Yu *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2016) 2016:7
DOI 10.1186/s13636-016-0085-x

EURASIP Journal on Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks

Yang Yu[1*†], Wenwu Wang[2†] and Peng Han[1]

Abstract

Time-frequency (T-F) masking is an effective method for stereo speech source separation. However, reliable estimation of the T-F mask from sound mixtures is a challenging task, especially when room reverberations are present in the mixtures. In this paper, we propose a new stereo speech separation system where deep neural networks are used to generate soft T-F mask for separation. More specifically, the deep neural network, which is composed of two sparse autoencoders and a softmax regression, is used to estimate the orientations of the dominant source at each T-F unit, based on low-level features, such as mixing vector (MV), interaural level, and phase difference (IPD/ILD). The dataset for training the networks was generated by the convolution of binaural room impulse responses (RIRs) and clean speech signals positioned in different angles with respect to the sensors. With the training dataset, we use unsupervised learning to extract high-level features from low-level features and use supervised learning to find the nonlinear functions between high-level features and the orientations of dominant source. By using the trained networks, the probability that each T-F unit belongs to different sources (target and interferers) can be estimated based on the localization cues which is further used to generate the soft mask for source separation. Experiments based on real binaural RIRs and TIMIT dataset are provided to show the performance of the proposed system for reverberant speech mixtures, as compared with a model-based T-F masking technique proposed recently.

Keywords: Deep learning, Deep neural networks, Source separation, Soft mask

1 Introduction

Robust speech separation is an attractive research field and provides a useful front-end for many applications, e.g., hearing aids, mobile communication device, and automatic speech recognition system. Many methods have been applied to this problem, such as independent component analysis (ICA) [1–3], beamforming [4], and computation auditory scene analysis (CASA) [5, 6]. The performance of these algorithms, however, is still limited in complex acoustic environment, especially when room reverberation is present in the mixtures. This is in contrast to human auditory system which is skillful in listenting into a particular conversation in a cocktail party environment with the presence of background noise and interfering sound. There is a big performance gap between the human auditory system and machine-based listening system. An influential view in auditory scene analysis is that the human auditory system splits the sound mixtures to fragments (e.g., regions in the time-frequency plane), and the fragments which belong to the same acoustic source will be assigned to a same cluster. Based on this idea, time-frequency (T-F) masking technique has been proposed for speech source separation where the mask can be derived from various cues based on the analysis of temporal, spectral, or spatial features of the sources. Recently, a time-frequency masking technique has been proposed in [7] where the mixing vector (MV) [8] and interaural phase and level difference (IPD/ILD) [9] have been integrated using a Gaussian mixture model (GMM) whose

*Correspondence: nwpuyuy@nwpu.edu.cn
†Equal contributors.
[1] School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China
Full list of author information is available at the end of the article

parameters are estimated iteratively using an expectation maximization (EM) algorithm. These methods provide a nice probabilistic framework for incorporating complementary information to deal with the uncertainties in T-F assignment. However, the performance of these algorithms is also limited by the accuracy of model fitting especially when room reverberation is present.

The GMM is essentially a shallow architecture of neural network which contains at most one layer of nonlinear feature transformation and is shown to offer good performance in source separation for anechoic mixtures [10] or mixtures with a relatively low level of reverberation. The shallow architecture, however, has a limited representation ability, which may cause performance degradation when applied in the complex real-world problems, such as speech separation in highly reverberant environments. Recent studies in speech recognition have shown that a deep architecture with more hidden layers can increase the representation abilities of a neural network, and it can be used to build internal representation for rich sensory data [11–14]. The deep architecture is regarded as being similar to the hierarchical structures within human visual and auditory systems, where the raw image or speech waveforms are transformed to a high-order linguistic level by these hierarchical structures [15–18]. The deep structure has the potential to reduce the performance gap between the human auditory system and machine listening system, as shown in recent works in the area of natural language processing and speech recognition systems [19–22]. The success of deep neural networks (DNNs) in these applications inspires us to investigate its potential for improving the performance of stereo speech source separation algorithms.

In this paper, we focus on the multiuser stereo speech source separation in reverberation environments and present a new approach for T-F assignment and mask estimation based on DNNs [23]. The network is trained with the low-level of features (i.e., MV and ILD/IPD) extracted from a training dataset of observed speech signals. In the separation stage, the trained network is used to estimate the orientations (i.e., directions of arrivals) of the target and interferers which is further exploited to derive the source occupation probability (and thereby the mask) at each T-F unit of the mixture. Our experimental results show that the proposed method performs significantly better than the GMM/EM-based baseline method [7] in terms of both signal to distortion ratio (SDR) and perceptual evaluation of speech quality (PESQ).

The remainder of the paper is organized as follows. Section 2 briefly discusses the related works. Section 3 outlines the proposed system. Section 4 discusses the low-level features to be used as inputs to the network. Section 5 presents the details about the deep network, including its structure, the training method, and how it is used for separation. Section 6 shows the experiments using real RIRs and TIMIT data before the conclusion is drawn in Section 7.

2 Relation to prior work

Several recent works have explored the potential of using DNNs for monaural/stereo speech separation. In [24], Wang et al. explored the use of monaural features for classification-based speech segregation. To deal with noise in the mixtures, a group Lasso approach and SVM classifier have been applied for generating the ideal binary mask for noise cancellation by combining different features. The experimental results show that (1) the complementary feature set is shown to give stable performance in experiments and outperforms each of its components significantly and (2) the unit-level features give better performance than frame-level features in unmatched test condition. In [25], Xu et al. presented a regression-based speech enhancement framework using DNNs, and the restricted Boltzmann machines (RBMs) have been used to learn a deep generative model for pre-training. They found that (1) using the large training dataset could result in a good generalization capability in mismatched testing conditions and (2) the two and three hidden layer DNNs have the similar performance. In [26], Narayanan and Wang proposed a feature enhancement algorithm for improving noise robustness of automatic speech recognition systems. The algorithm estimates a smoothed ideal ratio mask in the Mel spectrogram domain using DNNs, which is then used to filter out noise before cepstral transformation. In [27, 28], Huang et al. proposed to jointly optimize the deep learning models (deep neural networks and recurrent neural networks) with an extra masking layer to enforce a reconstruction constraint. They used a discriminative training criterion for the neural networks to further enhance the monaural source separation performance. In [29], Jin and Wang proposed a supervised learning approach to monaural segregation of reverberant speech using the multiresolution cochleagram (MRCG) features [30].

In [31], Jiang et al. first introduced DNNs to stereo speech separation. Similar to the work in [25] and [26], the RBMs were used to get the initial parameters of the DNNs and the output of the DNNs are the estimated ideal binary mask (IBM). They found that the DNN-based algorithm with joint binaural and monaural features could achieve better results than the representative binaural separation algorithms, especially when reverberation is present in the environment, and the target and interfering sources are either collocated or close to each other.

Compared with the monaural segregation of reverberant speech in [29], the stereo speech separation in [31] tends to be more robust due to the use of spatial information. In [7], GMM is used to model the MV and IPD/ILD cues that contain spatial information and the EM algorithm is used to estimate the model parameters and to derive the T-F mask. The combination of IPD/ILD and MV improves the separation quality as compared with the use of either only IPD/ILD or only MV and is achieved by using a coarse search to find the optimum set of weighting parameters which adjust the contribution of these cues. However, the optimum set of weighting parameters varies with different acoustic environment, i.e., the level of reverberation. In addition, the GMM used in [7] is a classical shallow architecture, and its representation ability is limited and can cause the performance degradation when the reverberation is present in the mixtures.

In this paper, similar to [7] and [31], we also consider multiuser stereo source separation problem. Instead of using GMM and EM or the RBMs, however, we use DNNs (composed of sparse autoencoders and softmax classifier) to estimate the source occupation likelihood at each T-F point. More specifically, the sparse autoencoders are used to learn the general model and the combined features—IPD/ILD and MV were used as the input of the DNNs. In other words, the low level features, i.e., IPD/ILD and MV, are now modelled with DNNs composed of sparse autoencoders and softmax classifier, and the output of the DNNs is an estimated soft mask (ratio mask). The network parameters are obtained through training by a greedy layer-wise training method [32] based on a training dataset containing observed speech signals (with one source speech signal placed at a different direction with respect to the microphones). With the trained sparse autoencoder and softmax classifer, we extract high-level features (i.e., spatial information of the sources) from the low-level features of the mixtures and generate the soft mask based on the softmax regression. The weighting parameters which are used to adjust the contribution of different cues will be learned automatically by the deep neural networks. Hence, different from [7, 31], we improve the separation quality by using the deep neural networks to find the optimum set of parameters and weighting the contributions of the cues (IPD/ILD, MV) automatically.

3 System overview

Our proposed system consists of the following four stages: (1) extraction of the low-level features (i.e. MV and ILD/IPD) (details in Section 4), (2) training of the deep networks (details in Section 5.1), (3) estimation of the probabilities that each T-F unit of the mixtures belongs to different sources and generation of the soft mask (details in Section 5.2), and (4) reconstruction of the target signal from the soft mask and the mixture signal. The system architecture is shown in Fig. 1. It should be noted that the neural nets are trained using isolated utterances (utterances originating from a single direction, i.e., clean speech utterances convolved with the binaural room impulse responses (BRIRs) corresponding to that direction) rather than the mixtures in stage (2).

The inputs to the system are the stereo (left and right) channel mixtures. We perform short-time Fourier transform (STFT) to both channels and obtain the T-F representation of the input signals, $X_L(m,f)$ and $X_R(m,f)$ where $m = 1,\cdots,M$ and $f = 1,\cdots,F$ are the time frame and frequency bin indices, respectively. The low-level features, i.e., MV and IPD/ILD, are then estimated at each T-F unit (details in Section 4). Next, we group the low-level features into N blocks (only along the frequency bins f). Each block includes K frequency bins, for example, the n-th block contains the bins $((n-1)K+1,\cdots,nK)$, where $K = \frac{F}{N}$. We build N deep networks with each corresponding to one block and use them to estimate the direction of arrivals (DOAs) of the sources. Through unsupervised learning and the sparse autoencoder [11] in deep networks, high-level features (coded positional information of the sources) are extracted and used as inputs for the output layer (i.e., the softmax regression) of the networks. The output of softmax regression is a source occupation probability (i.e., the soft mask) of each block (through the ungroup operation, T-F units in the same block are assigned with the same source occupation probability) of the mixtures. Then, the sources can be recovered by the inverse STFT (ISTFT).

The key point of our proposed system is the training of deep networks and the generation of soft mask. From the view point of practical applications, we create a dataset of sensor signals with each containing only a single source (i.e., source speech convolved with RIRs) from different directions with respect to the sensors for training and use the orientations of single source as the ground truth (described in Section 5.2). With the training dataset, the deep networks are trained by using a greedy layer-wise training method [32]. With the trained deep networks, we can get the probability of each T-F block of the input mixtures associated with different DOAs. Using a predefined threshold, we can estimate the number of sources, the DOAs of the sources, and a matrix of probabilities which we call "Probability Mask" in Fig. 1. Through the ungroup operation, we assign the same probability to T-F units that are belonging to the same block. Then, we obtain the soft mask for speech separation from the Probability Mask.

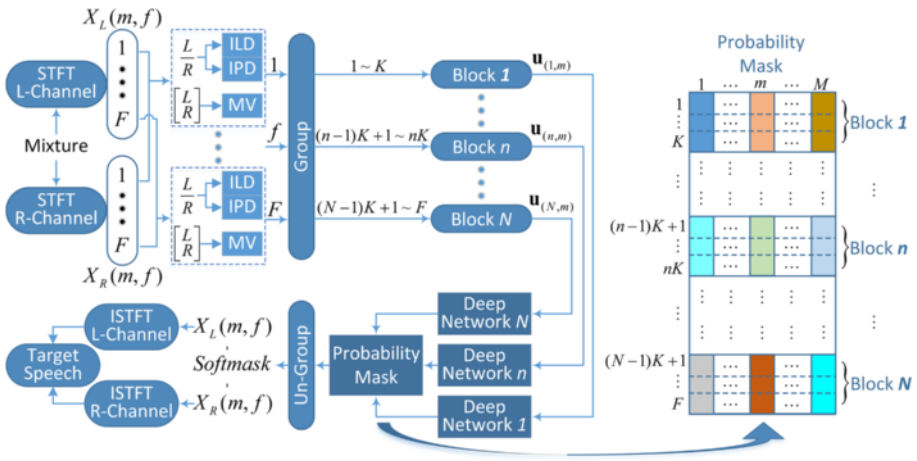

Fig. 1 System architecture for deep neural network-based time-frequency masking for stereo source separation

The N deep networks in our proposed system have the same architecture, and the details about the architecture and training method can be found in Section 5. Next, we discuss the low-level features used in our proposed system.

4 The low-level features for localization based separation

Many features can be used for stereo speech separation, such as IPD or ITD [33], ILD or interaural intensity differences (IID) [33], and the MV cue. It is widely acknowledged that ITD or IPD tends to be more robust in the low frequency range, whereas ILD or IID is more robust in the high-frequency range [34]. In [7], Alinaghi et al. found that the MV cues are more distinct compared to binaural cues (IPD/ILD) for the sources placed close to each other, whereas binaural cues IPD/ILD offer better separation results when the sources are distant from each other. These observations motivated Alinaghi et al. to combine these cues, introducing a new robust algorithm to improve the speech separation quality. We follow the work in [7] and use IPD/ILD and MV as the low-level features and the inputs to the neural networks. The nonlinear relationship between the source occupation probabilities and the input low-level features can be found by the deep networks and thus has the potential to further improve the speech separation quality. These low-level features are used to derive high-level features to be classified by sparse autoencoders.

The MV and the IPD/ILD cues can be calculated from the mixtures.

The MV [8] can be derived as

$$\mathbf{z}(m,f) = \frac{\mathbf{W}(f)\widetilde{\mathbf{x}}(m,f)}{\left\|\mathbf{W}(f)\widetilde{\mathbf{x}}(m,f)\right\|} \tag{1}$$

with $\widetilde{\mathbf{x}}(m,f) = \frac{[X_L(m,f),X_R(m,f)]^T}{\left\|[X_L(m,f),X_R(m,f)]^T\right\|}$, where $\mathbf{W}(f)$ is a whitening matrix, with each row being one eigenvector of $E(\widetilde{\mathbf{x}}(m,f)\widetilde{\mathbf{x}}^H(m,f))$, the superscript H is Hermitian transpose, and $\|\bullet\|$ is Frobenius norm.

ILD and IPD are the phase and amplitude difference between the left and right channel and calculated as follows [9]:

$$\alpha(m,f) = 20\log_{10}\left(\left|\frac{X_L(m,f)}{X_R(m,f)}\right|\right) \tag{2}$$

$$\phi(m,f) = \angle\left(\frac{X_L(m,f)}{X_R(m,f)}\right) \tag{3}$$

where $|\bullet|$ takes the absolute value of its argument, and $\angle(\bullet)$ finds the phase angle.

Concatenating the MV and ILD/IPD features, a feature vector can be obtained at each T-F unit, which is $\widetilde{\mathbf{u}}(m,f) = \left[\widehat{\mathbf{z}}^T(m,f), \alpha(m,f), \phi(m,f)\right]^T \in \mathbb{R}^6$. Since the inputs to the DNNs are real numbers, we use the real part and imaginary part of $\mathbf{z}$ as the features, i.e., $\widehat{\mathbf{z}}(m,f) = \left[\mathrm{Re}(z_L(m,f)), \mathrm{Im}(z_L(m,f)), \mathrm{Re}(z_R(m,f)), \mathrm{Im}(z_R(m,f))\right]$. Then, we group all the feature vectors $\widetilde{\mathbf{u}}(m,f)$ into N blocks (only along the frequency bins). For each block, we get a $6K$-dimensional feature vector $\mathbf{u}_{(n,m)} = \left[\widetilde{\mathbf{u}}^T(m,(n-1)K+1), \cdots, \widetilde{\mathbf{u}}^T(m,nK)\right]^T \in \mathbb{R}^{6K}$, where K is the number of the frequency bins in each block, as the input to the deep networks.

5 The deep networks

As described in Section 3, we group the low-level features into N blocks and build N individual deep networks which have the same architecture to classify the DOAs of the current input mixture in each block. The architecture of the deep network is shown in Fig. 2 and composed of deep autoencoder [11] and softmax classifier. More specifically,

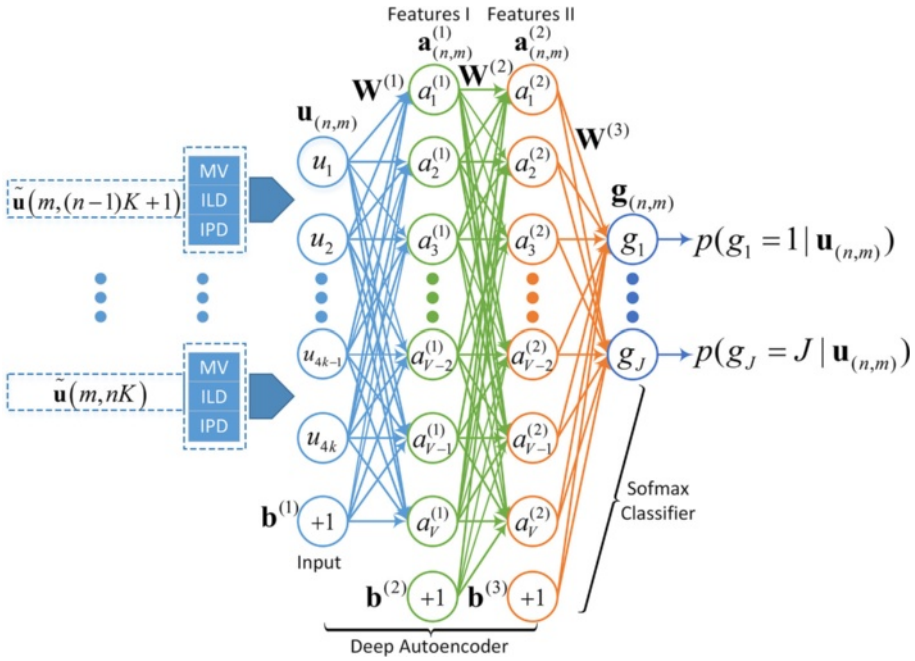

Fig. 2 The architecture for the deep neural networks in our proposed system. The deep neural networks are composed by the deep autoencoder for high-level feature extraction and the softmax classifier for soft mask generation and was trained by the greedy layer-wise training method

for stereo speech separation task, the target location is a natural choice of the output of the network. As shown in Fig. 3, we split the whole space to J ranges with respect to the sensors and separate the target and interferers based on different orientation ranges (DOAs with respect to the listener) where they are located. We apply the softmax classifier (to be discussed in Section 5.2) to perform the classification task and the inputs to the classifier, i.e., the high-level features: $\mathbf{a}^{(2)}$ which are extracted from the low-level features (ILD/IPD and MVs), are produced by the deep autoencoder. Assuming that the position of the target in the current input sample remains unchanged, the deep network estimates the probability $p\left(g_j = j|\mathbf{u}_{(n,m)}\right)$ of the orientation of the current input sample belonging to the orientation index j, where g_j is the j-th output unit of the network and the $\mathbf{u}_{(n,m)}$ is the m-th input sample of the n-th block (group). With the estimated orientation (obtained by selecting the maximum probability index) of each input sample, we cluster the samples which have the same orientation index to get the probability mask and obtain the soft mask from the probability mask through the ungroup operation. Note that each T-F unit in the same block is assigned the same probability. The number of sources can also be estimated from the probability mask by using a predefined probability threshold, typically chosen as 0.1 in our experiments (we only considered two or three sources and found empirically this value to be suitable).

5.1 Deep autoencoder

An autoencoder (shown in Fig. 4) is an unsupervised learning algorithm based on backpropagation. It aims to learn an approximation $\widehat{\mathbf{u}}$ of the input $\mathbf{u}$. It appears to be learning a trivial identity function; but by using some constraints on the learning process, such as limiting the number of neurons activated (sparsity constraint), it discloses some interesting structures about the data [11, 35, 36]. As shown in Fig. 4, the output of the autoencoders can be defined as $\widehat{\mathbf{u}} = \text{sigm}\left(\widehat{\mathbf{W}}^{(2)}\mathbf{a}^{(1)} + \widehat{\mathbf{b}}^{(2)}\right)$ with $\mathbf{a}^{(1)} = \text{sigm}\left(\mathbf{W}^{(1)}\mathbf{u} + \mathbf{b}^{(1)}\right)$, where the function $\text{sigm}(\mathbf{u}) = \frac{1}{1+\exp(-\mathbf{u})}$ is the logistic function, $\mathbf{W}^{(1)} \in \mathbb{R}^{V\times Y}$, $\mathbf{b}^{(1)} \in \mathbb{R}^{V}$, $\widehat{\mathbf{W}}^{(2)} \in \mathbb{R}^{Y\times V}$, and $\widehat{\mathbf{b}}^{(2)} \in \mathbb{R}^{Y}$, V is the number of hidden layer neurons, and Y is the number of input layer neurons, which is the same as that of the output layer neurons [37]. In our proposed system, we set $Y = 6K$, where K is the number of frequency bin in each block (group).

With the sparsity constraint, most of the neurons in the autoencoder are assumed to be inactive. More specifically, $\widehat{\rho}_v = \frac{1}{M}\sum_{m=1}^{M} a_v\left(\mathbf{u}_{(m)}\right)$ is the average activation of unit v with the m-th input sample $\mathbf{u}_{(m)}$, where $a_v = \text{sigm}\left(\mathbf{w}_v^{(1)}\mathbf{u} + b_v^{(1)}\right)$ denotes the activation value of the hidden layer unit v in the autoencoder and M is the number of training samples. With the sparsity constraint $\widehat{\rho}_v =$

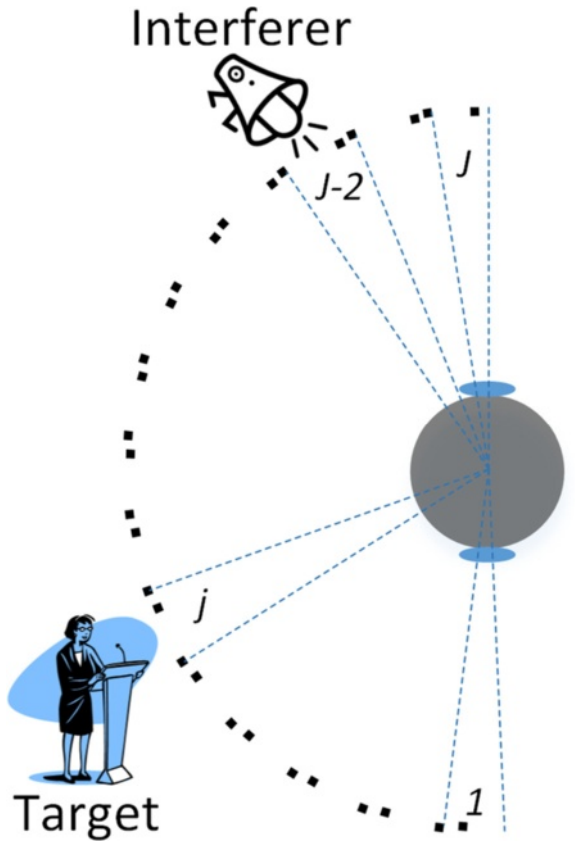

Fig. 3 Split the space to *J* ranges. We split the whole space to *J* ranges with respect to the sensors, and separate the target and interferers based on different orientation ranges (DOAs with respect to the listener) where they are located

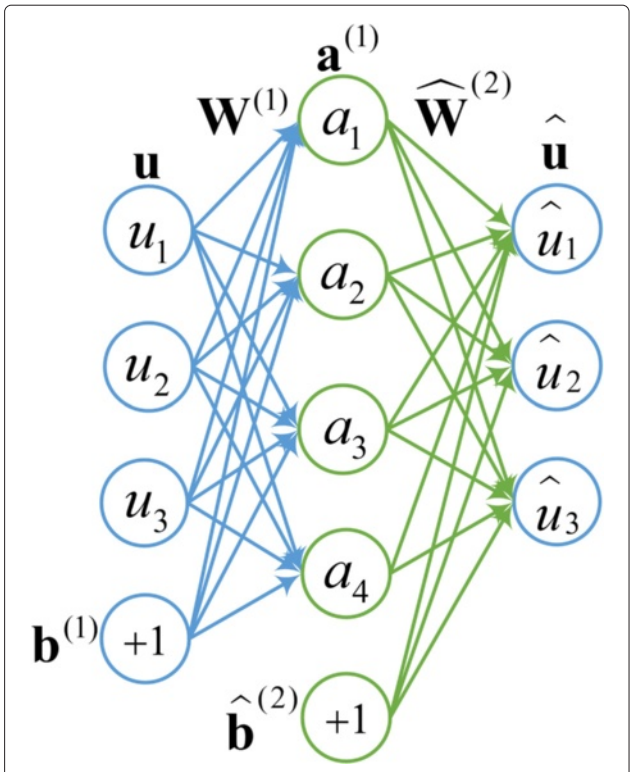

Fig. 4 A classical architecture of single layer autoencoder

ρ, the cost function $J_{\text{sparse}}(\mathbf{W},\mathbf{b})$ of the sparse autoencoder can be written as follows

$$J_{\text{sparse}}(\mathbf{W},\mathbf{b}) = \frac{1}{2}\|\mathbf{u}-\widehat{\mathbf{u}}\|^2 + \beta\sum_{\nu=1}^{V} KL(\rho \parallel \widehat{\rho}_\nu)$$
$$\sum_{\nu=1}^{V} KL(\rho \parallel \widehat{\rho}_\nu) = \sum_{\nu=1}^{V} \rho\log\frac{\rho}{\widehat{\rho}_\nu} + (1-\rho)\log\frac{1-\rho}{1-\widehat{\rho}_\nu} \tag{4}$$

where β controls the weight of the penalty term and ρ is a parameter preset before training, typically very small [37]. More details about the sparse autoencoder can be found in online lecture notes [38, 39].

In our proposed system, the cost function $J_{\text{sparse}}(\mathbf{W},\mathbf{b})$ is minimized using the limited memory BFGS (L-BFGS) optimization algorithm [40, 41] and the single-layer sparse autoencoder is trained by using the backpropagation algorithm.

After the finishing of the training of single-layer sparse autoencoder, we discard the output layer neurons, the relative weights $\widehat{\mathbf{W}}^{(2)}$, bias $\widehat{\mathbf{b}}^{(2)}$, and only save the input layer neurons $\mathbf{W}^{(1)}$ and $\mathbf{b}^{(1)}$. The output of the hidden layer—$\mathbf{a}^{(1)}$ are used as the input samples of the next single-layer sparse autoencoder. We can build a deep autoencoder by repeating these steps and stacking two or more layers of independently trained sparse autoencoders. The stacking procedure is shown on the right side of Fig. 5. The features II shown on figure are the high-level features and can be used as the training dataset for the softmax regression discussed next.

Many studies on deep autoencoders have shown that with the deep architecture (more than one hidden layer), more complex representation can be obtained from the simple low-level features. As a result, the underlying regularities of the data can be captured, leading to better performance, e.g., in recognition [23]. This motivates us to use deep autoencoder (two hidden layers) in our proposed system.

5.2 Softmax classifier

In our proposed system, the softmax classifier [37], based on softmax regression, was used to estimate the probabilities of the current input, i.e., the m-th sample $\mathbf{u}_{(n,m)}$ in the n-th block, belonging to the orientation index j, by

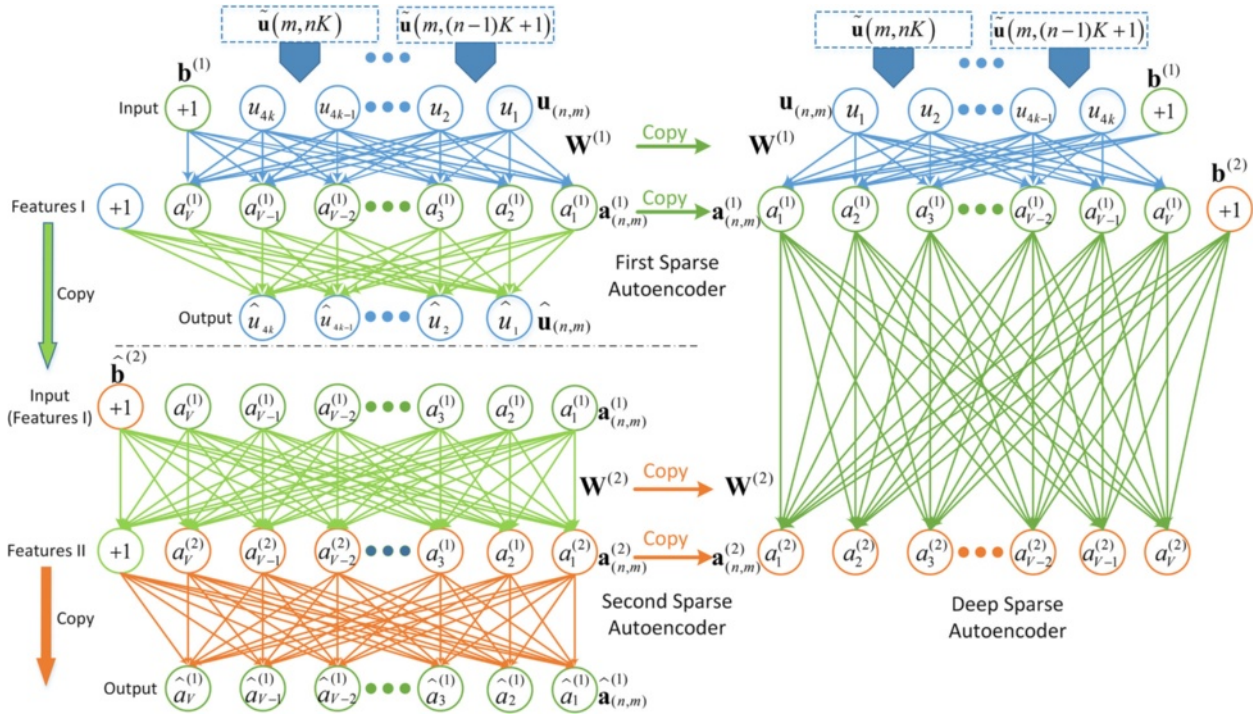

Fig. 5 The illustration of greedy layer-wise training and stacking. The *left part* of the figure shows the procedure of greedy layer-wise training, and the *right part* of the figure shows the procedure of stacking of sparse autoencoders

the deep autoencoder with the extracted high-level features $\mathbf{a}_{(n,m)}^{(2)}$ as inputs of the classifier. The architecture of the softmax classifier we used is shown in Fig. 6. In our proposed system, we represent the label of the training dataset as a one-hot vector (with 1 for the target class and 0 for others): $\mathbf{g}_{(n,m)} \in \mathbb{R}^J$. Then, the cross-entropy loss (cost function) of the softmax classifier can be written as follows:

$$J_{\text{softmax}}(\mathbf{W}^{(3)}) = -\frac{1}{M}\left[\sum_{m=1}^{M}(\mathbf{g}_{(n,m)})^T \mathbf{h}_{\mathbf{W}}\left(\mathbf{a}_{(n,m)}^{(2)}+\mathbf{b}\right)\right] + \frac{\lambda}{2}\sum_{j=1}^{J}\sum_{i=1}^{I}(w_{j,i})^2$$

$$\mathbf{h}_{\mathbf{W}}\left(\mathbf{a}_{(n,m)}^{(2)}\right) = \frac{1}{\sum_{j=1}^{J} e^{\mathbf{w}_j\left(\mathbf{a}_{(n,m)}^{(2)}+\mathbf{b}\right)}}\left[e^{\mathbf{w}_1^T\left(\mathbf{a}_{(n,m)}^{(2)}+\mathbf{b}\right)}\cdots e^{\mathbf{w}_J^T\left(\mathbf{a}_{(n,m)}^{(2)}+\mathbf{b}\right)}\right]^T \tag{5}$$

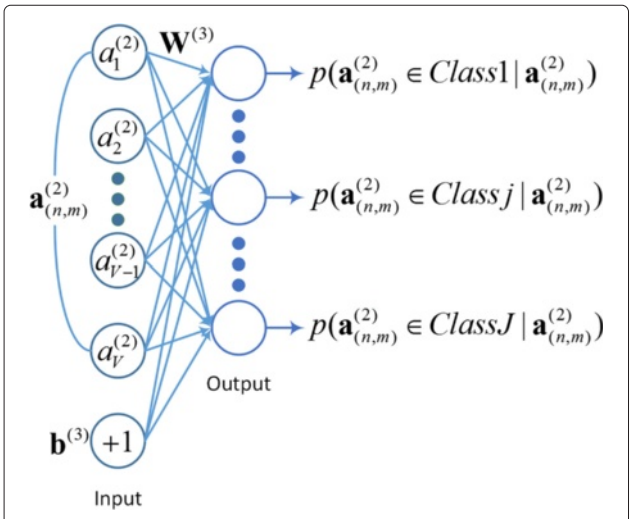

Fig. 6 A classical architecture of the softmax classifier

The softmax classifier can be trained by using the L-BFGS algorithm based on a dataset, in order to find an optimal parameter set $\mathbf{W}^{(3)}$ for minimizing the cost function $J_{\text{softmax}}\left(\mathbf{W}^{(3)}\right)$. In our proposed system, the dataset for softmax classifier training is composed by two parts. The first part is the input sample—$\mathbf{a}_{(n,m)}^{(2)}$ (features II), calculated from the last hidden layer of the deep autoencoder. The second part is the data label—$\mathbf{g}_{(n,m)} \in \mathbb{R}^J$, where the j-th element—g_j of $\mathbf{g}_{(n,m)}$ will be set to 1 when the input sample belongs to the source located in the range of DOAs of index j.

5.3 Stacking deep autoencoder and softmax classifier

We stack the softmax classifier and deep autoencoder together after the training is completed, as shown on the left part of Fig. 7. Finally, we use the training dataset and the L-BFGS algorithm to fine-tune the deep network with the initialized parameters $\mathbf{W}^{(1)}$, $\mathbf{b}^{(1)}$, $\mathbf{W}^{(2)}$, $\mathbf{b}^{(2)}$, $\mathbf{W}^{(3)}$, $\mathbf{b}^{(3)}$ obtained from the sparse autoencoders and softmax classifier training. The training phase of the sparse autoencoders and softmax classifier are called pre-training phase, and the stacking/training of the overall network, i.e., deep network, is called fine-tuning phase. In the pre-training phase, the shallow neural networks, i.e., sparse autoencoders and softmax classifier, are training individually, using the output of current layer as the input for the next layer. In the fine-tuning phase, we use the L-BFGS algorithm (i.e., a gradient descent method) to minimize the difference between the output of the deep network and the label of the training dataset. The gradient descent works well because the initialized parameters obtained from the pre-training phase include a significant amount of "prior" information about the input data through unsupervised learning [11].

6 Experiments

In this section, we first describe the generation of the datasets for training and testing and the setup of the training parameters of the deep networks. Similar to [7], different sentences from different speakers were convolved with real BRIRs to generate the stereo mixtures with room effects. The algorithms in [7] and [42] are used as baselines. We then apply both our proposed system and the basline algorithms to these mixtures to separate the target source. The separation quality is evaluated in terms of both signal distortion and perceptual speech quality.

6.1 Dataset generation

Similar to [7], the datasets that we used for training and testing are generated by the convolution of the original speech signal with real BRIRs. The original speech sources (target and interferer) were randomly selected from the TIMIT dataset which is a continuous speech corpus containing 6300 sentences: 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the USA [43]. 10 sentences spoken by 2 female speakers, who were randomly selected from the training usage set of the TIMIT, as the training dataset; another 30 sentences spoken by 4 male and 2 female speakers, who were randomly selected from the test usage set of the TIMIT, as the test dataset, where 10 sentences spoken by 2 males as the target source, 5 sentences spoken by 1 male, and 5 sentences spoken by 1 female as the interferer 1; and the remaining sentences which were spoken by 1 male and 1 female as the interferer 2. Details about the sentences and speaker IDs can be found in Table 1. All the sentences were normalized to have equal root mean square magnitude and cut to a same length (about 2.6 s for each sentence) for consistency.

The BRIR datasets used in our experiments were recorded using a dummy head and torso in five different types of room, named as X, A, B, C, and D at the University of Surrey, measured by Hummersone [44] and can be download from the website [45]. Room X is a very large room, and the reflections were truncated in the recordings

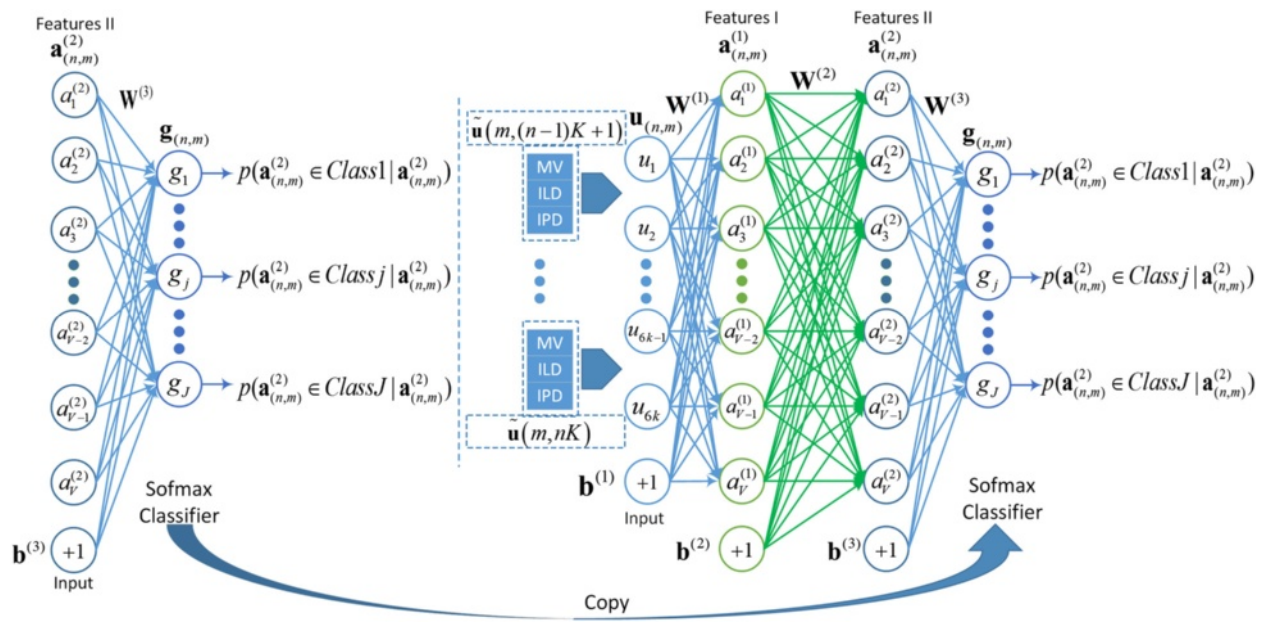

Fig. 7 The illustration of stacking deep autoencoder and softmax classifier

Table 1 Details about the speakers and the sentences, including speakers and sentences ID, their genders, dialect regions (DR), used in training or test dataset and used as the target or interferers

ID	Sex	DR	Sentence ID					Dataset	Source type
TBR0	F	1	21	201	381	111	291	Training	/
MEM0	F	1	297	333	207	387	117	Training	/
BJK0	M	2	95	275	5	185	365	Testing	Target
RMS1	M	7	407	137	317	47	227	Testing	Target
CTT0	M	5	28	208	388	118	298	Testing	Interferer 1
PKT0	F	3	8	188	368	98	278	Testing	Interferer 1
BDG0	M	3	383	113	293	23	203	Testing	Interferer 2
UTB0	F	5	124	304	34	214	394	Testing	Interferer 2

to produce anechoic recordings. We aim to evaluate the speech separation quality in reverberant environments. For this reason, in our experiments, we only used the BRIRs recorded in rooms A, B, C, and D. Different from other similar datasets, such as [46], this dataset has higher angular resolution and many different acoustic properties, which enabled us to evaluate the performance of the system over different acoustic environments with finer resolution. Table 2 shows the different acoustical properties of the rooms used in our evaluation. In each room, acoustic sources were placed 1.5 m away from the dummy head and had the same height as the dummy head, and the head related transfer function (HRTF) is applied in the BRIRs to mimic sound sources that would have been heard by human ears.

In our experiments, the training dataset, used to train the deep networks, is generated by the convolution of real BRIRs with the clean speech signals of two randomly selected speakers from the TIMIT dataset. More specifically, we use the speech signals observed at the microphones with a single source placed in different orientations with respect to the microphones, rather than the mixtures, to train the DNNs, and the orientations of the source are used as the ground truth. We consider speakers of different genders in training and test dataset for the evaluation of the generalization ability of the proposed system. More specifically, the training data set is generated by the convolution of clean New England female speech signals with all the real BRIRs (from $-90°$ to $+90°$ with a step of 5°), and the sentences spoken by the male speakers from a different dialect region are (DR in Table 1) are used as the target source in the test set.

Table 2 Room acoustic properties in initial time delay gap (ITDG), direct-to-reverberant ratio (DRR), and reverberation time (T60)

Room	Type	ITDG (ms)	DRR (dB)	T60 (s)
A	Medium office	8.73	6.09	0.32
B	Small class room	9.66	5.31	0.47
C	Large lecture theatre	11.9	8.82	0.68
D	Large seminar room	21.6	6.12	0.89

Different from the training dataset, the test set is composed by mixtures and used for the evaluation of speech separation quality, including determined and underdetermined cases, i.e., for two sources (target and interferer 1) and three sources (target, interferer 1 and interferer 2) with just two microphones as receivers. More specifically, similar to [7], the mixtures in the test dataset were generated by adding the reverberant target and interfering signals together which is equivalent to assuming superposition of their respective sound fields. The target and interfering signals are the randomly selected sentences from different male and female speakers, each convolved with the real BRIRs. For the determined case, the target source was located at 0° azimuth, and the azimuth of interferer 1 is varied from $-90°$ to $+90°$ with the step of 5°. For the underdetermined case, we add the speech signals from interferer 2 which was located at 30° to the mixtures of the determined case.

6.2 Experimental setting

Even though all the sources (including both the target and interferers) at different azimuths are recovered in our proposed system, the performance of the system is reported based on the quality of the recovered target located at 0° azimuth, with the azimuths of the interferers varied from $-90°$ to $+90°$ with step of 5°, similar to [7]. The sampling rate f_s used in signal sampling, STFT and ISTFT operation was 16 kHz ($f_s = 16$ kHz). We used a Hanning window of 2048 (128 ms) samples with 75 % overlap between the neighboring windows for the STFT. The frequency grouping parameters K and N are set to 16 and 128, respectively. Hence, we use 128 deep networks to generate the soft mask, with each deep network corresponding to a block. For each deep network, the input layer includes 96 units and $V = 256$ neurons for each of the hidden layers. $J = 37$ neurons were used in the output layer, corresponding to azimuths from $-90°$ to $+90°$ with a step of 5°.

The learning parameters are set as follows, the weight decay parameter $\lambda = 1 \times 10^{-4}$, the weight of the penalty term $\beta = 3$, and the sparsity parameter $\rho = 4 \times 10^{-3}$. The maximum number of iterations is set to 300. The parameters for the training of the softmax classifier are set as follows. The weight decay parameter $\lambda = 1 \times 10^{-4}$ and the maximum number of iterations was set to 200. In the fine-tuning phase, the weight decay parameter was changed to $\lambda = 3 \times 10^{-3}$.

For speech separation performance evaluation, we consider SDR [47] and PESQ [48] and the algorithms in [7, 42] as the baseline. In the evaluation, we consider both determined and undetermined cases and test the performance of our proposed system in different reverberation conditions, spatial diffuse noises, training dataset conditions, unseen rooms, block size K, and network types. The separation results including the comparison with the baseline methods are shown in Section 6.3.

6.3 Experimental results

In this section, we first test the performance of the proposed system under different training dataset configurations (the training set with full or half of the azimuths as discussed earlier) and different levels of reverberation for determined and underdetermined cases. Finally, we present the separation results for the mixtures corrupted by different levels of spatially diffuse noises (SNR = 5 and 10 dB).

6.3.1 Reverberation effect

We use the four reverberant rooms, i.e., rooms A, B, C and D, to evaluate the performance. The acoustical properties of the four rooms can be found in Table 2 in Section 6.1.

Figure 8 presents the SDRs of the separated signals with different DOAs of the interferer 1 and different rooms, for the determined case, where the deep networks used for soft mask generation were trained by the training set. Compared with the two baseline methods, we obtain at least 2-dB improvement in rooms B and D, when the interfering speech is placed far away from the target source. However, we obtain similar performance to the baseline in rooms A and C. It can be seen that with different reverberation times (T60s) and direction to reverberation ratios (DRRs), the proposed system performs generally more robust than the baseline methods and the performance of the proposed system does not decrease as much as the baseline methods when the level of room reverberation increases. Similar to [7], it can be seen that the separation

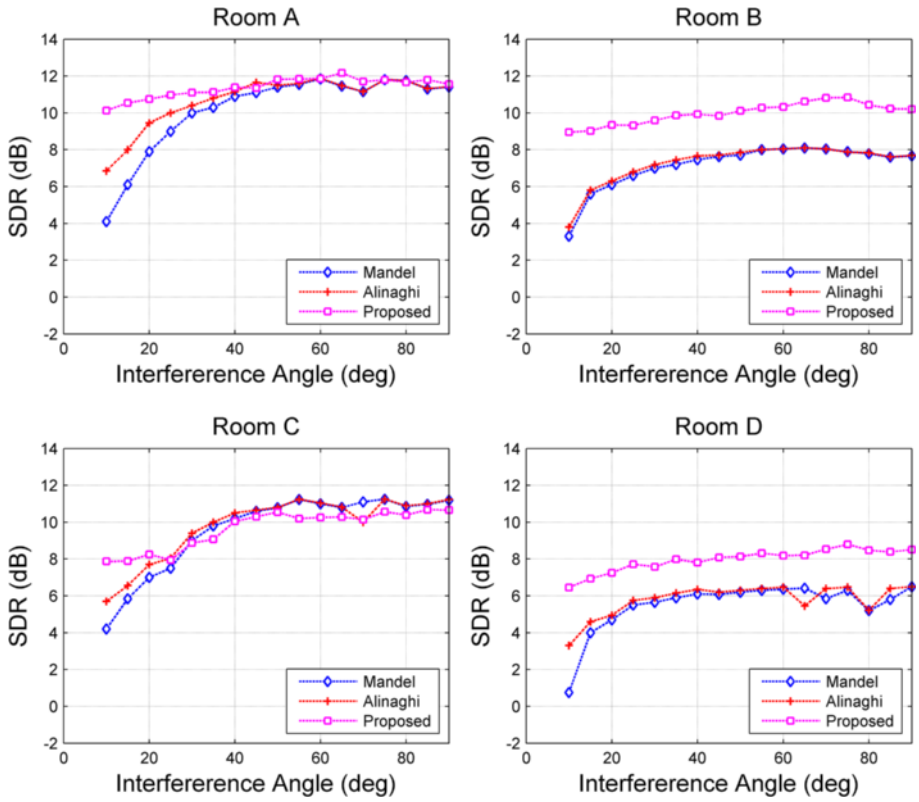

Fig. 8 SDRs performance comparison between the proposed system and the baseline methods among the rooms A, B, C, and D, in determined case, without noise

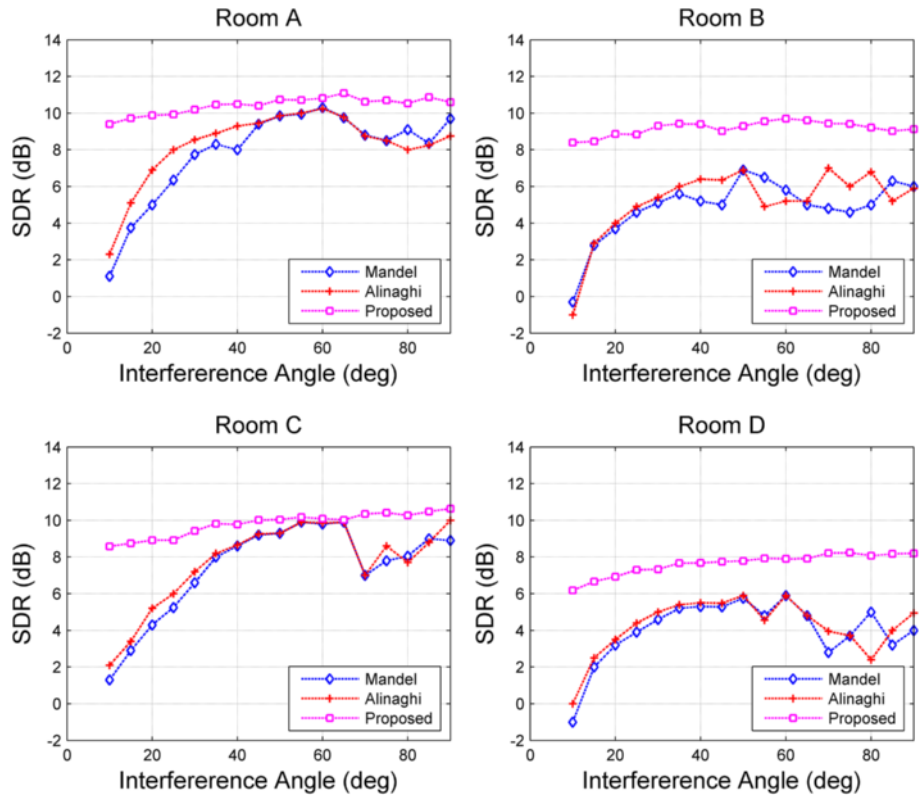

Fig. 9 SDRs performance comparison between the proposed system and the baseline methods among the rooms A, B, C, and D, in underdertermined case, without noise

quality of our proposed system depends on the acoustic parameters T_{60} and DRR.

The separation result for the underdetermined case is presented in Fig. 9. It can be seen that, compared with the two baseline methods, we obtain about 1-dB improvement for rooms B and D, and similar performance for rooms A and C, except for the situation that the target and interferer are close to each other. Compared with Fig. 8, it can be seen that the SDRs of the proposed system decrease for about 1 dB and the performance of our proposed system decreases with the increase in the number of the sources within the mixtures. Compared with Fig. 8, it can be seen that the SDRs of the proposed system decrease for about 2 dB and the PESQs of the proposed system decrease about 0.5. It can be seen that, similar to the baseline methods, the performance of our proposed system decreases with the increase in the number of sources within the mixtures.

From Figs. 8 and 9, we see that the proposed system is more robust to the acoustic parameters, i.e. the DRRs and T60s than the baseline methods, with at least 1 dB improvement in SDR. A summary of the PESQ results is represented in Table 3.

The comparison between the proposed system and the baseline methods suggests that the deep networks are able to provide more robust estimation results for the time-frequency mask even though blocking was used in our system.

6.3.2 Spatially diffuse noise

Similar to [7], we also evaluated the performance of the proposed system in the case of the mixtures corrupted by spatially diffuse noise. Same as Section 6.3.1, we repeat the experiments, but adding two different levels of noise

Table 3 Results of the baseline methods and the proposed method, for reverberant mixtures with the average over rooms A, B, C, and D in perceptual evaluation of speech quality (PESQ)

Case	Methods	Room A	Room B	Room C	Room D	Mean
Determined	Mandel	2.34	2.07	2.34	1.96	2.18
	Alinaghi	2.39	2.11	2.38	2.01	2.22
	Propsoed	2.34	2.22	2.30	2.14	2.25
Underdetermined	Mandel	2.1	1.85	2.14	1.81	1.98
	Alinaghi	2.15	1.87	2.18	1.84	2.01
	Propsoed	2.06	1.97	2.19	1.85	2.02

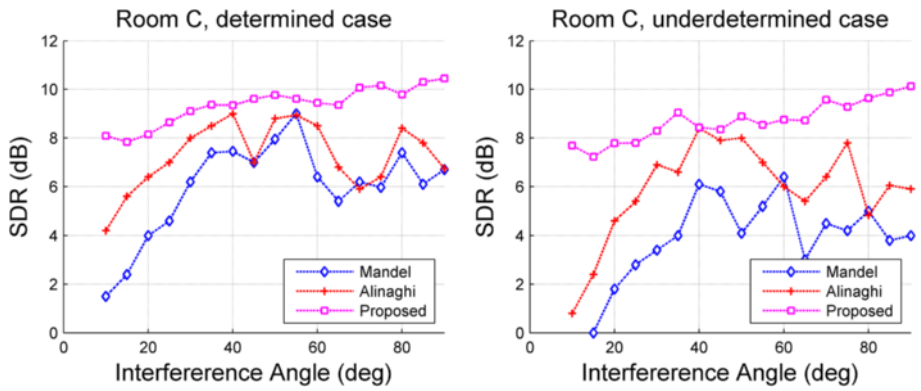

Fig. 10 SDR performance comparison between the proposed system and baseline methods in room C, with the SNR = 10 dB, for the determined and underdetermined cases

in the mixtures, i.e., the signal-to-noise-ratios (SNRs) were set to 5 and 10 dB (with respect to the mixture), respectively.

Figure 10 presents the SDRs comparison between the proposed system and the baseline methods in room C, with the SNR = 10 dB, for the determined and underdetermined cases. It can be seen that, for the determined case, the proposed system gives about 1 dB improvement in all of the azimuths, and for the underdetermined case, it also gives about 1 dB improvement in most of the azimuths. Similar to the results without noise, the performance of the deep network-based time-frequency masking technique also decreases with the increase in the number of sources presented in mixtures.

Figure 11 presents the SDRs for determined and underdetermined cases with SNR = 5 or 10 dB of the mixtures. We see that the SDRs of the proposed system decrease

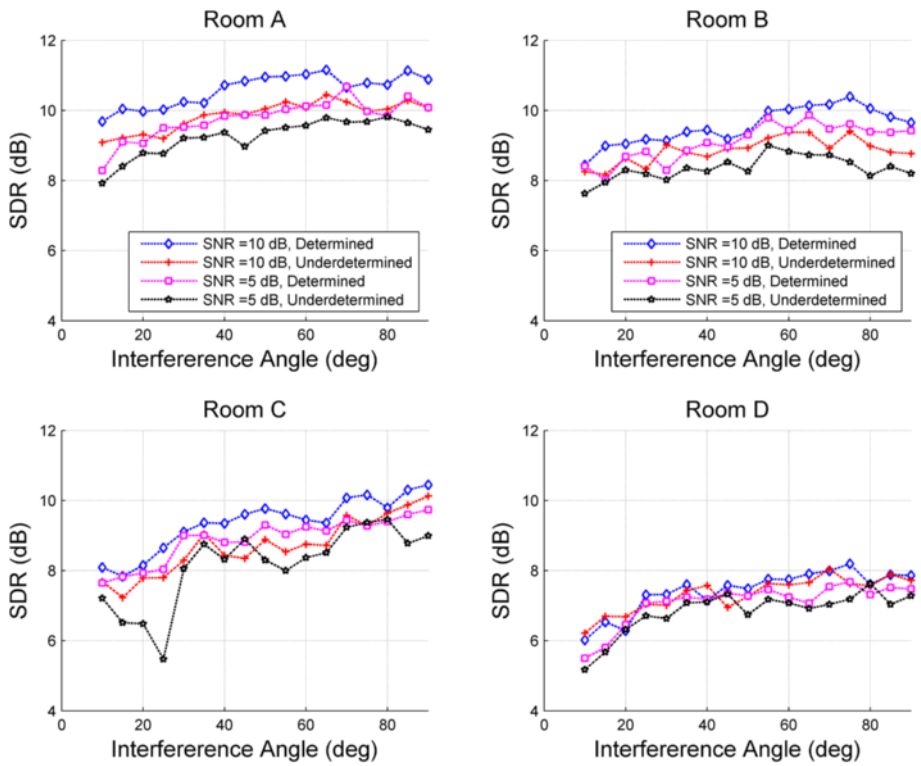

Fig. 11 SDRs for determined and underdetermined cases with SNR = 5 dB or 10 dB of the mixtures

about 1.5 dB in determined case and about 1 dB in underdetermined case when the SNR of mixtures is varied from 10 to 5 dB. Furthermore, as compared with the SDRs in Section 6.3.1 without noise, we can see that there is only about 1 dB performance drop when the SNR = 10 dB.

The Fig. 12 shows a separation example for room D, including the spectrogram of the mixture signals (Fig. 13a), original target source signals (Fig. 13b), separated target source signals (Fig. 13d), and the soft mask for separation (Fig. 13c), with the deep networks trained using the full training set, for the determined case (the interferer 1 was located at +15°).

6.3.3 Generalization to different rooms

In this subsection, we consider the generalization performance of our proposed system to unseen rooms in the determined and underdetermined cases. To this end, we selected each of the BRIRs recording from the four rooms in turn to generate the training set and use all the BRIRs to generate the test set. For instance, as shown in the top left plot of Fig. 13, we choose the BRIR of the room A to generate the training dataset and use all the BRIRs to generate the test dataset. The interferer is varied from −90° to +90° with a step of 5°. As shown in Fig. 13 (determined case) and Fig. 14 (underdetermined case), the system that was trained by the BRIRs of room D got the best generalization performance and the system trained by the BRIR of room A got the worst. Consider the different acoustic properties of these four rooms, we could find that the generalization performance of the proposed system increases with the complexity of the acoustic properties of the room. Compared with Figs. 8 and 9, the SDR performance of the proposed system trained by room D decreases about 4 dB in rooms A and C and increases about 1 dB in rooms B and D, both in determined and underdetermined cases.

6.3.4 Evaluation in different block size K

As mentioned in Section 3, we group K frequency bins to a block and use a corresponding DNN to generate a probability from the input features for each block. In this subsection, we evaluate the effect of different block size K for the determined case. As shown in Fig. 15, the system gives the best performance when the block size $K = 4$ and the SDR performance decreases with the increase of the K. However, we chose $K = 8$ in our proposed system,

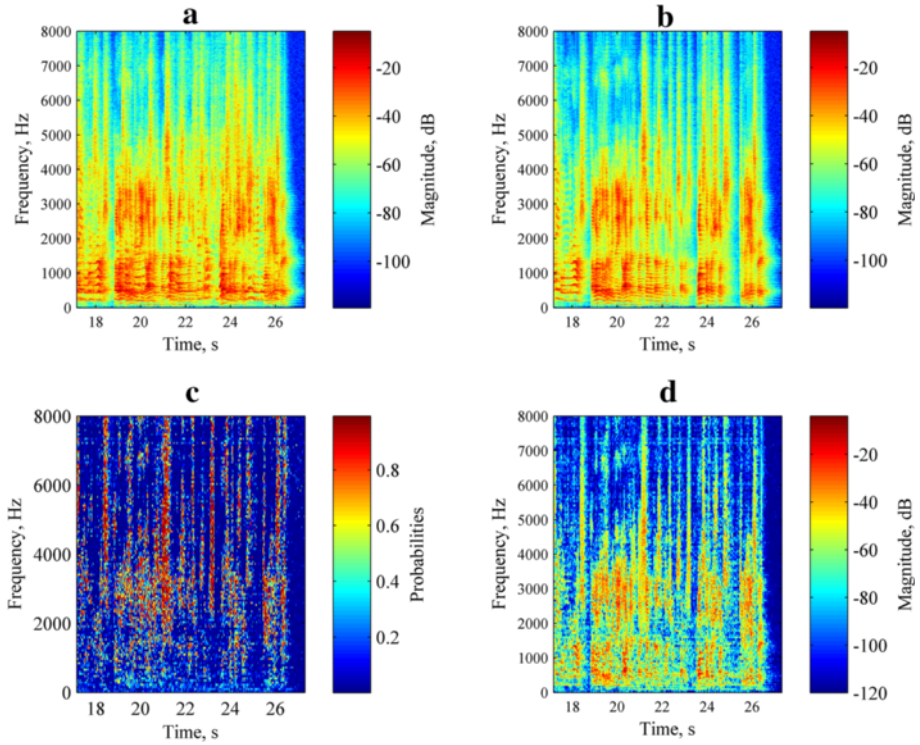

Fig. 12 A separation example for room D, including the spectrogram of the mixture signals, original target source signals, separated target source signals, and soft mask for separation. The mixtures include two sources (interferer 1 was located at +15°), without the noise. **a** Magnitude spectrogram of the mixture signals. **b** Magnitude spectrogram of the original target source signals. **c** The soft mask for separation. **d** Magnitude spectrogram of separated target source signals

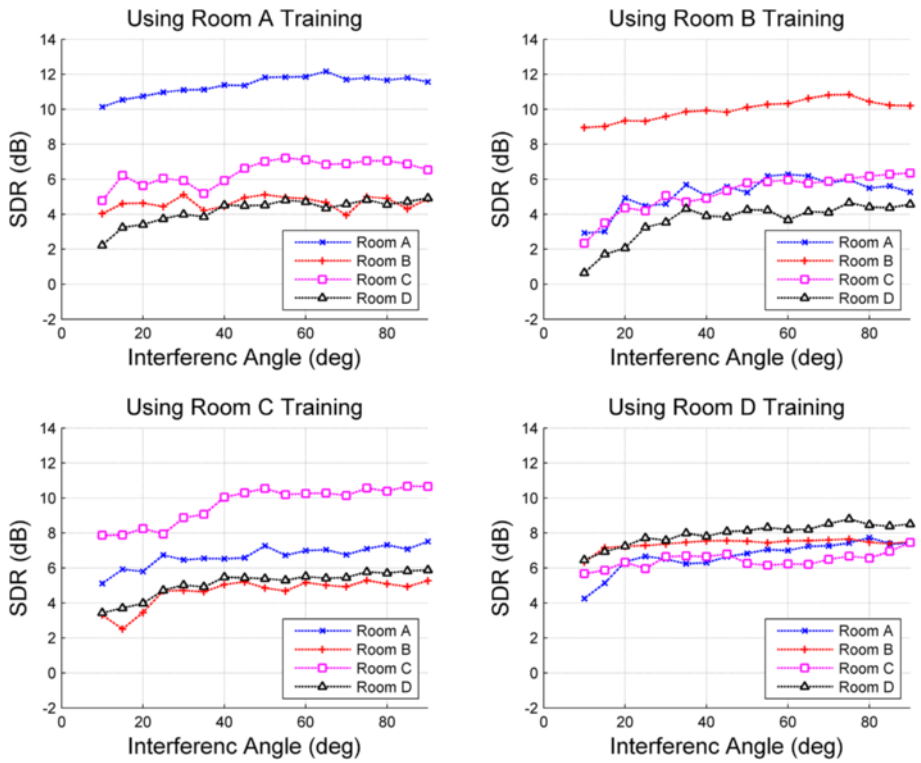

Fig. 13 SDRs to unseen rooms, in determined case. We select each of the four rooms to generate the training dataset and use the BRIRs from the four rooms (one by one respectively) to generate the test dataset. The interferer is varied from −90° to +90° with the step of 5°

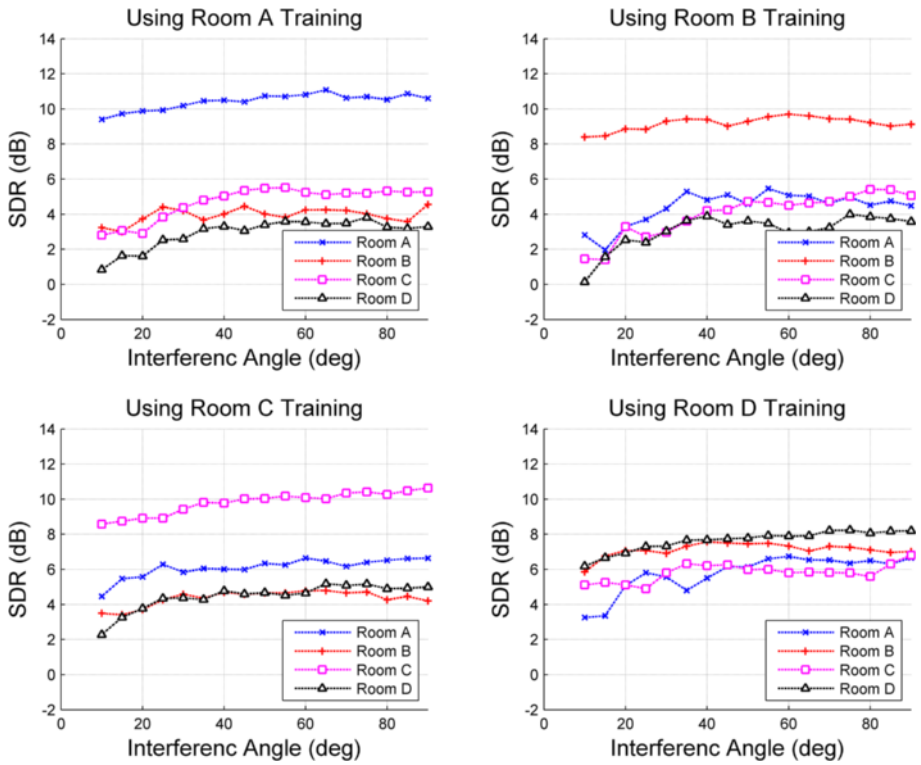

Fig. 14 SDR to unseen rooms, in underdetermined case. We select each of the four rooms to generate the training dataset and use the BRIRs from the four rooms (one by one respectively) to generate the test dataset. The interferer is varied from −90° to +90° with the step of 5°

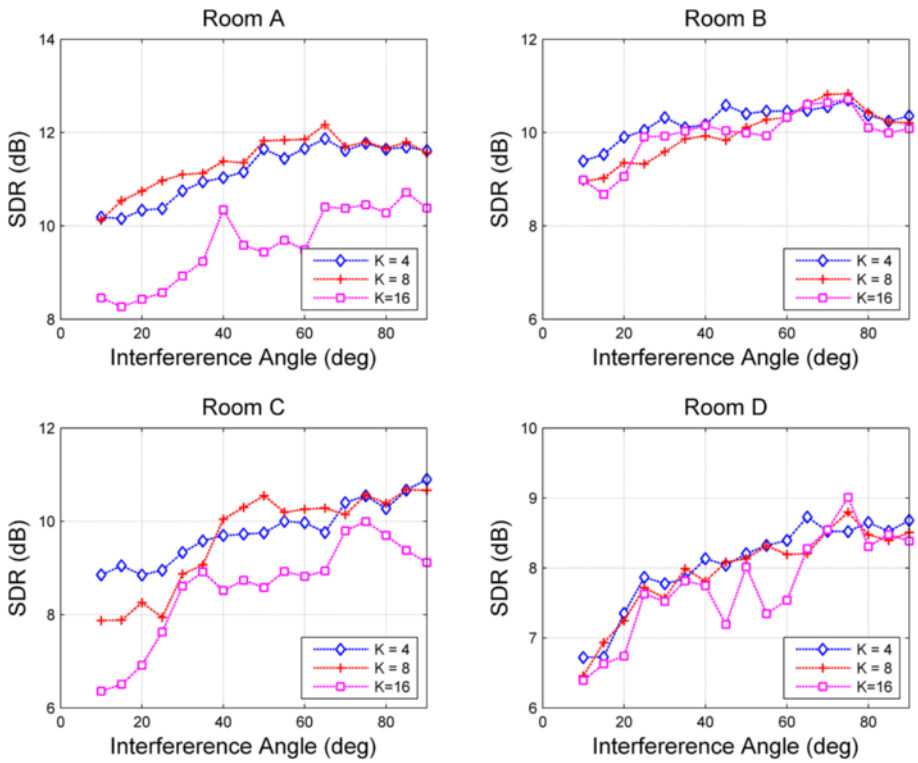

Fig. 15 SDR performance vs. different *K*, in determined case

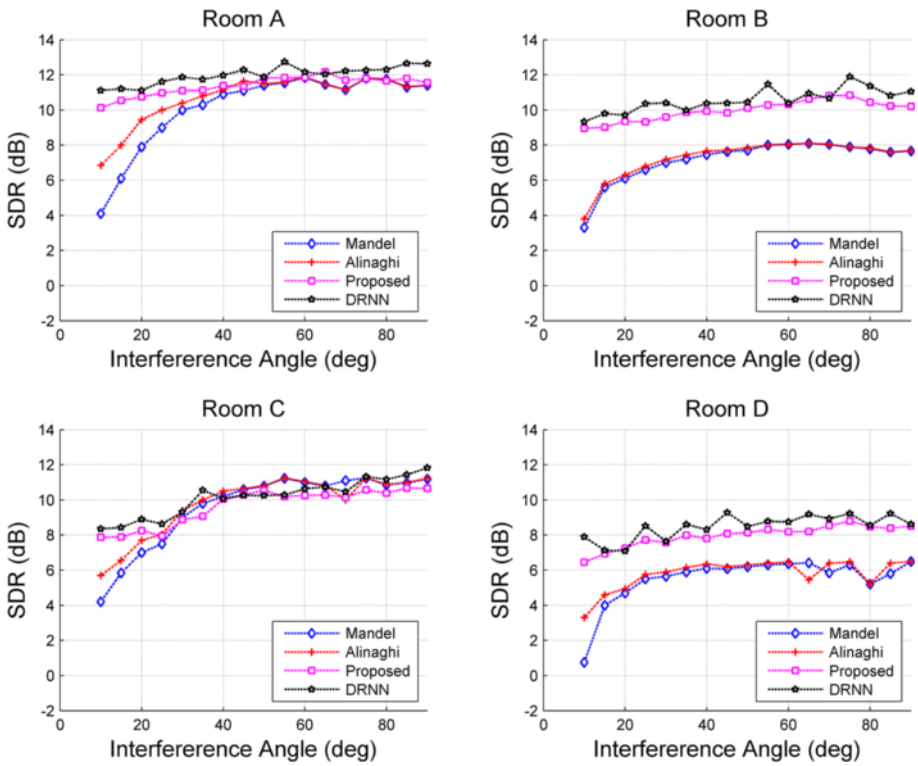

Fig. 16 SDRs comparisons among the proposed system, the RNNs method, Mandel method, and Alinaghi method, in underdetermined case

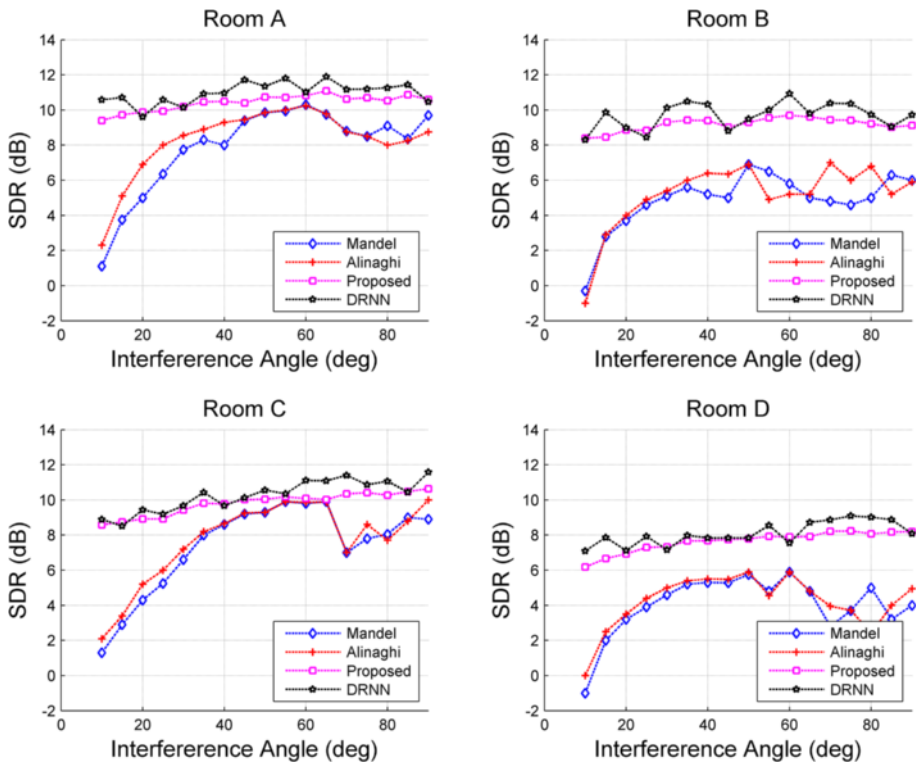

Fig. 17 SDRs comparisons among the proposed system, Mandel method, Alinaghi method, and the RNNs method, in underdetermined case

for the similar SDR performance and less computational complexity.

6.3.5 Evaluation in different neural network type

Room reverberation effect on speech signal can be regarded as signal extension in time. From this viewpoint, the recurrent neural networks (RNNs) may perform better than the DNNs in dealing with reverberation. In this subsection, we evaluate the use of the deep recurrent neural networks (DRNNs) in our system, instead of using DNNs. The DRNNs which were originally used by Huang et al. for monaural speech separation [27, 28] are employed here. The differences between the proposed system and the method in [27, 28] reside in the training dataset and ground truth. More specifically, we use the orientations of the sources as the ground truth and the isolated observed speech signals as training dataset, instead of using the separated source speech and mixture as the ground truth and training dataset. As shown in Figs. 16 and 17, we compare the SDR performance among the proposed system (DNNs method), the RNNs method, the Mandel method, and the Alinaghi method in four rooms, for the determined and underdetermind cases. It can be seen that the RNNs method get the best performance in all rooms, with about 1 dB improvement over the DNNs method. It is worth noting that the computational complexity of the RNNs method appears to be high and deserves further study in our future work.

7 Conclusions

We have presented a new localization-based stereo speech separation system using deep networks. Compared with GMM/EM-based algorithm in [7, 42], the deep network-based techniques provide better results in SDR and PESQ when room reverberation is presented in the mixtures. It is also shown that they are robust to spatially diffuse noise. In our future work, it would be interesting to compare the proposed method with other existing deep network-based separation algorithms such as [30, 31].

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This study was performed when the first author was an academic visitor in the Center for Video Speech and Signal Processing, University of Surrey, and he wishes to thank Qingju Liu, Philip JB Jackson, and Atiyeh Alinaghi for providing help for issues related to the algorithm in [7]. This research was supported partially by the Natural Science Basis Research Plan in Shaanxi Province of China (Program No.2014JQ8355). The authors wish to thank the anonymous reviewers for their helpful comments in improving this paper.

Author details
[1]School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an, China. [2]Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, UK.

Received: 15 December 2014 Accepted: 22 February 2016
Published online: 04 March 2016

References

1. P Comon, C Jutten, *Handbook of blind source separation: independent component analysis and applications*. (Academic Press, New York, 2010)
2. A Hyvärinen, J Karhunen, E Oja, *Independent component analysis*, vol. 46. (Wiley, New York, 2004)
3. A Hyvärinen, E Oja, Independent component analysis: algorithms and applications. Neural Netw. **13**(4), 411–430 (2000)
4. BD Van Veen, KM Buckley, Beamforming: a versatile approach to spatial filtering. IEEE ASSP Mag. **5**(2), 4–24 (1988)
5. D Wang, GJ Brown, *Computational auditory scene analysis: principles, algorithms, and applications*. (Wiley-IEEE Press, New York, 2006)
6. GJ Brown, D Wang, in *Speech Enhancement*. Separation of speech by computational auditory scene analysis (Springer, Berlin Heidelberg, 2005), pp. 371–402
7. A Alinaghi, PJ 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). doi:10.1109/TASLP.2014.2320637
8. H Sawada, S Araki, S Makino, Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment. IEEE Trans. Audio Speech Lang. Process. **19**(3), 516–527 (2011)
9. MI Mandel, RJ Weiss, DP Ellis, Model-based expectation-maximization source separation and localization. IEEE Trans. Audio Speech Lang. Process. **18**(2), 382–394 (2010)
10. G Kim, Y Lu, Y Hu, PC Loizou, An algorithm that improves speech intelligibility in noise for normal-hearing listeners. J. Acoust. Soc. Am. **126**(3), 1486–1494 (2009)
11. Y Bengio, A Courville, P Vincent, Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intel. **35**(8), 1798–1828 (2013). doi:10.1109/TPAMI.2013.50
12. J Schmidhuber, Deep learning in neural networks: an overview. ArXiv e-prints (2014). arxiv: url1404.7828
13. D Yu, L Deng, Deep learning and its applications to signal and information processing [exploratory dsp]. IEEE Signal Process. Mag. **28**(1), 145–154 (2011). doi:10.1109/MSP.2010.939038
14. Y Bengio, Learning deep architectures for ai. Found. Trends® Mach. Learn. **2**(1), 1–127 (2009)
15. J Baker, L Deng, J Glass, S Khudanpur, C-H Lee, N Morgan, D O'Shaughnessy, Developments and directions in speech recognition and understanding, part 1 [dsp education]. IEEE Signal Process. Mag. **26**(3), 75–80 (2009)
16. J Baker, L Deng, S Khudanpur, C-H Lee, JR Glass, N Morgan, D O'Shaughnessy, Updated minds report on speech recognition and understanding, part 2 [dsp education]. IEEE Signal Process. Mag. **26**(4), 78–85 (2009). doi:10.1109/MSP.2009.932707
17. L Deng, in *Computational Models of Speech Pattern Processing*. Computational models for speech production (Springer, Berlin Heidelberg, 1999), pp. 199–213
18. L Deng, in *Mathematical Foundations of Speech and Language Processing*. Switching dynamic system models for speech articulation and acoustics (Springer, New York, 2004), pp. 115–133
19. L Deng, J Li, J-T Huang, K Yao, D Yu, F Seide, M Seltzer, G Zweig, X He, J Williams, Y Gong, A Acero, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Recent advances in deep learning for speech research at microsoft, (2013), pp. 8604–8608. doi:10.1109/ICASSP.2013.6639345
20. J-T Huang, J Li, D Yu, L Deng, Y Gong, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers, (2013), pp. 7304–7308. doi:10.1109/ICASSP.2013.6639081
21. SM Siniscalchi, J Li, C-H Lee, Hermitian polynomial for speaker adaptation of connectionist speech recognition systems. IEEE Trans. Audio Speech Lang. Process. **21**(10), 2152–2161 (2013). doi:10.1109/TASL.2013.2270370
22. SM Siniscalchi, D Yu, L Deng, C-H Lee, Exploiting deep neural networks for detection-based speech recognition. Neurocomputing. **106**, 148–157 (2013)
23. G Hinton, L Deng, D Yu, GE Dahl, A Mohamed, N Jaitly, A Senior, V Vanhoucke, P Nguyen, TN Sainath, B Kingsbury, Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. **29**(6), 82–97 (2012). doi:10.1109/MSP.2012.2205597
24. Y Wang, K Han, D Wang, Exploring monaural features for classification-based speech segregation. IEEE Trans. Audio Speech Lang. Process. **21**(2), 270–279 (2013)
25. Y Xu, J Du, L-R Dai, C-H Lee, An experimental study on speech enhancement based on deep neural networks. Signal Process. Letters, IEEE. **21**(1), 65–68 (2014)
26. A Narayanan, D Wang, in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Ideal ratio mask estimation using deep neural networks for robust speech recognition (IEEE, 2013), pp. 7092–7096
27. P-S Huang, M Kim, M Hasegawa-Johnson, P Smaragdis, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep learning for monaural speech separation (IEEE, 2014), pp. 1562–1566
28. P 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. **PP**(99), 1–1 (2015). doi:10.1109/TASLP.2015.2468583
29. Z Jin, D Wang, A supervised learning approach to monaural segregation of reverberant speech. IEEE Trans. Audio Speech Lang. Process. **17**(4), 625–638 (2009). doi:10.1109/TASL.2008.2010633
30. J Chen, Y Wang, D Wang, in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. A feature study for classification-based speech separation at very low signal-to-noise ratio, (2014), pp. 7039–7043. doi:10.1109/ICASSP.2014.6854965
31. Y Jiang, D Wang, R Liu, Z Feng, Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 2112–2121 (2014). doi:10.1109/TASLP.2014.2361023
32. Y Bengio, P Lamblin, D Popovici, H Larochelle, Greedy layer-wise training of deep networks. Adv. Neural Inf. Process. Syst. **19**, 153 (2007)
33. J Blauert, *Spatial hearing: the psychophysics of human sound localization*, (1997)
34. FL Wightman, DJ Kistler, The dominant role of low-frequency interaural time differences in sound localization. J. Acoust. Soc. Am. **91**(3), 1648–1661 (1992)
35. M Ranzato, FJ Huang, Y-L Boureau, Y LeCun, in *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*. Unsupervised learning of invariant feature hierarchies with applications to object recognition, (2007), pp. 1–8. doi:10.1109/CVPR.2007.383157
36. M Ranzato, Unsupervised learning of feature hierarchies. PhD thesis, NEW YORK UNIVERSITY (2009)
37. CM Bishop, et al., *Pattern recognition and machine learning*, vol. 1. (Springer, New York, 2006)
38. A Ng, Lecture notes on sparse autoencoders. http://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf
39. A Ng, Sparse autoencoder. CS294A Lecture notes. **72**, 1–19 (2011)
40. J Dean, G Corrado, R Monga, K Chen, M Devin, M Mao, A Senior, P Tucker, K Yang, QV Le, et al., in *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Large scale distributed deep networks, (NIPS, 2012), pp. 1223–1231
41. Y Chauvin, DE Rumelhart, *Backpropagation: theory, architectures, and applications*. (Psychology Press, London, 1995)
42. MI Mandel, RJ Weiss, DPW Ellis, Model-based expectation-maximization source separation and localization. Audio Speech Lang. Process. IEEE Trans. **18**(2), 382–394 (2010). doi:10.1109/TASL.2009.2029711
43. JS Garofolo, LF Lamel, WM Fisher, JG Fiscus, DS Pallett, DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc n.1. NASA STI/Recon Technical Report N. **93**, 27403 (1993)
44. C Hummersone, A psychoacoustic engineering approach to machine sound source separation in reverberant environments. PhD thesis, University of Surrey (2011)
45. Binaural room impulse responses captured in real rooms. http://iosr.surrey.ac.uk/software/index.php
46. BG Shinn-Cunningham, N Kopco, TJ Martin, Localizing nearby sound sources in a classroom: binaural room impulse responses. J. Acoust. Soc. Am. **117**(5), 3100–3115 (2005)

47. E Vincent, R Gribonval, C Fevotte, Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. **14**(4), 1462–1469 (2006). doi:10.1109/TSA.2005.858005
48. L Di Persia, D Milone, HL Rufiner, M Yanagida, Perceptual evaluation of blind source separation for robust speech recognition. Signal Process. **88**(10), 2578–2583 (2008)