Wang and Bao *EURASIP Journal on Audio, Speech, and Music Processing* (2019) 2019:20
https://doi.org/10.1186/s13636-019-0164-x

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Speech enhancement methods based on binaural cue coding

Xianyun Wang and Changchun Bao*

Abstract

According to the encoding and decoding mechanism of binaural cue coding (BCC), in this paper, the speech and noise are considered as left channel signal and right channel signal of the BCC framework, respectively. Subsequently, the speech signal is estimated from noisy speech when the inter-channel level difference (ICLD) and inter-channel correlation (ICC) between speech and noise are given. In this paper, exact inter-channel cues and the pre-enhanced inter-channel cues are used for speech restoration. The exact inter-channel cues are extracted from clean speech and noise, and the pre-enhanced inter-channel cues are extracted from the pre-enhanced speech and estimated noise. After that, they are combined one by one to form a codebook. Once the pre-enhanced cues are extracted from noisy speech, the exact cues are estimated by a mapping between the pre-enhanced cues and a prior codebook. Next, the estimated exact cues are used to obtain a time-frequency (T-F) mask for enhancing noisy speech based on the decoding of BCC. In addition, in order to further improve accuracy of the T-F mask based on the inter-channel cues, the deep neural network (DNN)-based method is proposed to learn the mapping relationship between input features of noisy speech and the T-F masks. Experimental results show that the codebook-driven method can achieve better performance than conventional methods, and the DNN-based method performs better than the codebook-driven method.

Keywords: Monaural speech enhancement, Codebook, Deep neural network, Binaural cue coding

1 Introduction

For speech communication system in natural environment, the background noise will cause an impairment to speech signal. Thus, it is necessary to reduce the effect of the noise by speech enhancement. The purpose of speech enhancement is to improve quality and intelligibility of speech by suppressing background noise.

Many speech enhancement methods were developed in the last few decades, such as spectral-subtractive algorithm [1, 2], Wiener filtering [3] and statistical model-based methods [4, 5]. These methods could achieve a good performance for stationary noise, but their performance becomes worse when the non-stationary noise is concerned. The main problem is that the estimation of non-stationary noise is a difficult task [6–8], for example, the estimation method of noise power spectrum [9] in a large buffer limits its ability to track rapid changes of noise energy [6–8].

In order to solve the problem of non-stationary noise estimation, some supervised approaches were proposed by using a priori knowledge of speech and noise. For example, in the auto-regressive hidden Markov model (ARHMM) method [10] and codebook-driven Wiener filtering methods [6, 8, 11, 12], the spectral shapes of speech and noise were considered as a priori information used for the pre-training. In the ARHMM-based methods, the spectral shapes of speech and noise are represented by the HMM model, and the speech signal is reconstructed by combining spectral gains of speech and noise. Since the spectral gain of noise is obtained on a short-frame basis, the quick changes of noise energy of non-stationary noise can be followed. In codebook-driven methods, the auto-regressive (AR) coefficients of speech and noise are used to train two shape codebooks of spectra by vector quantization method [13]. Once the AR gains of speech and noise are estimated by maximum likelihood (ML) technique [11] or the Bayesian minimum mean squared error (MMSE) technique [6, 8] or maximum posteriori probability (MAP) technique

* Correspondence: b201402001@emails.bjut.edu.cn; baochch@bjut.edu.cn
Speech and Audio Signal Processing Laboratory, Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China

[12], the spectral envelope of speech and noise could be obtained. Since the AR gain of noise is estimated on a short-frame basis, the codebook-driven methods could better track the energy changes of non-stationary noise to some extent [6, 8].

In addition to above-mentioned methods [6, 8, 10–12] using spectral envelope of speech and noise as a priori information, the spectral details of speech and noise were used as a priori knowledge in [14–25]. In [14, 15], Gaussian mixture model (GMM) was used to train the log spectra of speech and noise, and speech signal was estimated by a MMSE estimator. In the MMSE estimation, the weighted sum of the posterior probabilities of all Gaussian pairs was used to the MMSE estimators of speech. Later, in [16], in view of the lack of temporal dynamics in the GMM-based methods, a layered HMM model was incorporated to model the relationship between adjacent frames. In the references from [17] to [25], log spectrum of speech or the T-F mask was trained by deep neural network model for restoring spectral details. Hereinto, in [24, 25], the generalization ability of the DNN-based methods was also discussed, and these studies have shown that large-scale training with a wide variety of noises is helpful to noise generalization.

Considering the fact that the prior information of speech and noise can improve speech quality, our former works [26, 27] have shown an effectiveness of using binaural inter-channel cues between speech and noise to enhance speech. In previous studies based on the cue parameter [28–39], the binaural inter-channel cues [28–37] have been used to estimate ideal T-F mask in binaural computational auditory scene analysis (CASA) systems and have shown a good performance in binaural speech processing. In the BCC technique [40–42], the binaural inter-channel cues were viewed as the side information, which was combined with a down-mixed audio signal to recover the left channel and right channel audio signals. The down-mixing signal is a mono signal generated from the left and right channel signals. According to the principle of the BCC, the BCC technique can recover the input signals of the left and right channels, when the down-mixing signal and the inter-channel cues between the left and right channel signals are given. Based on this, for single-channel speech enhancement, when the noisy signal is seen as the down-mixing signal constructed by speech and noise, the exact inter-channel cues between speech and noise are obtained as well. We could exploit the BCC framework to extract clean speech from noisy speech. Compared with the original BCC framework [41, 42] that the left channel and the right channel correspond to left microphone and right microphone, respectively, in the BCC scheme used in speech enhancement, the left channel and the right channel correspond to speech signal and noise signal, respectively. In this paper, the noisy signal is seen as the down-mixing signal of speech and noise based on the BCC framework [40–42]; two kinds of the T-F mask estimation methods are proposed to estimate clean speech from noisy speech by extracting and training the inter-channel cues between speech and noise.

For the first T-F mask estimation, the inter-channel cues between speech and noise are trained as a priori codebook similar to [26]. However, in [26], multiple frequency sub-bands have the same inter-channel correlation (ICC) between speech and noise, which limits the ability of reducing noise. In the proposed codebook-based method, the shared correlation of multiple frequency sub-bands is avoided by modifying the calculation of the inter-channel cues, that is, in this paper, the ICC between speech and noise is considered in each frequency sub-band. In addition, in the proposed method, the pre-enhanced cues following the pattern involved in [22] are extracted to generate a vector with 28 dimensions to replace a vector with 2 dimensions in [26], which help to improve the accuracy of selecting the optimal code-vectors by the weighted mapping technique [43]. Once the pre-enhanced cues and exact cues are extracted, they can be combined one by one for training a codebook offline. In online enhancement stage, by comparing distance between the pre-enhanced cues online and cues stored in the codebook, the exact cues can be obtained by a combinatorial mapping and used to generate the T-F masking estimator for enhancing noisy speech.

Since the codebook-driven method is much sensitive to the pre-enhancing method, and the inter-channel cues with unboundedness is detrimental to the supervised methods based on the gradient descent [17], the second technique cancels the pre-enhancing module and uses the DNN to directly learn the mapping relationship between input features of noisy speech and the T-F mask synthesized by exact inter-channel cues.

The rest of this paper is organized as follows. In Section 2, the relationship between the BCC and speech enhancement is described. In Section 3, we discuss the details of the proposed method. Experimental results are provided in Section 4, and Section 5 provides the conclusions.

2 The BCC framework and speech enhancement system

2.1 A brief description of the BCC framework

The binaural cue coding (BCC) [40] is a kind of stereo coding method. Its principle is shown in Fig. 1. The BCC method is composed of the encoder and decoder and adopts the coding way combining the down-mixed process and side information [40]. In the BCC, the binaural inter-channel cues are usually selected as the BCC parameters, which are termed as side information. In Fig. 1, to satisfy the coding way of the down-mixed

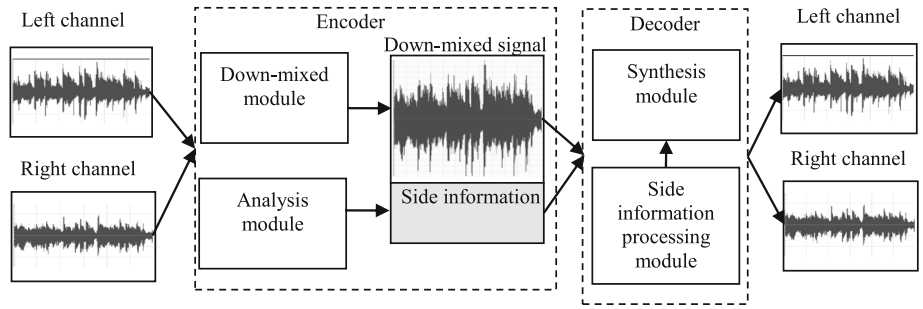

Fig. 1 Block diagram of the BCC

process and side information, the down-mixed module, analysis module, side information processing module, and synthesis module are considered. At the encoder, the signal of left channel and signal of right channel are down-mixed by the down-mixed module for producing a mono signal, and analyzed by the analysis module for generating the inter-channel cues. The down-mixed mono signal is the sum of the signals coming from the left and right channels. The binaural inter-channel cues, ICLD, ICC, and inter-channel time difference (*ICTD*) [30, 42] are extracted respectively as follows:

$$\mathrm{ICLD}(i,l) = 10\ \log 10\left[\frac{\sum_{n=0}^{N-1}\left|X_{il,L}(n)\right|^2}{\sum_{n=0}^{N-1}\left|X_{il,R}(n)\right|^2}\right] \tag{1}$$

$$\mathrm{ICC}(i,l) = \frac{\left|\sum_{n=0}^{N-1} X_{il,L}(n)X_{il,R}(n)\right|}{\sqrt{\sum_{n=0}^{N-1}\left|X_{il,L}(n)\right|^2}\sqrt{\sum_{n=0}^{N-1}\left|X_{il,R}(n)\right|^2}} \tag{2}$$

$$\mathrm{ICTD}(i,l) = \underset{\delta}{\mathrm{argmax}}\frac{\sum_{n=0}^{N-1} X_{il,L}(n)X_{il,R}(n-\delta)}{\sqrt{\sum_{n=0}^{N-1}\left|X_{il,L}(n)\right|^2}\sqrt{\sum_{n=0}^{N-1}\left|X_{il,R}(n-\delta)\right|^2}} \tag{3}$$

where N is frame length and δ denotes time delay. $X_{il,\ L}$ and $X_{il,\ R}$ indicate left channel and right channel signals at the ith sub-channel of the lth frame, respectively.

At the decoder, the side information processing module is used to extract the binaural inter-channel cues, which are combined with the down-mixed signal to obtain the left channel signal and right channel signal by the synthesis module, that is, given the T-F components $Y_{il}$of mono signal, the T-F components $X_{il,\ L}$ of left channel can be calculated as [42]

$$X_{il,L}(i,l) = F_L(i,l)\cdot G_L(i,l)\cdot Y_{il} \tag{4}$$

Similarly, we can obtain the T-F components of the right channel. In Eq. (4), $G_L(i,l)$ denotes the phase modification function related to time delay. In this paper, the matching data set between speech and noise is used in training, so when the speech and noise are viewed as left channel and right channel signals, the time delay may be neglected so that it is not concerned in this paper. $F_L(i, l)$ is used to determine the amplitude modification of the T-F component, which depends on the ICLD and ICC. For convenience, the frame index l in the following equations is omitted. Thus, the $F_L(i)$ can be obtained by the inter-channel cues as follows:

$$F_L(i) = 10^{(\mathrm{ICLD}(i)+r(i))/20}\cdot F_R(i) \tag{5}$$

with

$$F_R(i) = \frac{1}{\sqrt{1+10^{(\mathrm{ICLD}(i)+r(i))/10}}} \tag{6}$$

$$r(i) = [1-\mathrm{ICC}(i)]\cdot\tau(i) \tag{7}$$

where $\tau(i)$ is the random Gaussian sequence with zero mean and variance 1.

2.2 The transfer from the BCC to speech enhancement

In the original BCC framework [41, 42], the down-mixed mono signal is the sum of the collected signals from the left channel and right channel. For additive noise, since noisy speech is the sum of speech and noise signals, when speech and noise signals are regarded as the left channel signal and right channel signal, respectively, the noisy speech can be seen as a down-mixed mono signal composed of speech and noise. At this time, in order to observe the transfer from the original BCC framework to speech enhancement, the clean speech and noise signals are in the two separate channels to simulate the down-mixed process. Thus, the BCC shown in Fig. 1 can be transferred to a speech enhancement framework shown in Fig. 2. In Fig. 2, the functions of the

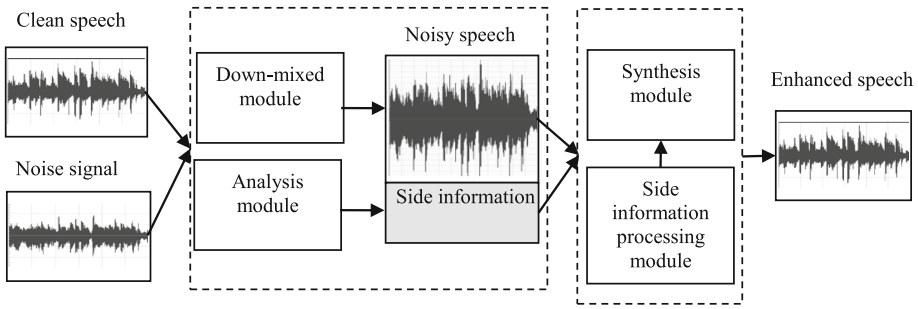

Fig. 2 Block diagram of speech enhancement framework based on the BCC

four modules (i.e., the down-mixed module, analysis module, side information processing module, and synthesis module) are the same as those in Fig. 1. From Fig. 2, we can see that once the inter-channel cues between speech and noise are given, the clean speech can be separated from noisy speech signal by using decoding principle of the BCC. This speech enhancement framework is based on the studies in [44–46], that is, there is a certain correlation between speech and noise components in time-frequency domain. Moreover, the studies in references [47] to [48] further confirmed that the level of normalized cross-correlation coefficient (NCCC) between noisy speech and noise approximates to 0.5 in voiced segments, which also implies a relatively strong correlation between speech and noise. So, we attempt to exploit level modifications of speech and noise to generate a ratio mask based on the correlation between speech and noise for speech restoration.

3 The proposed method based on binaural inter-channel cues

In the proposed speech enhancement method, the exact inter-channel cues between speech and noise are considered as a priori information of speech and noise. By extracting exact inter-channel cues between speech and noise, the T-F mask exploiting the BCC decoding can be obtained for speech restoration. In order to get exact inter-channel cues between speech and noise, two different techniques are considered in this paper. In the first technique that will be described in Section 3.1, the exact inter-channel cues and pre-enhanced inter-channel cues are trained as a priori codebook offline, and the estimations of the exact inter-channel cues are obtained online by a weighted mapping technique [43] for generating the T-F mask, which is termed as the codebook-based T-F mask estimation method. In the first technique, the selection of code vectors or the calculation of weights is liable to lead to errors, which directly affects the estimation accuracy of the exact inter-channel cues. Thus, for the second technique given in Section 3.2, we plan to use the DNN to directly predict exact inter-channel cues between speech and noise in the beginning. However, the study in [17] has shown that the compression training of the unbounded predictive target (e.g., ICLD) is detrimental to the supervised approaches based on the gradient descent [20]. So, in the second technique, the DNN is used to directly predict the T-F mask constructed by exact inter-channel cues, which is termed as the DNN-based T-F mask estimation method.

3.1 Codebook-based T-F mask estimation

Considering that the cocleagram is more separable than spectrogram [49], the proposed method is operated in the Gammatone auditory domain. In order to use the BCC framework to separate speech from noisy speech, the pre-enhancing technique [5] is used for initial estimation of speech and noise in order to modify the levels of speech and noise (Fig. 3).

Figure 3 shows a block diagram of codebook-based T-F estimation. The T-F mask estimation consists of two stages, i.e., the training phase and the enhancing phase. In the training phase, the noisy speech is firstly pre-enhanced to obtain pre-enhanced speech and estimated noise for generating the pre-enhanced inter-channel cues. Here, the purpose of using pre-enhancement is to ensure data matching between the enhancement and training phases. Next, clean speech, noise, pre-enhanced speech, and estimated noise are decomposed into T-F units by a Gammatone filter with 64 channels. Wherein, the frame length of each channel is 32 ms and overlapped by 16 ms. It must be noted that the pre-enhanced inter-channel cue vector θ_y is obtained from pre-enhanced speech and estimated noise, and the exact inter-channel cue vector θ_x is extracted from clean speech and noise. θ_y and θ_x are combined one by one to train a vector codebook. In the enhancing phase, the pre-enhanced

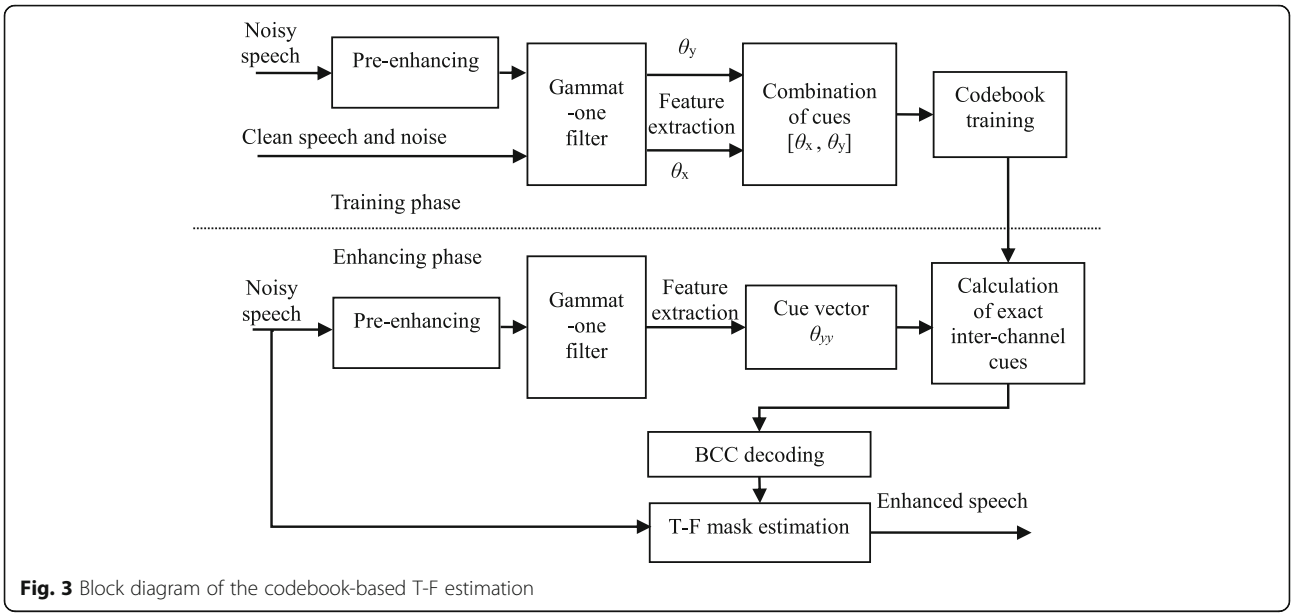

Fig. 3 Block diagram of the codebook-based T-F estimation

cue vector θ_{yy} is extracted from noisy speech. The weighted mapping technique [43] is used to obtain a weighted sum of the pre-enhanced cues chosen from the trained codebook. According to one-to-one relationship of θ_y and θ_x, the exact inter-channel cues are estimated. Given the estimated θ_x, the T-F mask can be obtained from the BCC decoder. With the estimated T-F masks, the speech signal is reconstructed.

In order to perform speech enhancement, there are three key problems to be solved. One is how to define exact inter-channel cues. The second one is how to estimate exact inter-channel cues from the trained codebook. The last one is how to obtain the T-F mask based on exact inter-channel cues.

3.1.1 Definition of the inter-channel cues

In order to facilitate symbol distinction of the cues between speech enhancement system and the BCC system, in this paper, the cue ICLD defined in the BCC is redefined as level difference of speech and noise (LDSN), and the cue ICC defined in the BCC is redefined as speech and noise correlation (SNC). For each frequency sub-band of each frame, the LDSN and SNC are denoted as exact inter-channel cues and calculated respectively by:

$$\mathrm{LDSN}(i) = 10\ \log 10\left[\frac{\sum_{n=0}^{N-1}|x_i(n)|^2}{\sum_{n=0}^{N-1}|w_i(n)|^2}\right] \tag{8}$$

and

$$\mathrm{SNC}(i) = \frac{\left|\sum_{n=0}^{N-1} x_i(n)w_i(n)\right|}{\sqrt{\sum_{n=0}^{N-1}|x_i(n)|^2}\sqrt{\sum_{n=0}^{N-1}|w_i(n)|^2}} \tag{9}$$

where N is frame length, i is the frequency sub-band index, and $x(n)$ and $w(n)$ denote clean speech and noise signals, respectively. For the LDSN, it is actually equivalent to the sub-band SNR in [50]. Obviously, in this paper, the SNC is computed in each frequency sub-band so that the shared correlation of multiple frequency sub-bands in [26] can be avoided.

For the calculation of the pre-enhanced inter-channel cues, the pre-enhanced speech $x_p(n)$ and pre-enhanced noise $w_p(n)$ are firstly obtained with the pre-enhanced method [5]. Then, pre-enhanced speech and estimated noise are decomposed into the sub-band signals [51] by the Gammatone filter, respectively. For the frequency sub-band i of frame l, the pre-enhanced LDSN is calculated in time-frequency domain with the following steps [22]:

(1) Compute the DFT of $x_p(n)$ and $w_p(n)$ as $X_p(k) = X_{p,k} \cdot \exp(j\phi_y(k))$ and $W_p(k) = W_{p,k} \cdot \exp(j\phi_y(k))$, where $X_{p,k}$ and $W_{p,k}$ denote the magnitude spectra of $x_p(n)$ and $w_p(n)$. $\phi_y(k)$ is the phase of the corresponding noisy speech.
(2) Estimate LDSN: $\varepsilon_k(l) = \alpha \cdot \varepsilon_k(l-1) + (1-\alpha) \cdot (X_{p,k})^2/(W_{p,k})^2$, where α is the weighting parameter and $\varepsilon_k(l-1)$ is the SNR value of previous frame.

Since the calculation of the LDSN concerns a high-dimension vector caused by the DFT for each frequency

sub-band, the dimension number is reduced to 10 from 256 by a similar polynomial model [22] in this paper.

The pre-enhanced SNC is defined in the critical bands and calculated as follows:

$$\mathrm{SNC}(b) = \frac{\left|\sum_{k=A_b}^{A_{\mathrm{up}}(b)} X_p(k)\cdot W_p(k)\right|}{\sqrt{\left(\sum_{k=A_{\mathrm{low}}(b)}^{A_{\mathrm{up}}(b)} X_p(k)\cdot X_p^*(k)\right)\left(\sum_{k=A_{\mathrm{low}}(b)}^{A_{\mathrm{up}}(b)} W_p(k)\cdot W_p^*(k)\right)}} \tag{10}$$

where $b \in [0, 18)$ is the critical band index and $A_{\mathrm{up}}(b)$ and $A_{\mathrm{low}}(b)$ are the upper and lower frequency bound of the bth critical band [52], respectively.

Thus, for the pre-enhanced signals, the LDSN cue and SNC cue are represented by 10 dimensional vector and 18 dimensional vector, respectively. Combining 10 LDSN cues and 18 SNC cues, we can build a vector with 28 dimensions to generate the pre-enhanced cues.

3.1.2 Estimation of exact inter-channel cues

In this part, the weighted codebook mapping (WCBM) algorithm [43] is selected to obtain exact inter-channel cues from the trained codebook. In the trained stage, the exact inter-channel cues and the pre-enhanced inter-channel cues are combined one by one and trained as a codebook, that is, for each code-vector of the trained codebook, it consists of exact cues with a 2-D vector θ_x and the pre-enhanced cues with a 28-D vector θ_y. Here, the exact cue vector θ_x = [LDSN, SNC] and offline pre-enhanced cue vector θ_y = [LDSN_{y0}, LDSN_{y1},..., LDSN_{y9}, SNC_{y0}, SNC_{y1},..., SNC_{y17}], where the [LDSN_{y0}, LDSN_{y1}, ..., LDSN_{y9}] is a vector with 10 dimensions obtained by dimension reduction of the pre-enhanced LDSN based on a polynomial model [22] and the [SNC_{y0}, SNC_{y1},..., SNC_{y17}] is a vector with 18 dimensions obtained from 18 critical band of the pre-enhanced SNC based on Eq.(10). In this paper, the size of codebook is Q = 256, i.e., the codebook is comprised of 256 code-vectors. Figure 4 gives the block diagram of extracting exact inter-channel cues from the trained codebook. For each frequency sub-band of each frame, the pre-enhanced cue vector θ_{yy} is extracted from noisy speech in the enhanced stage. Here, online pre-enhanced cue vector θ_{yy} = [LDSN_{yy0}, LDSN_{yy1},..., LDSN_{yy9}, SNC_{yy0}, SNC_{yy1},..., SNC_{yy17}] is similar to offline pre-enhanced cue vector θ_y. However, different from the [LDSN_{y0}, LDSN_{y1},..., LDSN_{y9}] and [SNC_{y0}, SNC_{y1},..., SNC_{y17}] generated from the pre-enhanced speech and noise of the trained stage, the [LDSN_{yy0}, LDSN_{yy1},..., LDSN_{yy9}] and the [SNC_{yy0}, SNC_{yy1},..., SNC_{yy17}] are generated from the pre-enhanced speech and noise of the enhanced stage. Then, by comparing Euclidean distance (ED) between the θ_{yy} and the pre-enhanced cue vector θ_y stored in the trained codebook, we only choose M code-vectors with relatively smaller ED from Q code-vectors [26]. The pre-enhanced cue vectors of the M code-vectors are defined as θ_{y1}, θ_{y2}, ..., θ_{yM}. In order to obtain the weights of M code-vectors [26], the qth ED between θ_{yq} and θ_{yy} is first given as follows (Fig. 4):

$$E_d(q) = \mathrm{sqrt}\left(\sum_{j=1}^{28}\left(\theta_{yq}(j)-\theta_{yy}(j)\right)^2\right) \qquad \forall q\in[1,Q] \tag{11}$$

Then, the EDs corresponding to M code-vectors are expressed as $E_d(1)$, $E_d(2)$, ... , $E_d(M)$. And the degree ρ_m of the mth member of M code-vectors is defined as follows [26]:

$$\rho_m = \left[\frac{(E_d(m))^2}{\sum_{m=1}^{M}(E_d(m))^2}\right]^{-1} \qquad \forall m\in[1,M] \tag{12}$$

Subsequently, the mth weight of M code-vectors is given by:

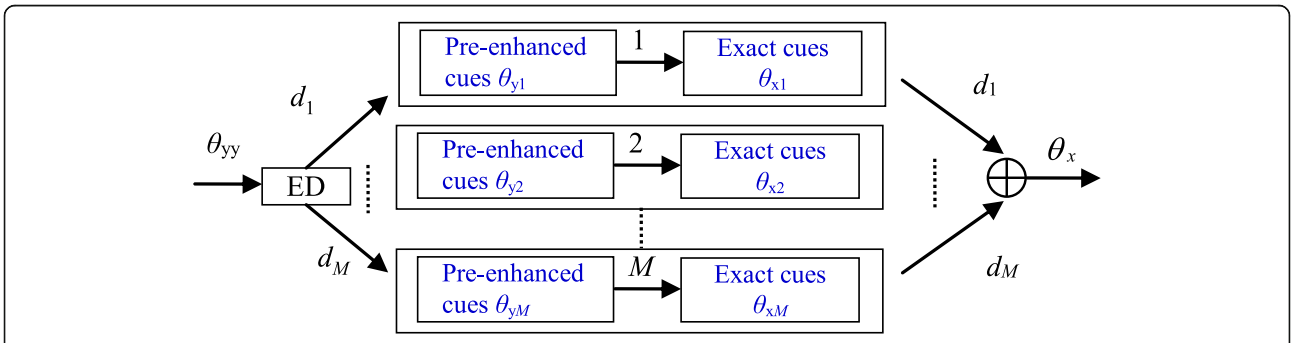

Fig. 4 Block diagram of extracting exact inter-channel cues from the codebook

$$d_m = \frac{\rho_m}{\sum_{m=1}^{M} \rho_m} \tag{13}$$

According to one-to-one mapping between exact cues and pre-enhanced cues in each code-vector, all weights derived by the Eq. (13) can be used to estimate exact cues as follows:

$$\hat{\theta}_x = \sum_{m=1}^{M} d_m \cdot \theta_{x,m} \tag{14}$$

where $\theta_{x,\ m}$ is the mth vector of M exact cue vectors chosen from M code-vectors.

3.1.3 T-F Mask estimation based on exact inter-channel cues

Once the exact cues including LDSN and SNC are estimated in the frequency sub-band i, by exploiting the BCC decoding principle, we have:

$$\mathrm{LDSN1}(i) = \mathrm{LDSN}(i) + r(i) \tag{15}$$

with

$$r(i) = [1 - SNC(i)] \cdot \tau(i) \tag{16}$$

Finally, with exact inter-channel cues obtained above, the proposed ratio mask based on the cues (RMC) is given as follows:

$$RMC(i) = \left(\frac{F_x(i)}{F_x(i) + F_w(i)}\right)^{1/2} \tag{17}$$

where based on amplitude modifications of left channel and right channel signals from the Eq.(5) and Eq.(6), $F_x(i)$and $F_w(i)$are defined as the energy modifications of speech and noise, respectively. They are given by:

$$F_w(i) = \left(\frac{1}{\sqrt{1 + 10^{\mathrm{LDSN1}(i)/10}}}\right)^2 \tag{18}$$

$$F_x(i) = \left(10^{\mathrm{LDSN1}(i)/20} * \frac{1}{\sqrt{1 + 10^{\mathrm{LDSN1}(i)/10}}}\right)^2 = \frac{10^{\mathrm{LDSN1}(i)/10}}{\left(1 + 10^{\mathrm{LDSN1}(i)/10}\right)} \tag{19}$$

3.2 DNN-based T-F mask estimation

In the aforementioned codebook-based method and the studies [26, 27], some intermediate parameters (e.g., code-vector selection or exact cue estimation) need to be selected for generating ratio mask or speech gain. However, it is not easy to ensure a high accurate degree in obtaining these intermediate parameters. Furthermore, the performance of the pre-enhancing module may limit the improvement of speech quality for these methods. In order to solve the problems, we attempt to reduce intermediate link of estimating ratio mask by directly constructing a mapping relationship between noisy speech and the ratio mask constructed by exact inter-channel cues. As we all know, the DNN has a very good learning ability to fit the mapping function between the input features and training targets. Thus, the DNN is investigated in this sub-section to learn the mapping relationship between input features of noisy speech and the proposed ratio mask.

Figure 5 shows a block diagram of the DNN-based T-F estimation. It also contains two stages. One is the offline training stage, and another one is the online enhancing stage. In the training stage, with the Gammatone filter, the T-F representation of the speech and noise are obtained. Then, the exact inter-channel cues are extracted to generate a T-F mask. For the acoustic features in the input of the DNN, a set of robust features [23] (i.e., amplitude modulation spectrogram (AMS), relative spectral transform and perceptual linear prediction (RASTA-PLP), Mel frequency cepstral coefficients (MFCC), and Gammatone filterbank power spectra (GF)) are obtained from noisy speech.

For the DNN-based methods, the training target plays a very important role on the performance of speech restoration. The ideal ratio mask (IRM) [23] is commonly used as the training target. However, it does not focus on the correlation between noise and clean speech. In this paper, the T-F mask obtained by level modifications of speech and noise incorporates the correlation between noise and clean speech. In constructing the correlation, considering that the introduction of the random number $\tau(i)$ may weaken the periodicity of speech to some extent, a multiplicative combination between the proposed RMC (i.e., Eq. (17)) and the IRM [23] is used as the desired output of the DNN. The combined mask (CM) is represented as follows:

$$\mathrm{CM}(i) = \mathrm{RMC}(i) \cdot \mathrm{IRM}(i) \tag{20}$$

In the enhancing stage, the well-trained DNN can be seen as a non-linear mapping function to directly predict the proposed T-F mask given input features of noisy speech. With the masking estimation, the clean speech can be separated from noisy speech (Fig. 5).

4 Experimental results

In this section, we attempt to give some experiments to evaluate the performance of the proposed scheme. In the experiments, three reference methods are considered, namely the pre-processed method [5] is selected as the first reference method (named as Ref.1), a codebook-based speech gain estimator [26] is considered as the second

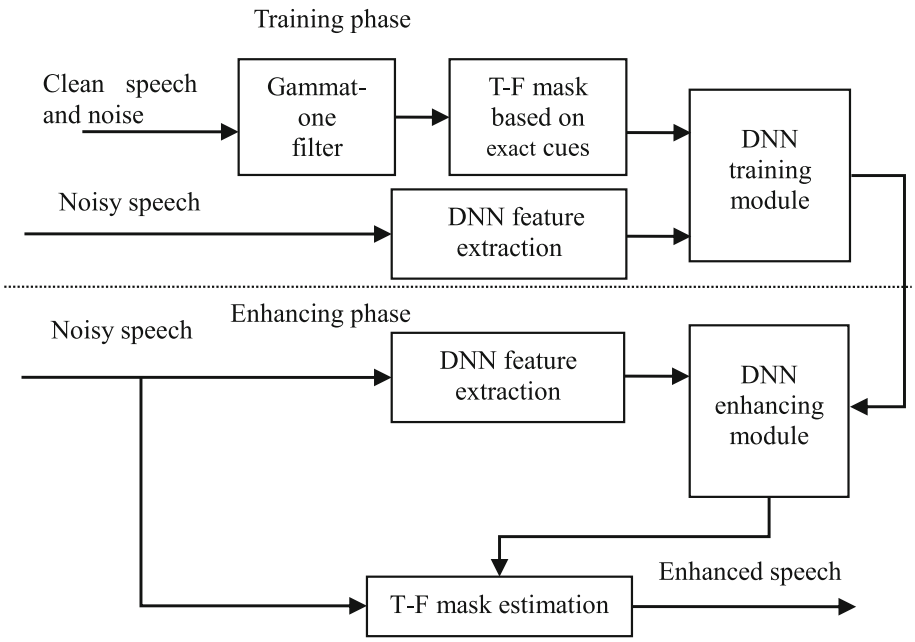

Fig. 5 Block diagram of the DNN-based T-F estimation

reference method (named as Ref.2), and a classical DNN-based IRM method [23] is considered as the third reference method (named as Ref.3). Our codebook-based technique is named as ProC, and the DNN-based technique is named as ProN. For Ref.2, it used the inter-channel cues and speech presence probability (SPP) to obtain spectral gain of speech. Moreover, in the Ref.2 and the proposed codebook-based technique, the pre-processed module is MMSE spectral amplitude (MSA) method [5] and the estimated noise is obtained from the minima controlled recursive averaging (MCRA) method [53] in the pre-processed module. The 8-bit inter-channel cue codebook was trained by LBG algorithm [13]. The number M of candidate code-vectors is 50. Considering that the minimum statistics (MS) method in [9] is a more robust approach in different SNR conditions [54], a method of combining the MSA and MS is also used as the reference method (named as Ref.4) to observe the robustness of the proposed method. For the Ref.3 and the proposed DNN-based technique, a set of the features (i.e., AMS, RASTA-PLP, MFCC, and GF) were used as input features of the DNN. In the training stage, 2 h of utterances from different talker were selected from TIMIT training set [55]. The speech signal was down-sampled to 8 kHz. Four different types of background noises (i.e., white noise, babble noise, f16 noise, and factory noise) were chosen from NOISEX-92 databases [56]. Aside from the aforementioned four training noises, two types of the unseen noises (i.e., factory2 noise and street noise) were used for mismatch evaluation. In the test stage, our method was evaluated with 240 noisy speech signals composed of 40 clean speech signals from the TIMIT test set [55] mixed with six noises for each input signal-to-noise ratio (SNR) condition. The input SNR is set to − 5dB, 0 dB, 5 dB, and 10 dB, respectively. The frame size is 32 ms (256 samples) with 50% overlap. For the DNN model in the Ref.3 and ProN, the four hidden layers (each with 1024 nodes) with sigmoid activation functions were used in the DNN model. The backpropagation algorithm with dropout regularization (dropout rate 0.2) was used to train the networks. The adaptive gradient descent along with a momentum term was used as the optimization technique. The momentum rate is 0.5 for the first 5 epochs and 0.9 for the rest epochs. The mean squared error was used as the cost function for the DNN training. The number of output units corresponds to the dimensionality of the training target. Some evaluations are performed for speech enhancement as follows.

4.1 The comparison of the RMC and combined mask

In this section, the performances of the RMC and combined mask (named as ProN) are discussed, when they are used as the desired output of the DNN. For the multiplicative combination between two masks, the study in [18] has shown that the combined mask can reduce the disadvantages of their respective masks. In the proposed DNN-based method, considering that the introduction of the random number $\tau(i)$ in the RMC may weaken the periodicity of speech to some extent, we select a multiplicative combination between RMC and the IRM [23] as the desired output of the DNN. Table 1 lists the average PESQ and STOI scores of the RMC and ProN for six noises. From Table 1, the RMC and ProN can all obtain better PESQ and STOI results

Table 1 Comparison on average STOI and PESQ

Methods		− 5 dB	0 dB	5 dB	10 dB
Noisy	PESQ	1.53	1.85	2.16	2.50
	STOI	0.6407	0.7522	0.8336	0.8793
RMC	PESQ	2.18	2.53	2.86	3.13
	STOI	0.7426	0.8064	0.8629	0.9094
ProN	PESQ	2.25	2.60	2.91	3.17
	STOI	0.7431	0.8068	0.8633	0.9097

than noisy speech. As a comparison, the combined mask gives a higher PESQ and STOI values than the RMC. This confirms that the multiplicative combination can help to improve the prediction ability of the RMC.

4.2 The comparison between the proposed method and other methods

4.2.1 The SSNR evaluation

The segmental signal-to-noise ratio (SSNR) [8] is often applied to evaluate the de-noising performance of speech enhancement method. It is defined as follows:

$$\mathrm{SSNR} = \frac{1}{N_{\mathrm{um}}} \sum_{j=1}^{N_{\mathrm{um}}} 10 \log_{10} \left(\frac{\sum_{n=1}^{N} x^2(n)}{\sum_{n=1}^{N} [x(n) - \hat{x}(n)]^2} \right) \quad (21)$$

where N_{um} is the number of frames, N is the length of frame, $x(n)$ is clean speech signal, and $\hat{x}(n)$ is the enhanced speech signal. Higher values of SSNR improvement is an indication of higher speech quality.

Table 2 shows the SSNR improvement (SSNRI) obtained by processing noisy signals with the proposed methods and reference methods at different input SNRs for the white, babble, F16, and factory noises, respectively. As seen from Table 2, the enhanced speech signals from the Ref.1 and Ref.4 get relatively lower SSNRI results than the other methods. The reason for this case may be the inaccurate estimation of noise power spectrum, which could result in more residual noise in the enhanced speech. Compared with the Ref.4, the Ref.1 can suppress more noise, which is similar to the case in [57].

For the Ref.2 algorithm considering the shared correlation from multiple frequency bands, it views the inter-channel cues with shared correlation as the prior information of speech and noise and generates effective performance in reducing noise. Moreover, we can find that the DNN-based Ref.3 method can consistently generate higher SSNRI results than the Ref.1, Ref.2, and Ref.4. In the proposed codebook-based technique, given the mono noisy signal, we can use the T-F mask based on the LDSN and SNC to estimate clean speech and obtain

Table 2 Test results of SSNR improvement

Noise type	Input SNR	Methods					
		Ref.1	Ref.2	Ref.3	Ref.4	ProC	ProN
White	− 5 dB	11.15	13.081	13.72	11.09	13.77	16.01
	0 dB	9.69	11.10	12.56	9.61	12.07	14.05
	5 dB	7.97	9.25	10.73	7.93	10.06	11.97
	10 dB	6.35	7.41	8.67	6.33	8.07	10.01
Babble	− 5 dB	9.71	10.10	13.42	9.65	10.60	15.05
	0 dB	8.58	9.29	11.87	8.49	9.59	13.55
	5 dB	7.04	8.03	10.05	6.96	8.24	11.21
	10 dB	5.62	6.05	7.93	5.33	6.60	9.02
F16	− 5 dB	10.78	12.59	14.01	10.65	14.34	16.14
	0 dB	9.45	11.05	12.60	9.33	12.07	14.32
	5 dB	7.83	9.93	10.57	7.76	11.60	12.11
	10 dB	6.19	7.75	8.79	6.11	8.86	10.12
Factory	− 5 dB	10.50	12.01	13.29	10.39	12.88	15.88
	0 dB	8.47	10.32	11.99	8.38	11.74	13.97
	5 dB	6.73	8.36	10.02	6.65	9.30	11.55
	10 dB	4.88	6.49	8.41	4.81	7.48	9.33

a better performance than the Ref.1, Ref.2, and Ref.4. However, the ProC achieves a poorer result than Ref.3 and Ref.N because of the introduction of the errors from the intermediate link between noisy feature and the T-F mask. From the results of the SSNRI, with the proposed DNN-based framework, we can reduce more residual noise than the Ref.3 in all conditions.

For each input SNR, the average SSNRI values of different methods for four types of noise are presented in Fig. 6. From Fig. 6, we can find that the average SSNRIs of the Ref.1 and Ref.4 are relatively lower than that of the supervised methods. In comparison, the proposed DNN system and Ref.3 get much higher average SSNR

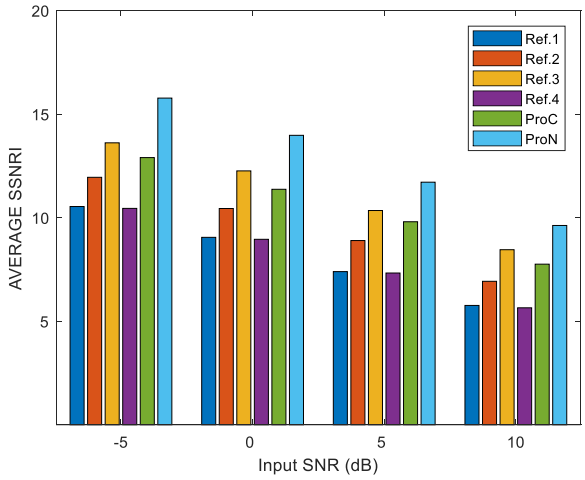

Fig. 6 Average test results of the SSNRI

improvement than the Ref.2 and ProC, which may mean that the reduction of the intermediate link between noisy feature and T-F mask could help to suppress background noise. Moreover, we can see that the proposed ProN method performs better than the Ref.3 in reducing noise.

4.2.2 The PESQ evaluation

The perceptual evaluation of speech quality (PESQ) [58] is an objective evaluation of speech quality, which is often used to evaluate the quality of the restored speech. The higher the PESQ, the better the speech quality.

Table 3 shows the PESQ results for the proposed method and reference methods at different noisy conditions and input SNR levels. From Table 3, we can see that the Ref.1 method is slightly better than the Ref.4 in most SNR cases because of good preference of the MCRA compared with the MS [59]. The Ref.1 and Ref.4 generate poorer PESQ results than the supervised methods. For the Ref.2 and ProC algorithm, in order to obtain spectral gain and T-F mask, several appropriate code-vectors termed as the intermediate parameters and the pre-enhanced module were concerned. The estimation error of these intermediate parameters and an inappropriate pre-enhanced method may enlarge the inaccuracy of estimating spectral gain and T-F mask so that the Ref.2 and ProC have more insufficient ability to cope with speech restoration compared to the Ref.3 and ProN methods. In addition, the DNN-based methods, Ref.3 and ProN, outperform the codebook-driven method since the DNN could more effectively model nonlinear interaction between training target and the acoustic features of noisy speech than vector quantization. In the ProN method, the level modifications including the correlation between speech and noise are used to generate a training target. Compared with the Ref.3 without correlation between speech and noise, the proposed ProN can provide better speech quality in terms of the PESQ.

Table 3 Test results of PESQ

Noise type	Input SNR	Methods						
		Noisy	Ref.1	Ref.2	Ref.3	Ref.4	ProC	ProN
White	− 5 dB	1.206	1.312	1.559	2.003	1.301	1.724	2.111
	0 dB	1.410	1.640	1.875	2.308	1.635	2.067	2.377
	5 dB	1.646	2.085	2.329	2.586	2.084	2.354	2.701
	10 dB	1.977	2.467	2.527	2.859	2.491	2.612	3.002
Babble	− 5 dB	1.360	1.430	1.650	1.981	1.398	1.699	2.080
	0 dB	1.630	1.895	2.019	2.270	1.854	2.075	2.381
	5 dB	2.010	2.308	2.427	2.572	2.289	2.481	2.682
	10 dB	2.390	2.607	2.709	2.859	2.601	2.717	2.992
F16	− 5 dB	1.293	1.355	1.615	2.049	1.301	1.7113	2.112
	0 dB	1.569	1.860	2.049	2.362	1.833	2.143	2.432
	5 dB	1.908	2.289	2.403	2.651	2.286	2.495	2.777
	10 dB	2.218	2.640	2.759	2.939	2.643	2.785	3.101
Factory	− 5 dB	1.179	1.340	1.505	2.010	1.299	1.518	2.109
	0 dB	1.470	1.820	1.941	2.303	1.787	2.033	2.412
	5 dB	1.820	2.280	2.385	2.582	2.276	2.450	2.711
	10 dB	2.237	2.534	2.623	2.869	2.533	2.742	3.012

From the average PESQ given in Fig. 7, we can see that, by comparing with the PESQ of noisy speech, all methods can improve the PESQ result to some extent. Furthermore, the proposed ProN method produces the highest average PESQ scores than the other methods in different input SNR conditions.

4.2.3 The speech intelligibility test

The short-time objective intelligibility (STOI) [60] is used to evaluate our system and reference methods for the intelligibility. The STOI is shown to be highly correlated to human speech intelligibility. Table 4 gives the average STOI comparison of different methods under different input SNR conditions. As shown in Table 4, compared with the noisy speech, the Ref.1, Ref.2, and Ref.4 methods do not consistently improve STOI results. Because the MCRA has a higher estimation error compared with the MS under high SNR conditions [7, 57], it may cause speech distortion and make the Ref.4 outperform the Ref.1 in relatively high SNR cases. Similar to the results of the SSNRI and PESQ, the Ref.3 and ProN algorithms reducing intermediate link between noisy feature and T-F mask can achieve higher results in all cases compared with the Ref.1, Ref.2, and RroC. In the Ref.3, IRM is used as the training target, which does not consider the correlation between noise and clean speech. As a comparison, the proposed ProN system gives a relatively higher average STOI value than the Ref.3 at different input SNR conditions, which may mean that incorporating the correlation between noise and speech into the T-F mask could help to improve speech intelligibility of the enhanced speech.

4.2.4 The speech spectrogram comparison

In this subsection, in order to describe the details and structure of speech, we give the speech spectrograms of the enhanced speech obtained by the proposed methods and reference methods. In this part, Fig. 8 provides the speech spectrograms of noisy speech (speech signal is mixed with babble noise at 0 dB input SNR) and the enhanced speech generated by the various methods. From the Fig. 8, we can see that the main structure of speech signal can be recovered by all algorithms, compared with the structure of noisy speech. For the Ref.1 and Ref.4, some speech regions are discarded and more residual noise is retained in the enhanced speech. The reason is

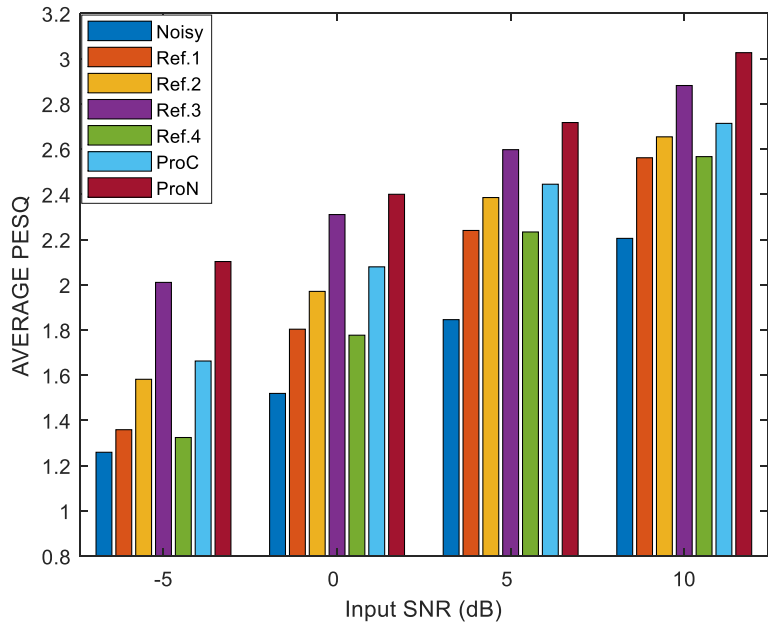

Fig. 7 The average test results of PESQ

likely that the spectral information belonging to the parts of speech is suppressed, and noise spectral information is retained because of inaccurate noise estimation.

The Ref.2 uses a codebook-based technique to obtain exact inter-channel cues for generating a spectral estimator of speech and can suppress more speech energy loss to some extent than the Ref.1 method. Moreover, the Ref.2 also achieves a good performance in noise reduction. Compared with the Ref.2, more residual noise is reduced in the enhanced speech obtained by the proposed ProC method, which may show that the inter-channel cues without the shared correlation can more effectively restore clean speech signal from noisy speech.

For the Ref.3 and ProN algorithms, the DNN model is used to model the nonlinear interaction between the T-F mask and the acoustic features of noisy speech. Compared with the Ref.1, Ref.2, and ProC methods, the Ref.3 and ProN can remove more background noise and maintain more speech energy, for example, at about the 1.3 s, more harmonic structure is retained. However, there are still more speech distortions in enhanced speech processed by two DNN-based methods, especially in the lower speech energy area, for example, between 1.5 and 1.7 s, more speech energy is seen as noise to be suppressed. This may be because of the limited training set and the limited learning ability of the DNN model. Thus, it may have a potential way to reduce the problem of speech distortion in future work, when long short-term memory model (LSTM) with temporal dependencies and large-scale training with many speakers and numerous noises are considered [25]. As a comparison, the Ref.3 and ProN can achieve similar performance in the speech restoration. However, more background noise can be suppressed in the proposed ProN method because of the consideration of the correlation between speech and noise.

Table 4 Comparison on average STOI

Enhancement methods	− 5 dB	0 dB	5 dB	10 dB
Noisy	0.6056	0.7394	0.8318	0.8776
Ref.1	0.6238	0.7411	0.8221	0.8545
Ref.2	0.6425	0.7589	0.8309	0.8767
Ref.3	0.7384	0.8050	0.8632	0.9115
Ref.4	0.6233	0.7399	0.8221	0.8548
ProC	0.6441	0.7631	0.8335	0.8826
ProN	0.7407	0.8062	0.8685	0.9156

4.2.5 Noise generalization ability test

In order to measure the robustness of noise environment of the proposed method, two types of the unseen noises (i.e., factory2 noise and street noise) are used for mismatch evaluation. Table 5 lists the average results of PESQ and STOI of the unseen noises for different methods under different input SNR conditions. The enhanced speech processed by the Ref.1 and Ref.4 methods can obtain better PESQ results than the noisy speech. For the Ref.2 and ProC, the ability to handle unseen noise is not weakened to a certain extent probably because of the usage of an unsupervised pre-enhanced

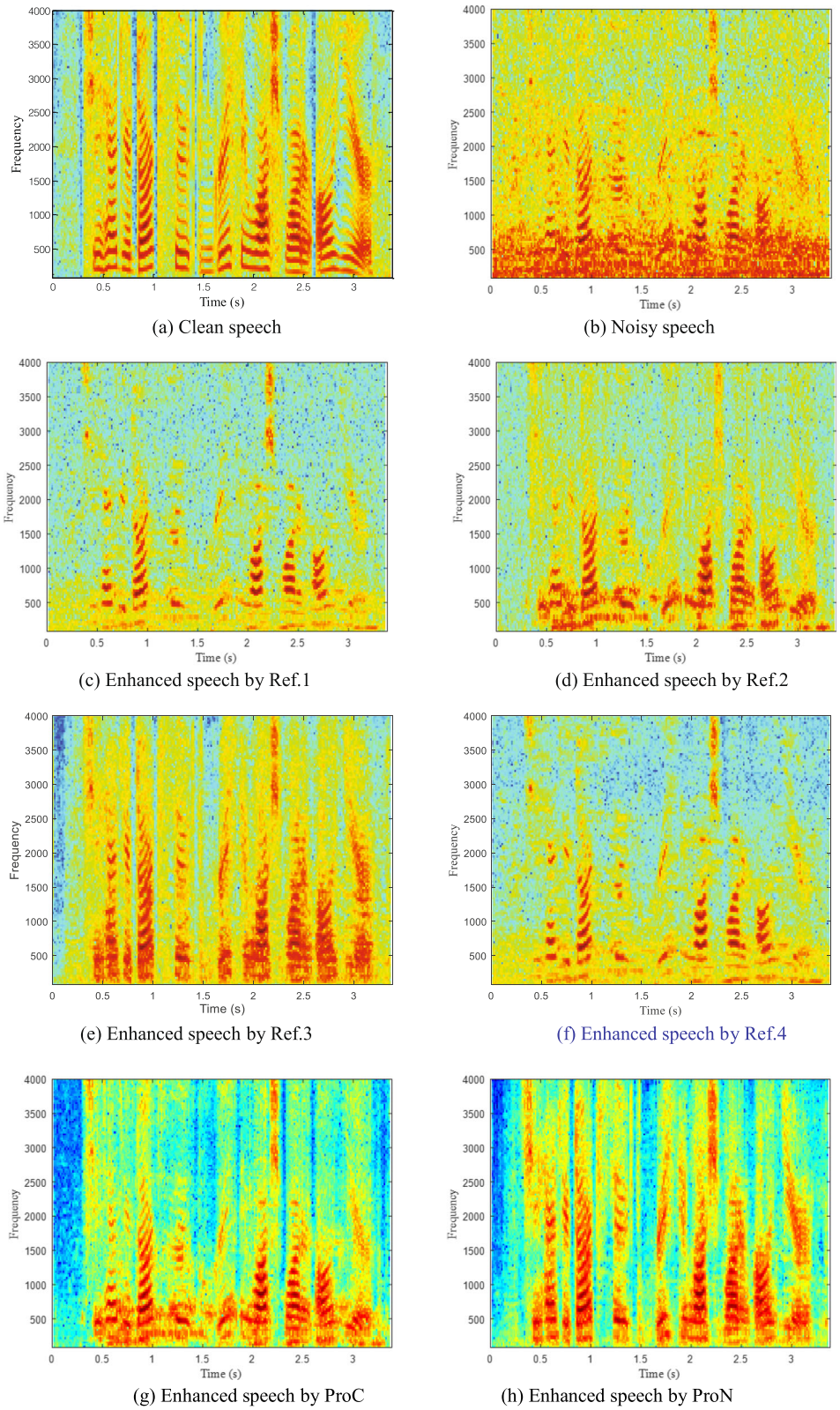

Fig. 8 Spectrogram comparison of different methods. **a** Clean speech. **b** Noisy speech. **c** Enhanced speech by Ref.1.**d** Enhanced speech by Ref.2. **e** Enhanced speech by Ref.3. **f** Enhanced speech by Ref.4. **g** Enhanced speech by ProC. **h** Enhanced speech by ProN

Table 5 Comparison on average STOI and PESQ

Methods		− 5 dB	0 dB	5 dB	10 dB
Noisy	PESQ	1.8146	2.1822	2.4782	2.8086
	STOI	0.6758	0.7651	0.8355	0.8811
Ref.1	PESQ	2.0631	2.4202	2.7037	2.9561
	STOI	0.6831	0.7462	0.8025	0.8516
Ref.2	PESQ	2.2145	2.5300	2.7873	3.0683
	STOI	0.6892	0.7471	0.8066	0.8615
Ref.3	PESQ	2.3813	2.7855	3.0747	3.3164
	STOI	0.7389	0.8055	0.8572	0.9033
Ref.4	PESQ	2.0589	2.4178	2.6987	2.9544
	STOI	0.6815	0.7458	0.8026	0.8518
ProC	PESQ	2.2884	2.6073	2.8644	3.1294
	STOI	0.7001	0.7668	0.8234	0.8805
ProN	PESQ	2.4114	2.8101	3.0902	3.3322
	STOI	0.7456	0.8075	0.8581	0.9039

method. However, compared with noisy speech, the Ref.1, Ref.2, Ref.4, and ProC methods do not perform well in terms of the STOI. A more possible reason is that the a priori information of speech and noise is not considered in the Ref.1 and Ref.4, and the higher accuracy of code-word selection is difficult for Ref.2 and ProC. Thus, more speech distortion existed in the enhanced speech processed by the Ref.1, Ref.2, and ProC methods, which may be detrimental to the improvement of speech intelligibility. For the Ref.3 and ProN methods using DNN model, two systems outperform the other three systems in terms of STOI and PESQ. Moreover, the STOI and PESQ improvements of two systems are also higher than the results of noisy speech in all input SNR conditions. In the ProN, the level modifications of speech and noise including the correlation between noise and speech are considered to generate a desired output for the DNN training. Compared with the Ref.3 method using the IRM, the ProN achieves a better performance in the STOI and PESQ.

4.2.6 Discussion

From the aforementioned experimental results, we can see that the first proposed technique achieves better performance of speech restoration than the pre-enhanced method and related codebook-based referenced method, and the second proposed technique performs better than all referenced methods. Herein, we discuss the advantages of the proposed methods against to the referenced methods.

As a classic unsupervised method, the MSA does not use a priori information about speech and noise. Generally, it requires noise estimation technique to estimate noise power spectrum [9, 53] from the noisy speech for achieving speech restoration. Here, the Ref.1 is obtained by combining the MSA and MCRA [53], and Ref.4 is obtained by combining the MSA and MS [9]. However, most of noise estimation techniques are hard to obtain noise power spectrum on a short-frame basis so that it is not helpful to track the rapid change of noise energy, which causes the performance of speech enhancement to be limited.

For the Ref.2 and ProC methods, the inter-channel cues are viewed as a priori information of speech and noise and are trained in the form of codebook. In these methods, since the idea of combining the down-mixed process and side information in BCC is used to achieve speech enhancement, the noise power spectrum is not needed in restoring target speech, which helps to reduce the problem of the methods depending on the noise power spectrum. In the Ref.2, multiple frequency bands share the same correlation between speech and noise so that the ability of reducing background noise is limited largely. Thus, to address the problem of the shared correlation from multiple frequency bands in the Ref.2, the ProC technique modifies the calculation of exact inter-channel cues. Moreover, in order to improve the accuracy of the code-vector selection, the ProC follows the pattern involved in [22] to calculate the pre-enhanced cues. However, in the Ref.2 and ProC methods, some appropriate code-vectors (termed as the intermediate parameters) need to be estimated in advance and a pre-enhanced module also need to be given, so the performance of these methods could be sensitive to the pre-enhanced method and the accuracy of intermediate parameters estimation.

Considering that the learning machine based on the DNN has a strong learning capacity in modeling the nonlinear interaction between training target and the acoustic features of noisy speech, the ProN technique uses the DNN model to directly learn the mapping relationship between the input features of noisy speech and the T-F mask based on exact inter-channel cues, namely, the pre-enhanced module and the intermediate link between noisy features and learned target are canceled. In the DNN-based Ref.3 method, IRM is used as the training target, which can help to improve the speech intelligibility and quality of target speech. However, it does not take into account the correlation between noise and speech. According to the studies in [44–48, 61], the correlation between speech and noise is helpful to improve speech quality. Thus, the paper attempts to use the inter-channel cues to incorporate the correlation between noise and speech into T-F mask for improving the quality of the enhanced speech.

5 Conclusions

In this paper, we present a single-channel speech enhancement system based on the inter-channel cues. In this system, the mechanism of processing left channel

and right channel signals from the BCC was exploited. In our work, the clean speech and noise signals are considered as the left channel and right channel signals of the BCC, respectively, and the noisy speech is considered as the down-mixed mono signal of the BCC. In order to achieve noise reduction based on the BCC, two techniques are proposed. In the codebook-based technique, when the clean signal and the corresponding noisy signal are given, the exact inter-channel cues and pre-enhanced cues can be extracted, respectively. This technique views the exact cues and pre-enhanced cues as the a priori information of speech and noise and trains them in the form of the codebook. With the weighted codebook mapping method in this technique, the exact cue parameters can be estimated to obtain a T-F mask for performing the single-channel speech enhancement. Considering that the errors from the intermediate link between noisy features and T-F mask may enlarge the inaccuracy of the output mask in the first technique, the second technique used a DNN model to directly learn the mapping relationship between noisy speech and the T-F mask based on exact inter-channel cues. Experiments showed that the proposed methods can achieve an effective improvement in speech quality and speech intelligibility.

Abbreviations

AMS: Amplitude modulation spectrogram; ARHMM: Auto-regressive hidden Markov model; BCC: Binueral cue coding; CASA: Computational auditory scene analysis; CM: Combined mask; DFT: Discrete Fourier transform; DNN: Deep neural network; GF: Gammatone filterbank power spectra; GMM: Gaussian mixture model; ICC: Inter-channel correlation; ICLD: Inter-channel level difference; ICTD: Inter-channel time difference; IRM: Ideal ratio mask; LDSN: Level difference of speech and noise; MAP: Maximum a posteriori; MFCC: Mel frequency cepstral coefficients; ML: Maximum likelihood; MMSE: Minimum mean square error; NCCC: Normalized cross-correlation coefficient; PESQ: Perceptual evaluation of speech quality; RASTA-PLP: Relative spectral transform and perceptual linear prediction; SNC: Speech and noise correlation; SSNR: Segmental signal-to-noise ratio; STOI: Short-time objective intelligibility; T-F: Time-frequency; WCBM: Weighted codebook mapping

Acknowledgements

This work was supported by the National Natural Science Foundation of China (i.e., grant no. 61831019 and grant no.61471014).

Authors' contributions

All authors read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (i.e., grant no. 61831019 and grant no.61471014).

Availability of data and materials

The utterances used in this paper are from the TIMIT database [55]. Noise signals are chosen from NOISEX-92 databases [56] in the experiment.

Competing interests

The authors declare that they have no competing interests.

Received: 6 January 2019 Accepted: 26 September 2019
Published online: 11 December 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). https://doi.org/10.1109/TASSP.1979.1163209
2. H. M. Goodarzi, S. Seyedtabaii, "Speech enhancement using spectral subtraction based on a modified noise minimum statistics estimation," International Joint Conference on INC, IMS and IDC, 2009, Seoul, South Korea. DOI: https://doi.org/10.1109/NCM.2009.272
3. P. C. Loizou, Speech Enhancement: Theory and Practice. Boca Raton, 670 FL, USA: CRC Press, 2007. ISBN: 9781420015836
4. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process **33**, 443–445 (1985). https://doi.org/10.1109/TASSP.1985.1164550
5. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean square error short-time spectral amplitude estimator. IEEE Trans Acoust Speech Signal Process **ASSP-32**(6), 1109–1121 (1984). https://doi.org/10.1109/TASSP.1984.1164453
6. S. Srinivasan, J. Samuelsson, W.B. Kleijn, Codebook-based Bayesian speech enhancement for nonstationary environments. IEEE Trans Audio Speech Lang Process **15**(2), 441–452 (2007). https://doi.org/10.1109/tasl.2006.881696
7. J.S. Erkelens, R. Heusdens, Tracking of nonstationary noise based on data-driven recursive noise power estimation. IEEE Trans Audio Speech Lang Process **16**(6), 1112–1123 (2008). https://doi.org/10.1109/tasl.2008.2001108
8. Q. He, C.C. Bao, F. Bao, Multiplicative update of auto-regressive gains for codebook-based speech enhancement. IEEE Trans. Audio Speech Lang Process **25**(3), 457–468 (2017). https://doi.org/10.1109/TASLP.2016.2636445
9. R. Martin, Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Process **9**(5), 504–512 (2001). https://doi.org/10.1109/89.928915
10. D.Y. Zhao, W.B. Kleijn, HMM-based gain modeling for enhancement of speech in noise. IEEE Trans Audio Speech Lang Process **15**(3), 882–892 (2007). https://doi.org/10.1109/TASL.2006.885256
11. S. Srinivasan, J. Samuelsson, W.B. Kleijn, Codebook driven short term predictor parameter estimation for speech enhancement. IEEE Trans Audio Speech Lang Process **14**(1), 163–176 (2006). https://doi.org/10.1109/TSA.2005.854113
12. X. Y. Wang and C. C. Bao, "Speech enhancement using a joint MAP estimation of LP parameters." Int. Conf. signal process., comm., comput., 2015. DOI: https://doi.org/10.1109/ICSPCC.2015.7338863
13. Y. Linde, A. Buzo, R.M. Gray, An algorithm for vector quantization design. IEEE Trans Commun ***C-28***(1), 84–95 (1980). https://doi.org/10.1109/tcom.1980.1094577
14. A. Reddy, B. Raj, Soft mask methods for single-channel speaker separation. IEEE Trans Audio Speech Lang Process **15**(6), 1766–1776 (2007). https://doi.org/10.1109/TASL.2007.901310
15. MH Radfar and RM Dansereau, "Single-channel speech separation using soft masking filtering." IEEE Trans. Audio, Speech, Lang. Process. vol. 15, no. 8, pp. 2299–2310, 2007. DOI: https://doi.org/10.1109/tasl.2007.904233
16. K. Hu, D.L. Wang, An iterative model-based approach to cochannel speech separation. EURASIP J Audio Speech Music Process **14**, 1–11 (2013). https://doi.org/10.1186/1687-4722-2013-14
17. Z. Wang, X. Wang, X. Li, Q. Fu, and Y. Yan, "Oracle performance investigation of the ideal masks," in IWAENC, pp. 1-5, 2016. DOI: https://doi.org/10.1109/IWAENC.2016.7602888
18. B. Yan, C. Bao, Z. Bai, "DNN-based speech enhancement via integrating NMF and CASA," International Conference on Audio, Language and Image Processing (ICALIP), 2018. DOI: https://doi.org/10.1109/ICALIP.2018.8455780
19. Y. Xu, J. Du, L. Dai, C. Lee, A regression approach to speech enhancement based on deep neural networks. IEEE Trans Audio Speech Lang Process **23**(1), 7–19 (2015). https://doi.org/10.1109/TASLP.2014.2364452
20. D. S. Williamson, Y. X. Wang, and D. L. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in Proc. ICASSP, pp. 5220-5224, 2016. DOI: https://doi.org/10.1109/ICASSP.2016.7472673
21. D. S. Williamson, Y. Wang and D. L. Wang, "Complex ratio masking for monaural speech separation," IEEE/ACM Trans. on Audio, Speech, and Lang. Process., vol. 24, pp. 483-492, 2016. DOI: https://doi.org/10.1109/TASLP.2015.2512042
22. M. Geravanchizadeh and R. Ahmadnia. "Monaural speech enhancement based on multi-threshold masking," In blind source separation, G.R. Naik, W.

Wang, Springer Berlin Heidelberg, pp.369-393, 2014. DOI: https://doi.org/10.1007/978-3-642-55016-4_13
23. Y.X. Wang, A. Narayanan, D.L. Wang, On training targets for supervised speech separation. IEEE Trans Audio Speech Lang Process **22**(12), 1849–1858 (2014). https://doi.org/10.1109/taslp.2014.2352935
24. J. Chen, Y. Wang, S.E. Yoho, D.L. Wang, E.W. Healy, Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises. J Acoustl Soc Am **139**(5), 2604–2612 (2016). https://doi.org/10.1121/1.4948445
25. D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: an overview." arXiv preprint arXiv:1708.07524, 2017. DOI: https://doi.org/10.1109/TASLP.2018.2842159
26. N. Chen, C. C. Bao and F. Deng, "Speech enhancement with binaural cues derived from a priori codebook." *in Proc.ISCSLP*, 2016. DOI: https://doi.org/10.1109/ISCSLP.2016.7918377
27. N. Chen, C. C. Bao and X. Y. Wang, "Speech enhancement based on binaural cues." in Proc.APSIPA, 2017. DOI: https://doi.org/10.1109/APSIPA.2017.8282017
28. T. May, S. van de Par, A. Kohlrausch, A binaural scene analyzer for joint localization and recognition of speakers in the presence of interfering noise sources and reverberation. IEEE Trans Audio Speech Lang Process **20**, 2016–2030 (2012). https://doi.org/10.1109/tasl.2012.2193391
29. Y. Jiang and R. S. Liu. "Binaural deep neural network for robust speech enhancement." in Proc. IEEE Int. Conf. Signal, Process., Communications, Computing, pp:692-695, 2014. DOI: https://doi.org/10.1109/ICSPCC.2014.6986284
30. Y. Jiang, D.L. Wang, R.S. Liu, Z.M. Feng, Binaural classification for reverberant speech segregation using deep neural networks. IEEE Trans Audio Speech Lang Process **22**(12), 2112–2121 (2014). https://doi.org/10.1109/TASLP.2014.2361023
31. S. Chandna, W. Wang, Bootstrap averaging for model-based source separation in reverberant conditions. IEEE/ACM Trans Audio Speech Lang Process **26**(4), 806–819 (2018). https://doi.org/10.1109/TASLP.2018.2797425
32. A. Zermini, Q. Liu, Y. Xu, M. D. Plumbley, D. Betts, and W. Wang, "Binaural and log-power spectra features with deep neural networks for speech-noise separation", in Proc. IEEE 19th International Workshop on Multimedia Signal Processing (MMSP 2017), Luton, UK, October 16-18, 2017. DOI: https://doi.org/10.1109/MMSP.2017.8122280
33. Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks", EURASIP Journal on Audio Speech and Music Processing, 2016:7, 18 pages, DOI https://doi.org/10.1186/s13636-016-0085-x, 2016.
34. A. Alinaghi, P. Jackson, Q. Liu, W. Wang, Joint mixing vector and binaural model based stereo source separation. IEEE/ACM Trans Audio Speech Lang Process **22**(9), 1434–1448 (2014). https://doi.org/10.1109/TASLP.2014.2320637
35. A. Alinaghi, W. Wang, and P. Jackson, "Integrating binaural cues and blind source separation method for separating reverberant speech mixtures," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), pp. 209-212, Prague, Czech Republic, May 22-27, 2011. DOI: https://doi.org/10.1109/ICASSP.2011.5946377
36. A. Alinaghi, W. Wang, and P.J.B. Jackson, "Spatial and coherence cues based time-frequency masking for binaural reverberant speech separation", in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013), pp. 684-688, Vancouver, Canada, May 26-31, 2013. DOI: https://doi.org/10.1109/ICASSP.2013.6637735
37. A. Alinaghi, P Jackson, and W. Wang, "Comparison between the statistical cues in BSS techniques and binaural cues in CASA approaches for reverberant speech separation", in Proc. IET International Conference on Intelligent Signal Processing (ISP 2013), London, UK, December 3-4, 2013. DOI: https://doi.org/10.1049/cp.2013.2076
38. Q. Liu, W. Wang, P. Jackson, and Y. Tang, "A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions", in Proc. European Signal Processing Conference (EUSIPCO 2017), Kos Island, Greece, August 28- September 2, 2017. DOI: https://doi.org/10.23919/EUSIPCO.2017.8081412
39. Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve permutation problem in convolutive BSS," Signal Processing, Special Issue on Latent Variable Analysis and Signal Separation, vol. 92, vol. 8, pp. 1916-1927, 2012. DOI: https://doi.org/10.1016/j.sigpro.2011.11.007
40. C. Faller, F. Baumgarte, "Binaural cue coding: a novel and efficient representation of spectral audio." IEEE ICASSP, Orlando, Florida, USA, pp. 1841-1844, 2002. DOI: https://doi.org/10.1109/ICASSP.2002.5744983
41. C. Faller, F. Baumgarte, "Binaural cue coding, part I: psychoacoustic fundamentals and design principles." IEEE Trans. Speech and Audio, Process., vol. 11, no. 6, pp. 509-519, 2003. DOI: https://doi.org/10.1109/TSA.2003.818109
42. C. Faller, F. Baumgarte, "Binaural cue coding, part II: schemes and applications." IEEE Trans. Speech and Audio, Process., vol. 11, no. 6, pp. 520-531, 2003. DOI: https://doi.org/10.1109/TSA.2003.818108
43. Y. Zhang, R. Hu. "Speech wideband extension based on Gaussian mixture model." Acta Acustica, vol. 34, no. 5, pp. 471-480, 2009. ISSN: 03710025
44. S. Liang, W. J. Liu, W. Jiang, and W. Xue. "The optimal ratio time-frequency mask for speech separation in terms of the signal-to-noise ratio." The Journal of the Acoustical Society of America, vol. 134, no. 5, 2013, pp. EL452-EL458, 2013. DOI: https://doi.org/10.1121/1.4824632
45. S. Liang, W.J. Liu, W. Jiang, W. Xue, The analysis of the simplification from the ideal ratio to binary mask in signal-to-noise ratio sense. Speech Comm **59**, 22–30 (2014). https://doi.org/10.1016/j.specom.2013.12.002
46. Y. Lu, P. Loizou, A geometric approach to spectral subtraction. Speech Comm **55**, 453–466 (2008). https://doi.org/10.1016/j.specom.2008.01.003
47. F. Bao, W.H. Abdulla, A new ratio mask representation for CASA-based speech enhancement. IEEE Trans Audio Speech Lang Process **27**(1), 7–19 (2019). https://doi.org/10.1109/TASLP.2018.2868407
48. F. Bao, W.H. Abdulla, A new IBM estimation method based on convex optimization for CASA. Speech Comm **97**, 51–65 (2018). https://doi.org/10.1016/j.specom.2018.01.002
49. B. Gao, W.L. Woo, S.S. Dlay, Unsupervised single-channel separation of nonstationary signals using gammatone filter-bank and itakura-saito nonnegative matrix two-dimensional factorizations. IEEE Trans Circuits Syst I **60**(3), 662–675 (2013). https://doi.org/10.1109/tcsi.2012.2215735
50. A. Narayanan, D.L. Wang, A CASA-based system for long-term SNR estimation. IEEE Trans Audio Speech Lang Process **20**(9), 2518–2527 (2012). https://doi.org/10.1109/TASL.2012.2205242
51. J. Chen, Y. Wang, D.L. Wang, A feature study for classification-based speech separation at low signal-to-noise ratios. IEEE/ACM Trans Audio Speech Lang Process **22**(12), 1993–2002 (2014). https://doi.org/10.1109/TASLP.2014.2359159
52. F. Deng, F. Bao, C.C. Bao, Speech enhancement using generalized weighted β-order spectral amplitude estimator. Speech Commun **59**, 55–68 (2014). https://doi.org/10.1016/j.specom.2014.01.002
53. I. Cohen, B. Berdugo, Noise estimation by minima controlled recursive averaging for robust speech enhancement. IEEE Signal Process Lett **9**(1), 12–15 (2002). https://doi.org/10.1109/97.988717
54. J. Taghia, N. Mohammadiha, J. Sang, et al. "An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments." 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011. DOI: https://doi.org/10.1109/ICASSP.2011.5947389
55. V. Zue, S. Seneff, J. Glass. (1990). "Speech database development at MIT: TIMIT and beyond," Speech Commun, vol. 9, no. 4, pp. 351–356, 1990. DOI: https://doi.org/10.1016/0167-6393(90)90010-7
56. A. Varga, H.J. Steeneken, 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 **12**(3), 247–251 (1993). https://doi.org/10.1016/0167-6393(93)90095-3
57. N. Fan, J. Rosca, R. Balan. "Speech noise estimation using enhanced minima controlled recursive averaging," 2007 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2007. DOI: https://doi.org/10.1109/ICASSP.2007.366979
58. Antony WR, John GB, Michael PH, Andries PH. "Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," IEEE International Conference on Acoustics,Speech and Signal Processing (ICASSP). 2001. https://doi.org/10.1109/ICASSP.2001.941023.
59. S. Rangachari, P. C. Loizou, Y. Hu. "A noise estimation algorithm with rapid adaptation for highly nonstationary environments," 2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2004. DOI: https://doi.org/10.1109/ICASSP.2004.1325983
60. C.H. Taal, R.C. Hendriks, R. Heusdens, et al., An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans Audio Speech Lang Process **19**(7), 2125–2136 (2011). https://doi.org/10.1109/tasl.2011.2114881
61. F. Bao and W. H. Abdulla. "Noise masking method based on an effective ratio mask estimation in Gammatone channels." APSIPA Trans. Signal, Information Process., vol. 7, e5, pp.1-12, 2018. DOI: https://doi.org/10.1016/j.specom.2018.01.002

Publisher's Note