm	Noise type				
	Babble	Buccaneer1	Buccaneer2	Destroyerengine	Destroyerops
	Preference %	Preference %	Preference %	Preference %	Preference %
Proposed algorithm	70.25	73.55	70.25	54.25	54.50
Contrast1	15.25	6.75	15.50	16.25	20.50
Contrast2	10.25	4.75	12.25	11.50	15.00
No preference	4.25	15.00	1.50	18.25	10.00
Algorithm	Noise type				
	m109	Machinegun	Pink	Volvo	White
	Preference %	Preference %	Preference %	Preference %	Preference %
Proposed algorithm	43.75	68.00	40.75	46.00	60.50
Contrast1	17.50	15.25	11.50	17.50	11.75
Contrast2	20.00	15.75	14.00	20.25	20.50
No preference	18.75	1.00	33.75	16.25	7.25
Algorithm	Noise type				
	f16	Factory1	Factory2	hfchannel	Leopard
	Preference %	Preference %	Preference %	Preference %	Preference %
Proposed algorithm	57.50	59.25	81.25	77.25	41.00
Contrast1	10.00	8.25	2.00	0.00	23.00
Contrast2	15.50	10.50	5.00	10.00	25.25
No preference	17.50	22.00	11.75	12.75	10.75

The adaptive coefficient α(i, f_c) is defined as [7]:

$$\alpha(i,f_c)=\frac{1}{1+\ \exp(-\mathrm{SNR}(i,f_c))} \tag{21}$$

Here, $SNR(i, f_c)$ is a signal-to-noise ratio of each frame, which is calculated by formula:

$$\mathrm{SNR}(i,f_c)=\frac{y^2(i,f_c)}{n^2(i,f_c)} \tag{22}$$

$y^2(i, f_c)$ and $n^2(i, f_c)$ denote the noisy speech and noise energy of the i^{th} frame and c^{th} sub-band signal, respectively.

Assuming that the first six frames are noise, the noise energy of the remaining five frames except the first frame is calculated by the (Eqs. 23, 24, and 25), which is used as the noise energy of the sixth frame.

$$\overline{n}^2(i,f_c)=\frac{1}{5}\sum_{a=0}^{4}n^2(i-a,f_c) \tag{23}$$

$$y^2(i,f_c)=\frac{1}{N}\sum_{t=0}^{N-1}(y_i(t,f_c))^2 \tag{24}$$

$$n^2(i,f_c)=\alpha(i,f_c)\times {n_{i-1}}^2(t,f_c)+(1-\alpha(i,f_c))\times {y_i}^2(t,f_c) \tag{25}$$

where $n(i, f_c)$ is the initial noise energy, N is the number of sampling point in one frame and is set to 320, and a is the frame index.

Figure 5 shows the waveform and spectrogram of an utterance tested based on IBM, IRM, and the proposed adaptive mask. In Fig. 5, compared with the enhanced speech with IRM, the enhanced speech with IBM has less noise; however, the quality of the enhanced speech is poor. The enhanced speech with IRM can keep more speech information but the intelligibility of the enhanced speech. Through analyzing the advantages and disadvantages of IBM and IRM, we proposed an adaptive mask combining with IBM and IRM. The enhanced speech with adaptive mask not only has less residual noise but also retains speech information well. So, the proposed adaptive mask is outperformed than IBM and IRM.

Table 6 The A/B test of the proposed algorithm and the contrast algorithm with street and office noise

Algorithm	Noise type	
	Street	Office
	Preference %	Preference %
Proposed algorithm	33.75	28.00
Contrast1	27.50	25.25
Contrast2	18.75	22.75
No preference	20.00	24.00

Table 7 The A/B test of the proposed algorithm and the contrast algorithm at different SNRs

Algorithm	SNR (dB)			
	10	5	0	− 5
	Preference %	Preference %	Preference %	Preference %
Proposed algorithm	75.00	48.21	46.55	46.87
Contrast1	10.56	16.03	21.20	18.63
Contrast2	11.44	21.36	28.45	21.30
No preference	3.00	14.40	3.80	13.20

Table 8 The MOS of the proposed algorithm and the contrast algorithm

Noise type	SNR (dB)	MOS (score)				Noise type	SNR (dB)	MOS (score)			
		Noisy speech	Contrast1	Contrast2	Proposed algorithm			Noisy speech	Contrast1	Contrast2	Proposed algorithm
Buccaneer1	10	2.40	2.89	2.90	3.15	hfchannel	10	1.81	3.95	3.81	4.24
	5	1.85	2.35	2.34	2.62		5	1.52	3.65	3.38	3.68
	0	1.51	1.93	2.02	2.20		0	1.33	2.31	2.21	2.85
	− 5	1.04	1.55	1.50	2.00		− 5	1.21	1.97	1.74	2.01
Buccaneer2	10	2.54	3.53	3.20	4.20	Leopard	10	2.12	3.54	3.41	3.75
	5	1.62	2.86	2.83	3.27		5	1.86	2.83	2.64	3.12
	0	1.34	2.34	2.32	2.56		0	1.80	2.13	2.21	2.74
	− 5	0.51	1.91	1.92	2.17		− 5	1.73	2.05	2.01	2.58
Babble	10	2.23	4.02	4.13	4.36	m109	10	2.70	3.97	4.15	4.35
	5	1.78	3.26	3.34	3.81		5	2.62	3.36	3.40	3.89
	0	1.42	2.82	2.56	3.05		0	1.97	2.99	2.85	3.14
	− 5	1.14	1.99	1.66	2.04		− 5	1.55	2.31	2.22	2.81
Destroyerengine	10	2.18	3.97	4.05	4.46	Machinegun	10	2.53	4.13	4.21	4.73
	5	1.77	2.73	2.41	3.53		5	2.02	3.78	3.99	4.30
	0	1.54	2.49	2.51	2.96		0	1.86	3.07	3.16	3.94
	− 5	1.39	2.33	2.42	2.74		− 5	1.25	2.71	2.83	3.02
Destroyerops	10	2.50	3.96	3.82	4.12	Pink	10	2.52	3.69	3.50	3.89
	5	2.02	3.70	3.16	3.82		5	2.11	2.38	2.43	2.58
	0	1.60	2.35	2.82	3.38		0	1.97	2.09	2.15	2.50
	− 5	1.24	2.92	2.57	2.97		− 5	1.65	1.81	1.92	2.32
f16	10	2.57	3.72	3.94	4.10	Volvo	10	2.30	3.64	3.83	4.26
	5	2..01	3.22	3.48	3.82		5	1.96	2.84	2.93	3.58
	0	1.80	3.01	3.15	3.45		0	1.65	2.29	2.37	3.00
	− 5	1.30	2.95	2.62	2.98		− 5	1.22	1.93	1.81	2.75
Factory1	10	2.21	3.86	3.79	3.99	White	10	2.47	2.57	2.83	3.10
	5	1.79	3.53	3.33	3.59		5	2.17	2.49	2.65	2.80
	0	1.44	2.88	2.95	3.10		0	1.90	2.11	2.31	2.56
	− 5	1.21	2.58	2.32	2.88		− 5	1.62	1.84	1.91	2.20
Factory2	10	2.53	3.84	3.91	4.39	Street	10	2.05	2.31	2.25	2.56
	5	2.16	3.34	3.52	3.95		5	1.50	2.04	2.14	2.32
	0	1.72	2.98	3.05	3.47		0	1.29	1.85	1.92	2.08
	− 5	1.39	2.16	1.92	2.97		− 5	0.85	1.60	1.65	1.90
Office	10	2.01	2.96	3.04	3.11	Average	10	2.33	3.56	3.57	3.93
	5	1.75	2.03	2.25	2.85		5	1.91	2.96	2.95	3.38
	0	1.45	1.96	1.98	2.17		0	1.62	2.45	2.50	2.89
	− 5	1.08	1.59	1.73	1.94		− 5	1.26	2.13	2.04	2.48

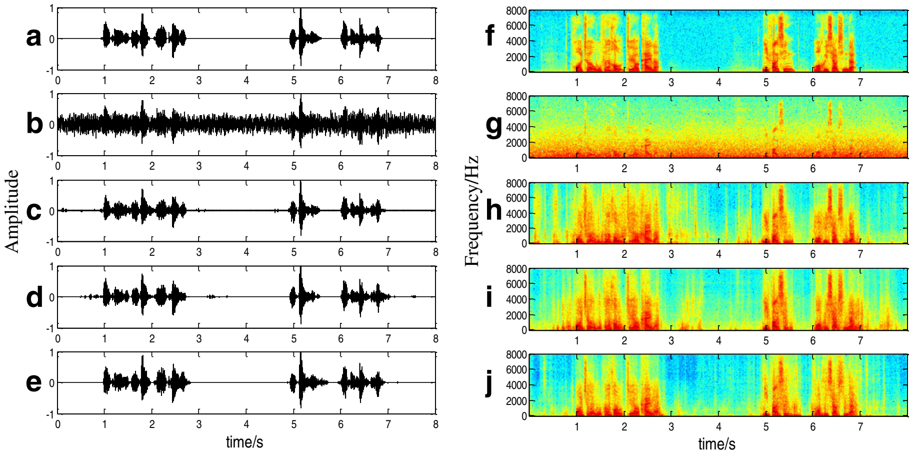

Fig. 7 Speech enhancement effect samples with − 5 dB factory2 noise. **a** Waveform of clean speech. **b** Waveform of noisy speech. **c** Waveform of enhanced speech with the contrast algorithm 1. **d** Waveform of enhanced speech with the contrast algorithm 2. **e** Waveform of enhanced speech with the proposed algorithm. **f** Spectrogram of clean speech. **g** Spectrogram of noisy speech. **h** Spectrogram of enhanced speech with the contrast algorithm 1. **i** Spectrogram of enhanced speech with the contrast algorithm 2. **j** Spectrogram of enhanced speech with the proposed algorithm

2.6 Algorithm implementation steps

The block diagram of complementation steps of the proposed algorithm is shown in Fig. 6. Figure 6 shows the processing pipeline of the proposed speech enhancement algorithm. In the training phase, we calculate the energy of each T-F unit of speech and noise and obtain the IBM, IRM, and signal-to-noise ratio corresponding to the noisy speech. The adaptive masking coefficient (a) is derived by the signal-to-noise ratio, which is used to weight IBM and IRM to get the adaptive mask (AM) as the training target for DNN. Then, the MRACC features of noisy speech are extracted as the inputs for deep learning. We train the DNN model and save the weights and thresholds of the DNN model after the training is completed. In this paper, the DNN architecture is 432-1024-1024-1024-1024-64. In the test phase, the MRACC feature vector of test sample is entered in the trained DNN network model to obtain an estimated adaptive mask, then the enhanced speech is synthesized by using the test sample and the estimated adaptive mask.

3 Results and discussions

3.1 Experimental data

In the experiment, clean utterances come from the NTT corpus. The sampling rate of data is set to 16 kHz. Three kinds of clean utterances are selected from the NTT corpus, including English, Chinese, and French. Each language library contains 96 sentences, which are produced by 8 speakers (4 male and 4 female speakers, and 12 utterances for each speaker). The length of each sentence is 8 s. Therefore, there are (96 × 3) 288 clean utterances. For each language, 76 clean sentences are randomly selected as the training data, and the remaining 20 sentences are tested as the test data. There are 17 noise types, namely buccaneer1, buccaneer2, babble, destroyerengine, destroyerops, f16, factory1, factory2, hfchannel, leopard, m109, machinegun, pink, volvo, white, office, and street selected from the NoiseX-92 database. The training set covers the first 15 noise types mentioned above. To evaluate the performance of the proposed algorithm in an unknown noise environment, office and street noise are used as the noise types that are not included in the training set. The 288 clean sentences are corrupted with abovementioned 17 noise types at 4 levels of SNR, i.e., 10 dB, 5 dB, 0 dB, and − 5 dB, to build a multi-condition data set.

In order to verify the effectiveness of the proposed algorithm, we select on training targets for supervised speech separation as the first contrast algorithm [20], and a feature study for classification-based speech separation at very low signal-to-noise ratio is considered as the second contrast algorithm [21].

3.2 Objective performance evaluation

The purpose of this test is to evaluate the performance of our proposed algorithm in complex noise environments. In this test, segment SNR (SegSNR), perceptual evaluation of speech quality (PESQ), log-spectral distortion (LSD), and short-time objective intelligibility

Table 9 The operation time comparison

Algorithm	Contrast1	Contrast2	Proposed algorithm
Time (s)	7.23	12.10	6.61

(STOI) are adopted as the objective measures of speech quality [31–33].

For the 17 noise types, the test results of SegSNR, PESQ, LSD, and STOI are shown in Tables 1, 2, 3, and 4, respectively.

It can be seen from Table 1 that for leopard noise with SNRs of 0 dB and − 5 dB, the SegSNR of the proposed algorithm is better than that of the contrast algorithm 1, but less than that of the contrast algorithm 2. For babble and m109 noise with SNR − 5 dB, the SegSNR of the proposed algorithm is better than that of the contrast algorithm 1, but less than that of the contrast algorithm 2. Compared with other noises, leopard and m109 noise have more complex time-frequency characteristics and the babble noise is similar to speech, so it is difficult to distinguish between speech and noise. But the average SegSNR under different SNRs of the proposed algorithm is all higher than that of the contrast algorithm. The reason is the MRACC feature in the proposed algorithm contains more phonetic information so that the speech signal will be separated from complex noise environments by DNN. So the proposed algorithm is better than the contrast algorithm in general.

As shown in Table 2, for babble and leopard noise with SNR − 5 dB, the LSD of the proposed algorithm is better than that of the contrast algorithm 1, but is a little weaker than that of the contrast algorithm 2. But for other noise types, compared with the contrast algorithm, the distortions all are reduced. And the average LSD under different SNRs of the proposed algorithm is all better than that of the contrast algorithm. Therefore, the distortion of enhanced speech based on the proposed algorithm is less than that based on the contrast algorithm on the whole.

We can know from Table 3, for 17 kinds of noise, the PESQ of the proposed algorithm is all greater than that of the contrast algorithm. Consequently, for the complex noise environment, the speech quality of enhanced speech based on our proposed algorithm is better than the contrast algorithm.

It can be seen from Table 4 that for babble noise, leopard noise, and m109 noise, the STOI of the proposed algorithm is similar to or slightly less than that of the contrast algorithm. But for other noises, the STOI of the proposed algorithm is a little better than that of the contrast algorithm and the average STOI of the proposed algorithm is all higher than that of the contrast lgorithm in every SNR condition. Therefore, the STOI of the proposed algorithm is slightly greater than the contrast algorithm overall.

3.3 Subjective performance evaluation

In order to test the performance of the proposed algorithm further, A/B test method, MOS (mean opinion score), waveform, and spectrogram are adopted as the subjective measures of speech quality. The A/B test method which is often used for page and process testing can reflect the user's preference for different versions of the page or process. Therefore, the A/B test method is adopted by this paper to test the subjective performance. Ten testers (five males and five females) are invited to conduct A/B test and MOS on the enhanced speech of the proposed algorithm and the comparison algorithm, respectively.

For the 17 noise types, SNR conditions include − 5 dB, 0 dB, 5 dB, and 10 dB; the test results of A/B test method are summarized in Tables 5, 6, and 7. The MOS score from 0 to 5 indicated that the speech quality is getting better. Table 8 presents the results of MOS at different SNRs across the 17 noise types.

We can see from Table 5 for 15 kinds of noise, the A/B test of the proposed algorithm is all higher than that of the contrast algorithm in every noise condition. In Table 6, for office noise and street noise, the A/B test of the proposed algorithm is also better than that of the contrast algorithm. Consequently, for the complex noise environment, the proposed algorithm has the stronger robustness. Therefore, the subjective speech quantity of the proposed algorithm is better than that of the contrast algorithm.

Table 7 shows that the A/B test of the proposed algorithm is higher than that of the contrast algorithm at different SNRs.

As shown in Table 8, for the 17 noise types, the MOS of the proposed algorithm is all higher than that of the compared method. Therefore, the subjective quality of enhanced speech based on the proposed algorithm is greater than that based on the contrast algorithm.

Figure 7 shows the waveform and spectrogram of the proposed algorithm and the contrast algorithm with factory2 noise at SNR = − 5 dB. We can know from Fig. 7, for factory2 noise with the SNR of − 5 dB, the proposed algorithm can eliminate most of the noise to a certain extent. But there is still a lot of noise in the contrast algorithm, which makes the listeners feel annoying. Consequently, the denoising effect of the proposed algorithm is better than that of the contrast algorithm. Therefore, the enhancement effect of the proposed algorithm is greater than that of the contrast algorithm on the whole.

According to the analysis of the above test results, we can come to the conclusion that the performances of SegSNR, LSD, PESQ, STOI, A/B test, and MOS of the proposed algorithm are greater than those of the compared method. Moreover, in the low SNR environments, the performance of the proposed algorithm is very excellent. Therefore, the proposed algorithm is more suitable for low SNR environments.

3.4 Algorithm complexity test

In order to test the algorithm complexity of this algorithm, the MATLAB operation time of each algorithm is shown in Table 9 in this paper. Each algorithm deals with all speech signals and then calculates the average length of time it takes for each speech to be processed. It can be seen that the operation time of the proposed algorithm is less than that of the contrast algorithm. After analysis, we can know that there are two reasons. Firstly, the complexity of the extraction process of MRACC feature in the proposed algorithm is lower than that of the contrast algorithm 1. Secondly, compared with the MRCG feature in contrast algorithm 2, the proposed MRACC feature dimension is reduced. Therefore, in a large number of experiments, the operation time of the proposed algorithm is lower than that of the contrast algorithm.

4 Conclusion

In this paper, a speech enhancement algorithm based on MRACC and adaptive mask with deep learning is proposed. In this algorithm, firstly, a new feature, MRACC, is presented. Compared with the MRCG feature, this feature uses power function instead of log function so that it can capture local information and spectro-temporal contexts, and DCT is employed to gather the power to the low frequency in this feature so that the dimension of feature is reduced according to the power's distribution. Therefore, the algorithm complexity of the proposed algorithm is reduced. Secondly, an adaptive mask which can track the noise changes is used for speech enhancement. Because the adaptive mask combines the advantages of IRM and IBM, so it has more accurate estimation on the target speech energy ratio with DNN. Thirdly, we adopt a DNN model with four hidden layers to estimate an adaptive mask. DNN has strong nonlinear processing ability, which could well describe the complex nonlinear relationship between noise and speech. So our proposed algorithm has better quality and intelligibility as well as lower algorithm complexity than the contrast algorithm overall.

Abbreviations

AM: Adaptive mask; BP: Backpropagation; CASA: Computational auditory scene analysis; CG: Channel cochleagram; DCT: Discrete cosine transform; DNN: Deep neural network; DNN-SVM: Deep neural network-support vector machine; DRNN: Deep recurrent neural networks; IBM: Ideal binary mask; ILMSAF: Improved least mean square adaptive filtering; IMRCG: Improved MRCG; IRM: Ideal ratio mask; LSD: Log-spectral distortion; LSTM-RNN: Long short-term memory recurrent neural network; MFCC: Mel-frequency cepstral coefficient; MMSE: Minimum mean square error; MRACC: Multi-resolution auditory cepstral coefficient; MRCG: Multi-resolution cochleagram; MSE: Mean squared error; PESQ: Perceptual evaluation of speech quality; SegSNR: Segment SNR; SNMF: Sparse non-negative matrix factorization; SNR: Signal-to-noise ratio; STOI: Short-time objective intelligibility

Funding

This work was supported by the Scientific Research Program of Beijing Municipal Commission of education (No. KM201510005007) and the National Natural Science Foundation of China(No.51477028).

Availability of data and materials

Please contact authors for data requests.

Authors' contributions

RL devised the algorithm, checked the experiment, and improved this paper. XS wrote the draft of this paper and did partial simulation experiments. YL programmed the code and did the simulation experiments. DY helped to check the codes. LD improved the English of this paper. All the authors wrote this paper together, and they have read and approved the final manuscript.

Ethics approval and consent to participate

This study does not involve human participants, human data, or human tissue.

Consent for publication

In the manuscript, there is no any individual person's data.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Beijing Key Lab of Computational Intelligence and Intelligent System, Faculty of Information Technology, School of Information and Communications Engineering, Beijing University of Technology, Beijing, China. [2]Electrical and Computer Engineering, Baylor University, Waco, TX 76798, USA.

Received: 29 September 2018 Accepted: 20 March 2019
Published online: 02 April 2019

References

1. S.F. Boll, Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. **27**(2), 113–120 (1979)
2. J.D. Chen, J. Benesty, Y.T. Huang, S. Doclo, New insights into the noise reduction Wiener filter. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1218–1234 (2006)
3. Loizou, *Speech Enhancement: Theory and Practice* (CRC Press, New York, 2007)
4. RC. Henddriks, R. Heusdens, J. Jensen, MMSE based noise PSD tracking with low complexity. Proc.IEEE Int. Conf. Acoustics, Speech, Signal Process, 4466–4469 (2010)
5. A. Ozerov, E. Vincent, F. Bimbot, A general flexible framework for the handling of prior information in audio source separation. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1118–1133 (2012)
6. N. Mohammadiha, P. Smaragdis, A. Leijon, Supervised and unsupervised speech enhancement using non-negative matrix factorization. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2140–2151 (2013)
7. L. Ruwei, B. Changchun, D. Huijing, Speech enhancement using adaptive threshold based on bi-orthogonal wavelet packet decomposition. Chin. J. Sci. Instrum. **29**(10), 2135–2140 (2008)
8. L. Ruwei, B. Changchun, D. Huijing, Speech enhancement algorithm based on wavelet transform. J Data Acquis Proc **24**(3), 362–368 (2009)
9. D.L. Wang, G.J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (IEEE Press, Piscataway, 2006)
10. Z. Weiqiang, G. Cong, Z. Qiao, K. Jian, H. Liang, L. Jia, T. Johnson Micheal, A speech enhancement algorithm based on computational auditory scene analysis. J Tian Jin Univ (Sci Technol) **48**(8), 663–669. (2015)
11. L. Wen, J. Nie, S. Liang, S. Zhang, X. Liang, Deep learning based speech separation technology and its developments. Zidonghua Xuebao/acta Automatica Sinica **42**(6), 819–833 (2016)

12. Y. Xu, J. Du, L.R. Dai, C.H. Lee, An experimental study on speech enhancement based on deep neural networks. IEEE Signal Process Lett. **21**(1), 65–68 (2014)
13. F. Weninger, F. Eyben, B. Schuller, in *IEEE International Conference on Acoustics Speech and Signal Processing*. Single-channel speech separation with memory-enhanced recurrent neural networks (IEEE Press, Florence, 2014), pp. 3737–3741
14. Y. Xu, J. Du, L.R. Dai, et al., A regression approach to soeech enhancement based on deep neural network. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(1), 7–19 (2015)
15. P.S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Joint optimization of masks and deep recurrent neural networks for monaural source separation. IEEE/ACM Trans. Audio Speech Lang. Process. **23**(12), 2136–2147 (2015)
16. T.T. Vu, B. Bigot, E.S. Chng, in *IEEE International Conference on Acoustics Speech and Signal Processing*. Combining non-negative matrix facorization and deep neural networks for speech enhancement and automatic speech recognition (IEEE Press, Shanghai, 2016), pp. 499–503
17. R. Li, Y. Liu, Y. Shi, W. Cui, ILMSAF based speech enhancement with DNN and noise classification. Speech Comm. **85**, 53–70 (2016)
18. Y. Wang, D.L. Wang, Towards scaling up classification-based speech separation. IEEE Trans. Audio Speech Lang. Process. **21**(7), 1381–1390 (2013)
19. A. Narayanan, D.L. Wang, *Ideal Ration Mask Estimation on Using Deep Neural Networks for Robust Speech Recognition* (IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013), pp. 1520–6149
20. Y.X. Wang, A. Narayanan, D.L. Wang, On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1849–1858 (2014)
21. J. Chen, Y. Wang, D.L. Wang, in *IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP)*. A feature study for classification-based speech separation at very low signal-to noise ratio (2014)
22. H.-W. Tseng, M. Hong, Z.-Q. Luo, in *IEEE International Conference on Acoustics,Speech and Signal Processing(ICASSP)*. Combing sparse NMF with neural network: a new classification-based approach for speech enhancement (2015)
23. Y. Jiang, W. Li, Y. Zu, in *The 9th International Congress on Image and Signal Processing BioMedical Engineering and Information (CISP-BMEI2016)*. A DNN parameter mask for the binaural reverberant speech segregation (2016)
24. L. Xu, J. Li, Y. Yan, in *Interspeech*. Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions (2017), pp. 1203–1207
25. H. Zhang, X. Zhang, G. Gao, in *Interspeech*. Multi-target ensemble learning for monaural speech separation [C]//INTERSPEECH. 1958-62, 2017
26. L. Sun, J. Du, L.-R. Dai, C.-H. Lee, *Multiple-Target Deep Learning for LSTM-RNN Based Speech Enhancement* (Hands-free Speech Communications and Microphone Arrays, HSCMA, 2017)
27. G. Zhexue, C. Zhongsheng, *Matlab Time Frequency Analysis Technology and its Application* (People's post and Telecommunications Press, Beijing, 2006)
28. Y.W. Yang, Y. Jiang, R.S. Liu, et al., in *Proc. Signal Processing, Communications and Computing (ICSPCC)*. A realtime analysis/synthesis Gammatone filterbank (2015), pp. 1–6
29. R. Li, D. Pan, S. Zhang, Speech enhancement algorithm based on sound source localization and scene matching for binaural digital hearing aids. J. Med. Biol. Eng. (2018). https://doi.org/10.1007/s40846-018-0412-z
30. C. Kim, *Signal Processing for Robust Speech Recognition Motivated by Auditory Processing*Ph.D. dissertation (Carnegie Mellon University, Pittsburgh, 2010)
31. T. Xiaoheng, Q. Jiwei, Z. Shuai, Objective evaluation method of speech quality based on auditory perceptual properties. J. Southwest Jiao Tong Univ. **48**(4), 756–760 (2013)
32. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136 (2011)
33. ITU-T Recommendation P. 862, *Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-to-End Speech Quality Assessment of Narrowband Telephone Networks and Speech Codecs* (International Telecommunication Union, Geneva, 2001)