Speech Communication 140 (2022) 87–97

Contents lists available at ScienceDirect

Speech Communication

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

Deep ad-hoc beamforming based on speaker extraction for target-dependent speech separation

Ziye Yang, Shanzheng Guan, Xiao-Lei Zhang *

School of Marine Science and Technology and Center for Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, China
Research & Development Institute of Northwestern Polytechnical University in Shenzhen, China

ARTICLE INFO

Keywords:
Ad-hoc microphone array
Speaker extraction
Channel selection
Deep ad-hoc beamforming

ABSTRACT

Recently, the research on ad-hoc microphone arrays with deep learning has drawn much attention, especially in speech enhancement and separation. An ad-hoc microphone array may cover such a large area where multiple speakers stand far apart and talk independently. Therefore, it is important to extract and trace a specific speaker in the ad-hoc array, which is called target-dependent speech separation, aiming to extract a target speaker from a mixed speech. However, this technique has not been explored yet. In this paper, we propose deep ad-hoc beamforming based on speaker extraction, which is to our knowledge the first work for target-dependent speech separation based on ad-hoc microphone arrays and deep learning. The algorithm contains three components. First, we propose a supervised channel selection framework based on speaker extraction, where the estimated utterance-level SNRs of the target speech are used as the basis for the channel selection. Second, we apply the selected channels to a deep learning based MVDR algorithm, where a single-channel speaker extraction algorithm is applied to each selected channel for estimating the mask of the target speech. We conducted an extensive experiment on WSJ0-adhoc corpus and Libri-adhoc40 corpus. Experimental results demonstrate the effectiveness of the proposed method in both simulation and real scenarios.

1. Introduction

Speech separation, also known as cocktail party problem, aims to separate target speech from interference background (Wang and Chen, 2018). It is often used as the front end of speech recognition for improving the accuracy of human–machine interaction. Conventional speech separation technologies include computational auditory scene analysis (Rouat, 2008), non-negative matrix factorization (Schmidt and Olsson, 2006; Virtanen, 2007), HMM-GMM (Virtanen, 2006; Stark et al., 2010), and minimum mean square error (Ephraim and Malah, 1985). Recently, deep learning based speech separation becomes a new trend (Wang and Wang, 2012, 2013; Wang et al., 2014; Zhang and Wang, 2017; Delfarah and Wang, 2019; Delfarah et al., 2020), which is the focus of this paper. According to whether speakers' information is known as a prior, deep-learning-based speech separation techniques can be divided into three categories, which are speaker-dependent (Bregman, 1994), target-dependent, and speaker-independent speech separation. Speaker-dependent speech separation needs to know the prior information of all speakers, which limits its practical applications. Nowadays, the research on speech separation is mostly speaker-independent and target-dependent.

Speaker-independent speech separation based on deep learning faces the speaker permutation ambiguity problem. In order to solve this problem, two techniques have been proposed. The first one is deep clustering (DPCL) (Hershey et al., 2016; Wang et al., 2018a; Yang and Zhang, 2019b). It projects each time–frequency unit to a higher-dimensional embedding vector by a deep network, and conducts clustering on the embedding vectors for speech separation. The method in Isik et al. (2016) implements an end-to-end training strategy of DPCL via leveraging soft clustering, which further improves the performance of DPCL. The second technique is permutation invariant training (PIT) (Yu et al., 2017; Kolbæk et al., 2017; Xu et al., 2018). The network of PIT directly estimates the speech of each speaker. For each training mixture, it picks the permutation of the speakers that has the minimum training error among all possible permutations to train the network.

Target-dependent speech separation based on deep learning aims to extract target speech from a mixture given some prior knowledge on the target speaker. The earliest speech separation method takes the target speaker as the training target (Zhang and Wang, 2016). It has to train a model for each target speaker, which limits its practical use. To prevent training a model for each target speaker, *speaker extraction* further takes speaker codes extracted from a speaker recognition system as part of

* Corresponding author.
E-mail addresses: 2015300797@mail.nwpu.edu.cn (Z. Yang), gshanzheng@mail.nwpu.edu.cn (S. Guan), xiaolei.zhang@nwpu.edu.cn (X.-L. Zhang).

https://doi.org/10.1016/j.specom.2022.04.002
Received 26 August 2021; Received in revised form 6 April 2022; Accepted 7 April 2022
Available online 19 April 2022

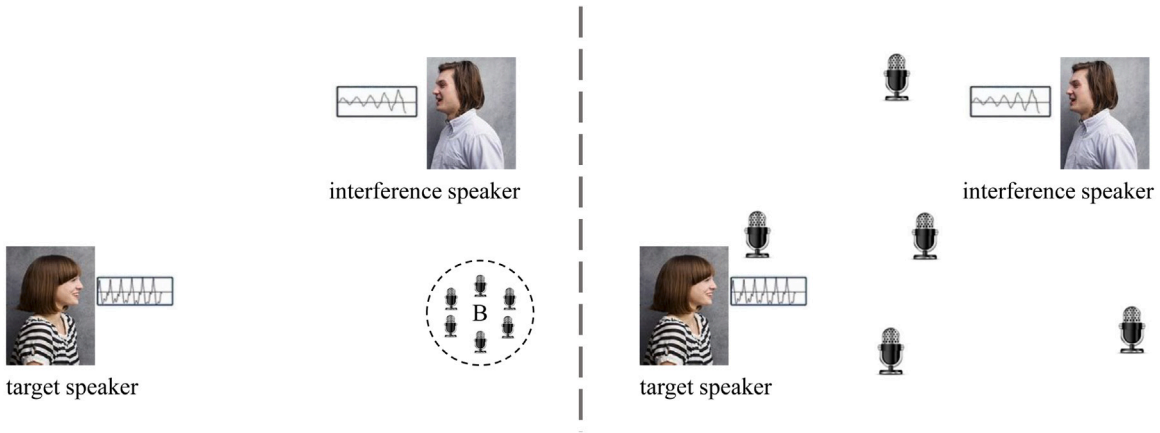

Fig. 1. The scenarios of speaker extraction under traditional microphone array and DABse.

the network input (Williamson and Wang, 2017; Žmolíková et al., 2017; Wang et al., 2018c; Xu et al., 2019; Xiao et al., 2019; Delcroix et al., 2019; Ochiai et al., 2019).

The aforementioned methods are all single-channel methods. Although they work well in clean scenarios, their performance degrades significantly in reverberant scenarios. To improve the performance of speech separation in reverberant scenarios, many multichannel methods were proposed in both fields of traditional algorithms and deep learning algorithms. Traditional algorithms can be categorized into NMF-based models (Ozerov and Févotte, 2009; Sawada et al., 2013; Kitamura et al., 2016, 2018) and probability models (Otsuka et al., 2013; Itakura et al., 2018) according to different source models. Besides, it is impressive that Higuchi et al. (2016) pointed out that the conventional steering vector estimators of beamforming might rely on inaccurate knowledge like the array geometry, and then addressed this problem by applying a complex Gaussian mixture model to estimate the time–frequency masks for estimating the steering vector (Higuchi et al., 2017). As for deep learning algorithms, there are currently two major forms. The first form combines spatial features that are extracted from microphone arrays, such as interaural time difference and interaural level difference, with spectral features as the input of single-channel speech separation networks (Wang et al., 2018b; Jiang et al., 2014; Araki et al., 2015; Pertilä and Nikunen, 2015). The second form uses a deep network to predict a mask for each speaker at each channel, and then conducts beamforming for each speaker (Nakatani et al., 2017). For brevity, we call this method *deep beamforming*. Žmolíková et al. (2019) proposed SpeakerBeam, which addresses the target speech extraction problem by utilizing the representation of a target speaker with a fixed microphone array. Gu et al. (2020) also proposed a multi-channel framework for this problem by making full use of the information of the target speaker. Some methods combined the above two forms for boosting their advantages together in reverberant scenarios, e.g. Yoshioka et al. (2018) and Yang and Zhang (2019a).

The aforementioned multichannel methods are only studied with traditional fixed arrays, such as linear arrays or spherical arrays. However, for far-field speech separation problems with high reverberation, they suffer significant performance degradation since the energy of speech signals gradually drops during their transmission through the air. How to maintain the estimated speech at the same high quality throughout an interested physical space is of broad interests. Ad-hoc microphone array, which is a group of randomly distributed microphones collaborating with each other, is a solution to the problem. Fig. 1 gives a comparison example where a target speaker extraction problem with a fixed array is on the left and that with an ad-hoc microphone array on the right. From the figure, we see that, compared with the fixed array that is far from the target speaker, the ad-hoc microphone array has several apparent advantages. First, an ad-hoc microphone array may put a number of microphones around the target speaker, which significantly reduced the probability of far-field speech processing. By channel selection, it might be able to form a local microphone array around the target speaker. At last, it may be able to incorporate application devices of various physical sizes.

In literature, ad-hoc microphone arrays have consistently been an important research topic (Jayaprakasam et al., 2017; Tavakoli et al., 2017; Zhang et al., 2017; Koutrouvelis et al., 2018). However, they face many practical problems due to the lack of important priors. Recently, Zhang (2018) addresses the difficulties of ad-hoc microphone arrays, such as lack of priors and insufficient estimation of variables, by deep learning for the first time. The proposed method, named deep ad-hoc beamforming (DAB), was originally designed for speech enhancement only, which predicts segment-level signal-to-noise-ratio (SNR) by deep neural networks for supervised channel selection. Later on, some speech separation methods based on ad-hoc microphone arrays were proposed. Luo et al. (2020) proposed a transform-average-concatenate strategy for a filter-and-sum network (Luo et al., 2019) to realize the channel reweighting/selection ability for ad-hoc microphone arrays. Because ad-hoc microphone arrays lack the prior of the number and spatial distribution of microphones, Wang et al. (2020) proposed a network architecture by interleaving inter-channel processing layers and temporal processing layers to leverage information across time and space alternately. Wang et al. (2021) further solved the problem of continuous speech separation by extending the method in Wang et al. (2020). Besides, speech recognition and speaker verification with ad-hoc microphone arrays also received much attention. Chen and Zhang (2021) proposed Scaling Sparsemax to address the channel selection problem of speech recognition with large-scale ad-hoc microphone arrays. Liang et al. (2021) proposed attention-based multi-channel speaker verification with ad-hoc microphone arrays for the missing prior information problem.

Among the above problems, deep learning based speech separation with ad-hoc microphone arrays has aroused our interest. And we find that existing methods are all speaker-independent. To our knowledge, target-dependent speech separation with ad-hoc microphone arrays are far from explored yet. In many applications, extracting and tracking target speech is of more interests than separating a mixture into its components. This is particularly the case for ad-hoc microphone arrays, where several speakers may locate far apart and talk independently.

In this paper, motivated from DAB (Zhang, 2018) for speech enhancement, we propose a target-dependent speech separation algorithm with ad-hoc microphone arrays, named DAB based on speaker extraction (DABse), which aims to address a different problem from DAB. The main problem to be solved is how to re-weight each channel in ad-hoc microphone arrays. Here, DABse predicts the speech quality of the target speaker received by each channel, i.e., SNR, for the weight

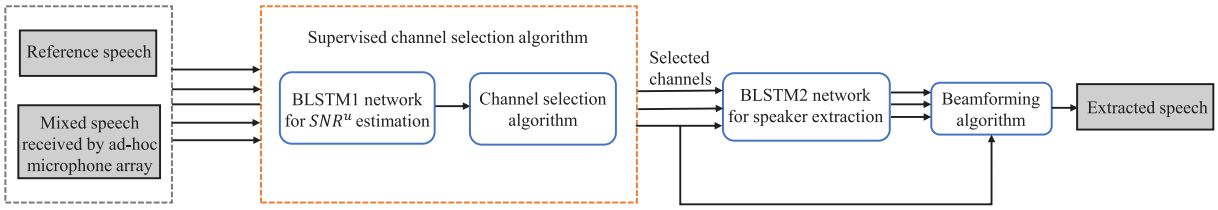

Fig. 2. Diagram of the proposed DABse system.

of each channel. However, compared with DAB, because the mixture contains multiple speakers, the network faces the speaker permutation ambiguity problem during the training stage, a well-known problem in speech separation as the aforementioned. In order to avoid this problem, we propose a new SNR estimation network, which adds codes of the target speaker to the network so as to make the network focus on the target speaker. Then, we apply a single-channel mask estimation network different from that of DAB. In addition, in our work, we tested DABse not only on simulated datasets but also on a semi-real dataset.

Our algorithm consists of three components: first, we propose a supervised channel selection based on speaker extraction, which applies bi-directional long short-term memory (BLSTM) networks to estimate the utterance-level SNR of the target speaker. Then, we employ the heuristic channel selection algorithms in Zhang (2018) to pick the channels with high SNRs. We further apply a single-channel speaker extraction algorithm to the selected channels for the mask estimation problem of the target speech. At last, we use the estimated masks to derive a beamformer for the target speaker, such as minimum variance distortionless response (MVDR) (Heymann et al., 2016). Experimental results on both a simulated WSJ0-adhoc corpus and a semi-real Libri-adhoc40 corpus show that the proposed DABse performs well in reverberant environments.

The rest of the paper is organized as follows. We present the deep ad-hoc beamforming system based on speaker extraction in Section 2. In Section 3, we present the experimental results. Finally, we conclude this study in Section 4.

2. Deep ad-hoc beamforming based on speaker extraction

We build the signal model for target-dependent speech separation based on ad-hoc microphone arrays. Suppose that a room contains a target speaker, an interference speaker, and an ad-hoc array of W microphones. Then, the mixed speech signal received by any single microphone of the ad-hoc array can be represented as:

$$y(t) = x_a(t) + x_i(t) + h(t) \tag{1}$$

where $x_a(t)$ and $x_i(t)$ are the direct speech of the target speaker and interference speaker at time t, and $h(t)$ is the early and late reverberation of the speech source signal. Note that, in our signal model, we ignore non-speech background noise.

We perform the short-time Fourier transform (STFT) to the signal (1), which results in:

$$Y(t,f) = X_a(t,f) + X_i(t,f) + H(t,f) \tag{2}$$

where $X_a(t,f)$ and $X_i(t,f)$ are the time–frequency units of the direct speech of the target speaker and interference speaker at time t and frequency f respectively, $H(t,f)$ is the time–frequency unit of the early and late reverberation. We can further define the direct speech as follows:

$$X_a(t,f) = c_a(f)S_a(t,f) \tag{3}$$

$$X_i(t,f) = c_i(f)S_i(t,f) \tag{4}$$

where $S_a(t,f)$ and $S_i(t,f)$ are the spectra of the target and interference speech at the source locations, and $c_a(f)$ and $c_i(f)$, which are complex numbers, are the time-invariant acoustic transfer functions from the speech sources to the microphone of the array.

Fig. 2 describes DABse. It first picks eligible channels from ad-hoc microphone arrays by a supervised channel selection algorithm based on speaker extraction for a target speaker. Then, it conducts deep-learning-based MVDR on the selected channels, where a separate single-channel speaker extraction network is used to estimate the mask of the target speaker.

2.1. Supervised channel selection based on speaker extraction

The main idea of the supervised channel selection based on speaker extraction is to select the channels with high SNR of the target speaker. The module contains two parts: a channel-weight estimation network, and a channel selection algorithm.

2.1.1. Channel-weight estimation network

The channel-weight estimation network aims to estimate the quality of the target speech for each channel. To make the channel-weight estimation network independent to the topology of ad-hoc microphone arrays, it needs to be trained in a single-channel fashion which is then applied to each channel separately in the test stage. Here we use a speaker extraction network to estimate the quality of the target speech, where an auxiliary network is to extract the identity feature of the target speaker. Fig. 3 shows the architecture of the channel-weight estimation network.

First of all, we need to define a training target. Many objective evaluation metrics are suitable to be used as the training target for evaluating the speech quality. As the first work of the target-dependent speech separation based on DAB, we take the simplest training target—a variant of SNR:

$$\mathrm{SNR}^u = \frac{\sum_t |x_a(t)|}{\sum_t |x_a(t)| + \sum_t |x_i(t)|}. \tag{5}$$

We name the variant of the SNR as the utterance-level SNR (SNR^u).

The network structure contains an auxiliary network and a main network. Suppose each target speaker has an auxiliary speech that is collected independently from the mixed speech $y(t)$. The auxiliary network takes the magnitude spectrum of the auxiliary speech $|A|$ as its input for extracting the identity embedding feature of the target speaker. It uses a BLSTM network to extract frame-level features from $|A|$,

$$[\mathbf{s}_1, \ldots, \mathbf{s}_g, \ldots, \mathbf{s}_G] = f(|A|;\theta). \tag{6}$$

where θ is the parameter of the BLSTM network, $[\mathbf{s}_1, \ldots, \mathbf{s}_g, \ldots, \mathbf{s}_G]$ is the output of the linear layer, and G represents the number of frames of the magnitude spectrum $|A|$. Then it uses a pooling layer to transform the frame-level features into an utterance-level embedding vector $\mathbf{v}$:

$$\mathbf{v} = \frac{1}{G}\sum_g s_g. \tag{7}$$

The main network contains a frame-level network, a pooling layer, and an utterance-level network from bottom-up. The frame-level network first transforms $|Y|$ to frame-level features, denoted as $[\mathbf{z}_1, \ldots, \mathbf{z}_b, \ldots, \mathbf{z}_B]$, then concatenates each frame-level feature $\mathbf{z}_b$ with $\mathbf{v}$, and finally extracts a target-dependent frame-level feature from $[\mathbf{z}_b^T, \mathbf{v}^T]^T$, where B

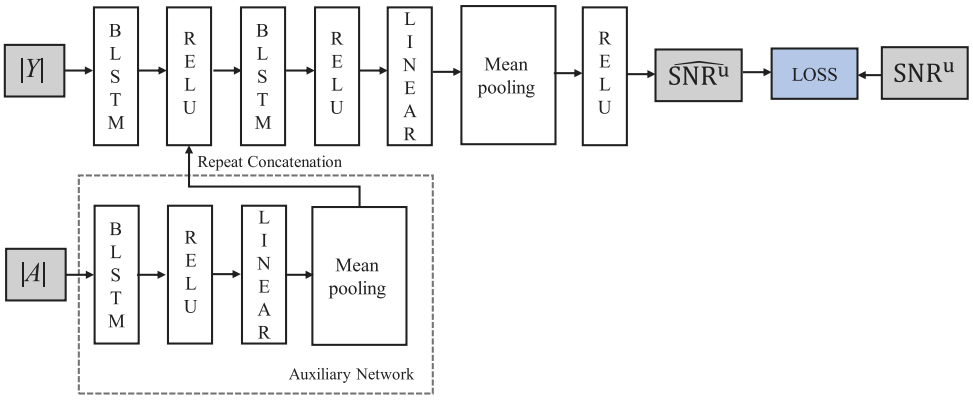

Fig. 3. Diagram of the SNR^u estimation network.

represents the number of frames of the magnitude spectrum $|Y|$. The pooling layer extracts a target-dependent utterance-level embedding from the target-dependent frame-level features. The utterance-level network is a regression network. It takes the utterance-level embedding as the input for predicting SNR^u of $y(t)$. It minimizes the mean-squared error:

$$J_1 = \|q - \text{SNR}^u\|_2^2 \tag{8}$$

where q is an estimate of SNR^u (denoted as $\widehat{\text{SNR}}^u$ in Fig. 3), which will be used as the channel weight for the channel-selection algorithm in the test stage.

The main network and auxiliary network are jointly trained by back-propagation. Both networks use mean pooling as the pooling layer, which averages the frame-level features along the time axis for the utterance-level embeddings.

2.1.2. Channel selection algorithms

The channel-selection algorithms in this section are used in the test stage only. Applying the channel-weight estimation network to each channel respectively gets a channel-weight vector $\mathbf{q} = [q_1, q_2, \dots, q_W]^T$. A channel-selection algorithm takes $\mathbf{q}$ as input, and outputs a channel-mask vector $\mathbf{p} = [p_1, p_2, \dots, p_W]^T$. Some channel-selection algorithms are described as follows.

- **Selecting one-best channel (1-best)**
  This algorithm selects the channel with the highest speech quality among the W channels:

$$p_j = \begin{cases} 1, & \text{if } q_j = \max_{1\le n\le w} q_n \\ 0, & \text{otherwise} \end{cases}, \forall j = 1, \dots, W \tag{9}$$

- **Selecting N-best channel with predefined number (fixed-N-best)**
  If the speakers are in a large room, and if the microphones are sufficiently dispersed, then selecting a number of microphones around the target speaker may yield better performance than using all microphones. This channel selection algorithm first sorts $\{q_1, q_2, \dots, q_W\}$ in descent order, denoted as $q'_1 \ge q'_2 \ge \cdots \ge q'_W$, and then picks the first N channels with the highest q:

$$p_j = \begin{cases} 1, & \text{if } q_j \in \{q'_1, q'_2, \dots, q'_N\} \\ 0, & \text{otherwise} \end{cases}, \forall j = 1, \dots, W \tag{10}$$

  where $N \le W$.
- **Selecting N-best channel where number is predetermined on-the-fly (auto-N-best)**
  This algorithm provides a method to determine N automatically. It first picks the 1-best channel by $q_* = max_{1\le n\le w} q_n$, and then calculates p_j by:

$$p_j = \begin{cases} 1, & \text{if } \frac{q_j}{q_*}\frac{1-q_*}{1-q_j} > \gamma \\ 0, & \text{otherwise} \end{cases}, \forall j = 1, \dots, W \tag{11}$$

  where $\gamma \in [0, 1]$ is a tunable hyperparameter.
- **Selecting soft N-best channel (soft-N-best)**
  Different from the auto-N-best algorithm, this algorithm re-weights the selected channels according to the quality of the target speech:

$$p_j = \begin{cases} q_j, & \text{if } \frac{q_j}{q_*}\frac{1-q_*}{1-q_j} > \gamma \\ 0, & \text{otherwise} \end{cases}, \forall j = 1, \dots, W \tag{12}$$

After obtaining the channel-mask vector $\mathbf{p}$, we re-weight the channels of the ad-hoc microphone array by the vector, and apply the selected channels to the next module. For the 1-best channel selection, we pick the signal $Y(t, f)$ of the best channel, and apply a nonlinear single-channel speaker extraction algorithm to the channel. For the other channel selection algorithms, we select N-channel signals, denoted as $\mathbf{Y}(\mathbf{t}, \mathbf{f}) = [Y_1(t, f), Y_2(t, f), \dots, Y_N(t, f)]$. After obtaining the corresponding estimated masks $\mathbf{M}(\mathbf{t}, \mathbf{f}) = [M_1(t, f), M_2(t, f), \dots, M_N(t, f)]$, we apply the estimated masks to a deep learning based beamforming algorithm.

2.2. Single-channel speaker extraction

The system diagram of the single-channel speaker extraction is shown in Fig. 4. Given some auxiliary information of the target speech, it estimates a ratio mask for the target speech. Then, the estimated magnitude spectrum of the target speaker is obtained by applying the ratio mask to the mixed speech as follows:

$$|\hat{X}_a(t, f)| = \hat{M}_a(t, f)|Y(t, f)| \tag{13}$$

where $|\hat{X}_a(t, f)|$ is the estimated magnitude of the target speech, $\hat{M}_a(t, f)$ is an estimated phase sensitive mask (PSM) (Erdogan et al., 2015) of the target speech.

The single-channel speaker extraction encodes the auxiliary information of the target speaker into an embedding in the same way as that in the channel-weight estimation network. The network uses the magnitude and temporal spectrum approximation loss (Xu et al., 2019). This loss not only integrates the merit of PSM and signal approximation (Huang et al., 2014), but also captures the dynamic information, i.e. the increment and acceleration, of the target speech:

$$\begin{aligned} J_2 = \frac{1}{T}\sum \Big(& \big\| |\hat{X}_a(t, f)| \\ & - |X(t, f)| \cos(\theta_y(t, f) - \theta_x(t, f)) \big\|_F^2 \\ & + w_d \big\| f_d |\hat{X}_a(t, f)| \\ & - f_d(|X(t, f)| \cos(\theta_y(t, f) - \theta_x(t, f))) \big\|_F^2 \\ & + w_c \big\| f_c |\hat{X}_a(t, f)| \\ & - f_c(|X(t, f)| \cos(\theta_y(t, f) - \theta_x(t, f))) \big\|_F^2 \Big) \end{aligned} \tag{14}$$

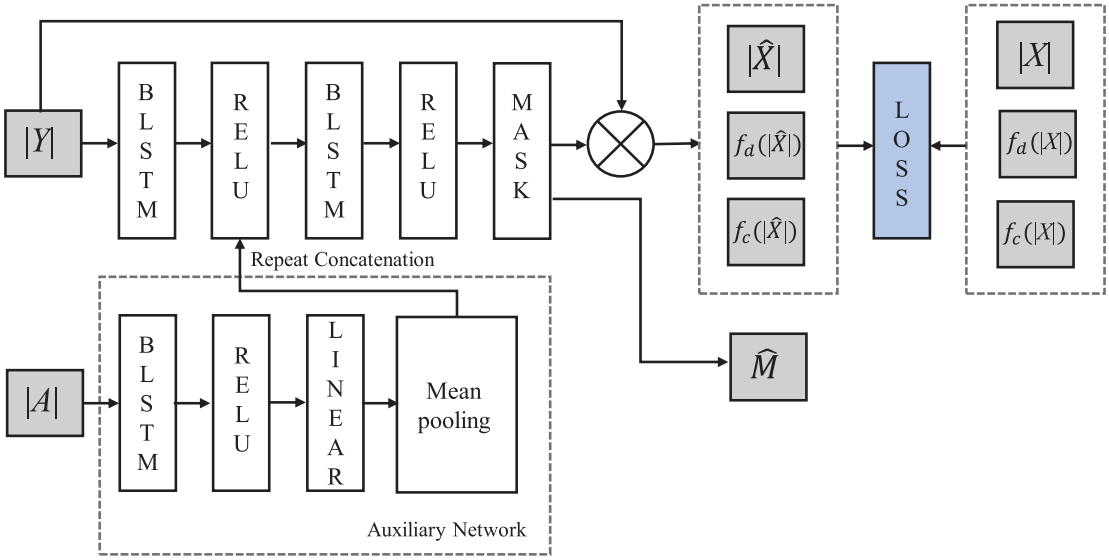

Fig. 4. Diagram of the single-channel speaker extraction network.

where $|X(t,f)|$ is the ground-truth magnitude of the target speaker. $\theta_y(t,f)$ and $\theta_x(t,f)$ are the ground-truth phases of the spectrum of the mixed speech and target direct speech, w_d and w_c are the weights, and $f_d(\cdot)$ and $f_c(\cdot)$ are two functions for calculating the increment and acceleration respectively (Furui, 1986).

2.3. Beamforming algorithm

The deep learning based MVDR (Heymann et al., 2016) is used to integrate the selected channels for the target speech. It contains two components—deep learning based single channel speaker extraction and MVDR. The deep learning based single-channel speaker extraction in Section 2.2 is used to generate an estimation for each channel, and uses the estimations to learn linear filters w_a for MVDR. MVDR, which is a linear beamforming algorithm, suffers less nonlinear distortions than the single channel speaker extraction algorithm. It uses the linear filters w_a to produce the target speech by:

$$\hat{X}_a(t,f) = \mathbf{w}_\mathbf{a}^H(f)\mathbf{Y}(t,f) \tag{15}$$

where the symbol ^{H} is the conjugate transpose operator, and $\hat{X}_a(t,f)$ is the estimated target speech.

The filter is derived by:

$$\mathbf{w}_\mathbf{a}(f) = \frac{\hat{\mathbf{\Phi}}_{\mathbf{ii\_all}}^{-1}(f)\hat{\mathbf{c}}_\mathbf{a}(f)}{\hat{\mathbf{c}}_\mathbf{a}^H(f)\hat{\mathbf{\Phi}}_{\mathbf{ii\_all}}^{-1}(f)\hat{\mathbf{c}}_\mathbf{a}(f)} \tag{16}$$

where $\hat{\mathbf{c}}_\mathbf{a}(f)$ is the first principal component of the spatial covariance matrix of the target speaker $\hat{\mathbf{\Phi}}_{\mathbf{aa}}(f)$ which is a N×N dimensional matrix:

$$\hat{\mathbf{c}}_\mathbf{a}(f) = \text{principal}(\hat{\mathbf{\Phi}}_{\mathbf{aa}}(f)) \tag{17}$$

$$\hat{\mathbf{\Phi}}_{\mathbf{aa}}(f) = \frac{1}{\sum_t \eta_a(t,f)} \sum_t \eta_a(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^H \tag{18}$$

where $\eta_a(t,f)$ is derived by (Zhang, 2018):

$$\eta_a(t,f) = \prod_{n=1}^{N} \hat{M}_{a,n}(t,f) \tag{19}$$

where $\hat{M}_{a,n}(t,f)$ is the estimated mask of the target speaker from the nth selected channel, and N is the number of the selected channels. $\hat{\mathbf{\Phi}}_{\mathbf{ii\_all}}(f)$ is the covariance matrix of the overall interference (Yin et al., 2018; Taherian et al., 2020), which is derived by :

$$\hat{\mathbf{\Phi}}_{\mathbf{ii\_all}}(f) = \mathbf{\Phi}_{\mathbf{yy}}(f) - \hat{\mathbf{\Phi}}_{\mathbf{aa}}(f) \tag{20}$$

where $\mathbf{\Phi}_{\mathbf{yy}}(f)$ is the covariance matrix of the noisy speech, calculated by:

$$\mathbf{\Phi}_{\mathbf{yy}}(f) = \sum_t \mathbf{Y}(t,f)\mathbf{Y}(t,f)^H \tag{21}$$

Finally, we can get the estimated target speech $\hat{x}_a(t)$ by inverse-STFT:

$$\hat{x}_a(t) = \text{iSTFT}(\hat{X}_a(t,f)) \tag{22}$$

3. Experiments

In this section, we present the datasets, experimental settings, and results in Sections 3.1, 3.2, and 3.3, respectively.

3.1. Datasets

We focused on the two-speaker and three-speaker speech separation problem, in which one speaker was regarded as a target speaker. For each mixture, we randomly generated 16 microphones in a randomly generated room that is 5 to 15 m long, 5 to 25 m wide, and 1 to 2.5 m high. Microphones and speech sources were placed randomly in the room. It is also necessary to mention that for each mixed speech, the locations of the microphones and the size of the room were randomly generated. Each speaker is at least 0.2 m away from the walls, and 0.3 m from the microphones. We created room impulse responses (RIRs) using the method in Allen and Berkley (1979). We convolved the clean speech signals with the RIRs and added the reverberant signals from both sources to produce the mixes speech. The reverberant condition T60 was generated from a Gaussian distribution randomly with a mean value of 0.25 s and a variance of 0.1 second^2. Additionally, we constrain T_{60} in the range of $[0.1, 0.4]$ s. To evaluate the effect of the number of microphones on performance, we repeated the above process except that the number of microphones was set to 8.

We first generated WSJ0-adhoc-2mix corpus[1] from the WSJ0 corpus (Garofolo et al., 1993) at a sampling rate of 8 kHz in the aforementioned environment for 16 and 8 microphones respectively. The mixed speech in the WSJ0-adhoc-2mix corpus contains the same content as that in the WSJ0-2mix corpus (Xu et al., 2019). Specifically, the original corpus '*si_tr_s*', composed of 50 male and 51 female speakers, was randomly mixed to generate the training set (20 000 utterances) and

[1] https://github.com/aaaceo890/distributed-multi-channel-data-generate.

Table 1
Parameter setting of the channel-weight estimation network.

Main network	
BLSTM	2 BLSTMs of 512 nodes in each direction
The feed-forward	A nonlinear layer with 512 ReLUs
hidden layers	A nonlinear layer with 256 ReLUs
Pooling layer	Mean pooling
Output layer	1-dimensional sigmoid function
Auxiliary network	
BLSTM	A BLSTM of 256 nodes in each direction
The feed-forward	A nonlinear layer with 256 ReLUs
hidden layers	A linear layer of 30 nodes
Pooling layer	Mean pooling
Initial learning rate	0.0005
Minibatch size	32
Training epochs	30–60
Optimizer	Adam algorithm

validation set (5000 utterances) of WSJ0-2mix at various SNR uniformly chosen between 0 dB and 5 dB. Similarly, the original datasets '*si_dt_*05' and '*si_et_*05', composed of 10 male and 8 female speakers, were randomly mixed to generate the test set (3000 utterances). Then, we generated WSJ0-adhoc-3mix corpus for 16 microphones from WSJ0-3mix corpus in the same way. Because the speakers of the test set were different from those of the training set and validation set, our experimental scenario was regarded as open condition evaluation. To study the effect of different gender combinations on performance, we grouped the test set into 'Female+Female' (F+F), 'Female+Male' (F+M), and 'Male+Male' (M+M) for evaluation.

Then, we evaluated our algorithm on the Libri-adhoc40 corpus (Guan et al., 2021) which was collected by adhoc microphone arrays of 40 strongly synchronized distributed nodes in a real office environment. This dataset has strong reverberation with little additive noise. In order to verify the generalization performance of DABse, we used Libri-adhoc40-simu corpus (Guan et al., 2021) for the network training and used part of the training set of the semi-real Libri-adhoc40 corpus for testing, which amounts to 20 000 utterances from 96 speakers as the training set, 4000 utterances from 91 speakers as the validation set, and 2000 utterances from 65 speakers as the test set. The speakers in the training and validation sets were chosen from the Libri-adhoc40-simu corpus, and the speakers in the test set from the Libri-adhoc40 corpus.

In all experiment scenarios, for the mixed speech, we set the first speaker as the target speaker, and the other speakers as the interference speakers. The utterance of the target speaker was regarded as the reference speech. At the same time, we randomly selected a different utterance of the same target speaker from the corresponding corpus as the auxiliary information of the target speaker.

3.2. Experimental settings

3.2.1. DABse

We set the frame length to 32 ms and the frame shift to 16 ms. A 129-dimensional spectrum was extracted from each frame by STFT with a pre-emphasis of a normalized square root hamming window.

The parameters of the channel-weight estimation network are described in Table 1. The parameters of the single-channel speaker extraction network were similar with those of the channel-weight estimation network, except that the output layer contains 129 units for predicting the ratio mask.

We denote DABse with a specific channel-selection algorithm as 'DABse+channel-selection', which results in the following four methods.

- **DABse+1-best.**
- **DABse+fixed-N-best.** We set N at $\sqrt{M}$.
- **DABse+auto-N-best.** We set γ at 0.5.
- **DABse+soft-N-best.** We set γ at 0.5.

Table 2
Comparison results of DABse with 16 microphones per ad-hoc microphone array and three baselines, where the average T60 of all environments is 0.25 s.

Comparison method	Gender	SDR (dB)	PESQ	STOI
Single-channel	M+M	2.84	1.91	0.71
	F+F	3.65	2.00	0.74
	F+M	5.45	2.18	0.77
Linear array	M+M	4.65	1.96	0.75
	F+F	4.02	1.91	0.74
	F+M	5.61	2.13	0.79
All-channels	M+M	5.47	2.20	0.79
	F+F	4.17	1.92	0.78
	F+M	5.68	2.14	0.80
DABse+1-best	M+M	4.84	1.98	0.78
	F+F	6.58	2.21	0.79
	F+M	7.72	2.26	0.81
DABse+fixed-N-best	M+M	3.51	1.86	0.77
	F+F	5.32	2.08	0.79
	F+M	7.96	2.30	0.83
DABse+auto-N-best	M+M	5.56	2.10	0.79
	F+F	6.66	2.23	0.81
	F+M	**8.47**	**2.34**	**0.84**
DABse+soft-N-best	M+M	5.17	2.06	0.76
	F+F	6.30	2.15	0.80
	F+M	8.11	2.30	0.83

Note that 'DABse+1-best' is a nonlinear speech separation method while the others are linear methods.

3.2.2. Baselines

We compared DABse with two extreme DABse variants:

- **Selecting one-random channel (single-channel)** We randomly select a channel from the W channels, and then conduct single-channel speaker extraction. This extreme case does not refer to channel selection, and is therefore irrelevant to the number of channels.
- **Selecting all channels (all-channels)** We use all channels for the multi-channel speaker extraction, i.e.:

$$p_j = 1, \forall j = 1, \ldots, W \tag{23}$$

- **Deep learning based MVDR with linear microphone array (linear array)** The method is similar with the baseline 'all-channels' except that the ad-hoc microphone array is replaced by a traditional microphone array. We choose a linear array for the far-field multi-channel speaker extraction. We set each linear array with 16 microphones and the distance between the microphones to 10 cm. To evaluate its performance for two-speaker speech extraction, we set its evaluation environment the same as WSJ0-2mix-adhoc.

Note that 'single-channel' is a nonlinear speech separation method while 'linear array' and 'all-channels' are both linear methods.

3.2.3. Evaluation metrics

The performance evaluation metrics include signal to distortion ratio (SDR) (Vincent et al., 2006), perceptual evaluation of speech quality (PESQ) (Rix et al., 2001), and short-time objective intelligibility (STOI) (Taal et al., 2011). SDR is a metric similar to SNR for evaluating the quality of enhancement. PESQ is a test methodology for automated assessment of the speech quality as experienced by a listener of a telephony system. STOI evaluates the objective speech intelligibility of time-domain signals. The higher the value of an evaluation metric is, the better the performance is.

Table 3
Comparison results of DABse with 8 microphones per ad-hoc microphone array and three baselines on the gender pair of F+M.

Comparison method	Beamforming method	SDR (dB)	PESQ	STOI
Single-channel	–	5.45	2.18	0.77
Linear array	MVDR	3.49	1.91	0.73
All-channels	MVDR	3.52	1.94	0.74
DABse+1-best	–	5.69	2.05	0.76
DABse+fixed-N-best	MVDR	5.79	2.28	0.77
DABse+auto-N-best	MVDR	5.84	2.17	0.79
DABse+soft-N-best	MVDR	**6.49**	**2.22**	**0.80**

Table 4
Results on different number of speakers in the mixed speech. The comparison methods are the 'DABse+auto-N-best' with 16 microphones per ad-hoc microphone array and the single-channel method on the gender pair of F+M. The average T60 is 0.25 s. The first two columns are the number of speakers.

Training	Test	Comparison method	SDR (dB)	PESQ	STOI
2	2	Single-channel	5.45	2.18	0.77
		DABse	**8.47**	**2.34**	**0.84**
2	3	Single-channel	1.32	1.83	0.69
		DABse	4.11	2.04	0.75
3	2	Single-channel	3.68	2.01	0.72
		DABse	5.79	2.20	0.79
3	3	Single-channel	3.13	1.98	0.73
		DABse	5.43	2.15	0.77
2 & 3	2	Single-channel	5.02	2.09	0.76
		DABse	8.13	2.31	0.82
2 & 3	3	Single-channel	4.17	2.08	0.76
		DABse	6.65	2.21	0.80

Table 5
Effect of hyperparameter N of 'DABse+fixed-N-best' on the gender pair F+M.

Number of selected microphones (N)	SDR (dB)	PESQ	STOI
N = 2	4.42	2.11	0.78
N = 4	**7.96**	**2.30**	0.83
N = 6	7.87	**2.30**	**0.85**
N = 8	7.53	2.27	0.83
N = 10	6.97	2.24	0.81
N = 12	6.23	2.22	0.81
N = 14	6.76	2.26	0.82
N = 16	5.86	2.14	0.80

3.3. Results on WSJ0-adhoc

Table 2 lists the comparison results of the DABse variants with the three baselines. From the table, we see that the last three channel selection algorithms outperform the other comparison methods in most cases. Among the three algorithms, 'DABse+auto-N-best' outperforms all comparison methods, followed by 'DABse+soft-N-best'. Although 'DABse+fixed-N-best' produces similar PESQ and STOI results with 'DABse+soft-N-best', its SDR score is poorer than the latter.

From the results listed in Tables 2 and 3, we can see that the performance of DABse with channel selection is better than that without channel selection, so we can draw a conclusion that the SNR^u estimators are closely related to the performance of DABse. To be more specific, from the Table 2, we find that even the simplest channel selection algorithm 'DABse+1-best' can produce better experimental results than all the baselines in most cases, which proves the necessity, correctness and effectiveness of the channel selection for DABse. For example, for the combination of F+M, 'DABse+1-best' achieves 2.27 dB higher than 'single-channel' in terms of SDR.

3.3.1. Effect of the number of microphones in the ad-hoc microphone array

In order to explore the influence of different number of microphones of ad-hoc microphone arrays on DABse, we conducted experiments with ad-hoc microphone arrays of 8 microphones on the F+M combination when $T60_{mean}$ was set at 0.25 s. And it is necessary to be mentioned that we set N at 3 ($\sqrt{8} \approx 3$) of 'DABse+fixed-N-best'.

The results are listed in the Table 3. From the table, we can find that it is the 'DABse+soft-N-best' rather than 'DABse+auto-N-best' that performs the best among all comparison methods. Besides, it is obvious that the ad-hoc microphone array with 16 microphones outperforms that with 8 microphones.

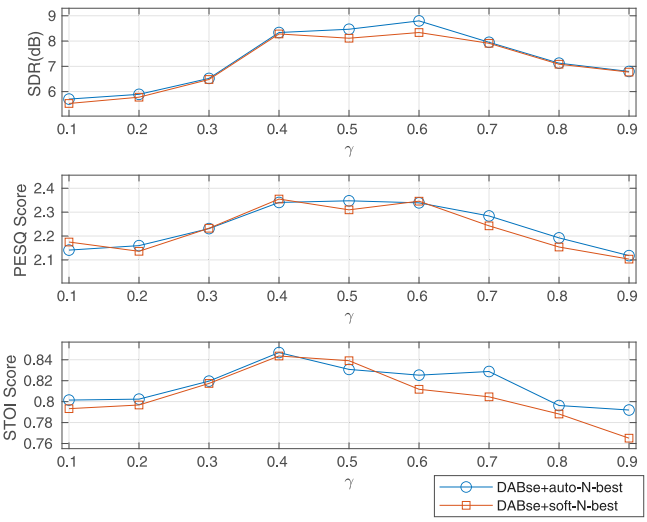

Fig. 5. Effect of hyperparameter γ of 'DABse+auto-N-best' and 'DABse+soft-N-best' on the gender pair F+M.

3.3.2. Effect of DABse on different gender combinations

Table 2 lists the effect of DABse on different gender combinations. From the table, we see that DABse and single-channel speaker extraction always achieve better performance on the gender pair of F+M. For the same gender combinations, it seems that they perform better on F+F than on M+M in most cases. In addition, we find that 'DABse+1-best' does not outperform 'all-channels' on M+M. The phenomena indicate that M+M might be more difficult for single-channel speaker extraction than the other gender combinations, and the effectiveness of DABse is strongly affected by the single-channel speaker extraction algorithm.

3.3.3. Effect of DABse on different number of speakers

In order to explore the influence of different number of speakers on DABse, we conducted an experiment on two-speaker and three-speaker separation problems. We trained the DABse system with three conditions, which are the two-speaker mixtures, three-speaker mixtures, and the mixtures with both two-speakers and three-speakers respectively. Then, we compared 'DABse+auto-N-best' with the single-channel method on the two-speaker and three-speaker mixed test data when $T60_{mean}$ was set to 0.25 s. Table 4 lists the results on the gender pair of F+M. From the table, we see that although the performance of both the comparison methods drop significantly when the number of speakers increases, DABse still outperforms 'single-channel' in all test conditions.

3.3.4. Effect of hyperparameters of DABse on performance

To study how the selected number of channels N affects the performance of 'DABse+fixed-N-best', we conducted an experiment on the F+M gender pair with N selected from $\{2, 4, 6, 8, 10, 12, 14, 16\}$ respectively. From the experimental results in Table 5, we see that the performance first gets improved and then decreased along with the increase of N, with the top SDR performance appearing at $N = 4$ and top STOI performance appearing at $N = 6$, which demonstrates the correctness of our experimental setting, that is, setting N at $\sqrt{M}$.

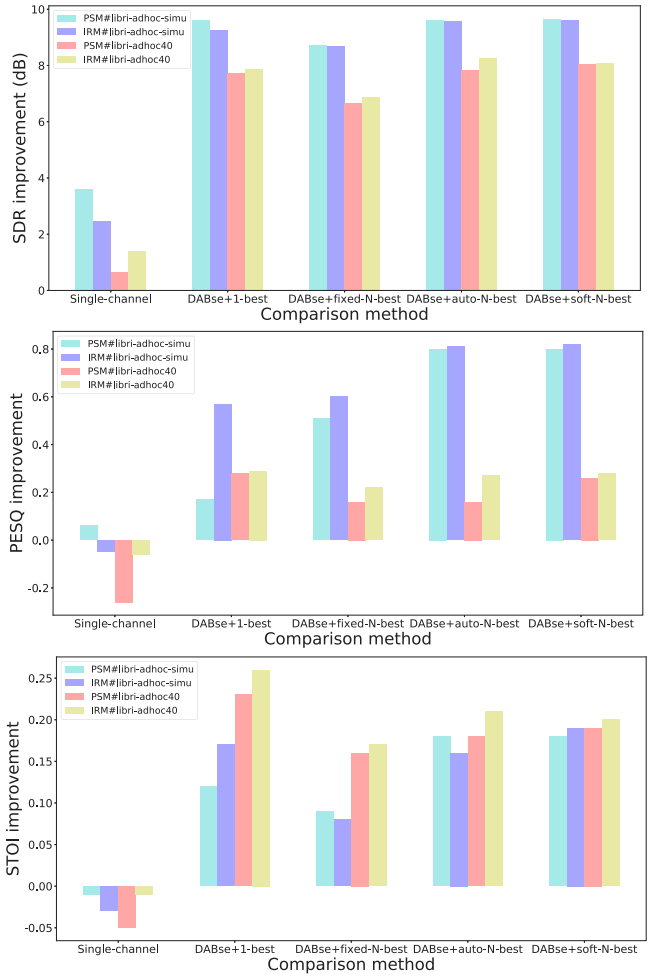

Fig. 6. Comparison results between PSM and IRM for the DABse variants and the 'single-channel' baseline on the Libri-adhoc-simu and Libri-adhoc40 corpora.

To explore how the hyperparameter γ affects the performance of 'DABse+auto-N-best' and 'DABse+soft-N-best', we conducted an experiment on the F+M gender pair with γ selected from 0.1 to 0.9. The experimental results are shown in Fig. 5. From the figure, we see that, when γ was set at 0.4 to 0.6, both 'DABse+auto-N-best' and 'DABse+soft-N-best' achieve top performance. Therefore, setting the default value of γ to 0.5 is reasonable.

3.4. Results on Libri-adhoc40

In order to study whether DABse is valid in real scenarios, we conducted experiments on the real dataset with the model trained on the simulation dataset. In this experiment, the parameter setting of DABse is the same as the previous experiments.

Fig. 6 shows the comparison results of the DABse variants with the baseline of 'single-channel'. From the bar charts, we see that DABse with the channel selection module outperforms the baseline in all the cases. Even the simplest channel selection method 'DABse+1-best' outperforms the 'single-channel' baseline that does not adopt channel selection. For example, 'DABse+1-best' with the IRM mask achieves nearly 6 dB higher than 'single-channel' on the Libri-adhoc40 corpus in terms of SDR.

From Fig. 6, we also find that although the performance of the DABse variants on the semi-real Libri-adhoc40 corpus is inferior to that on the Libri-adhoc-simu corpus, it is still significantly better than the 'single-channel' baseline on both datasets. Among four DABse variants, 'DABse+auto-N-best' outperforms all comparison methods on the Libri-adhoc40 corpus, followed by 'DABse+soft-N-best'. At the same time, we find that the results on the Libri-adhoc-simu and Libri-adhoc40 corpora basically have the same trend. Therefore, we can conclude that the advantage of the DABse model trained in the simulated environment can be generalized from the simulated test environment to the semi-real environment.

3.4.1. Effects of different masks on performance

In addition, we further compared PSM with IRM, which is defined in Eq. (24), to study how different masks affect the performance of DABse in both simulation and real scenarios. From Fig. 6, we see that the DABse with the IRM mask outperforms that with the PSM mask on the Libri-adhoc40 corpus. The same phenomenon can be found on the Libri-adhoc-simu corpus as well. Although the 'DABse+fix-N-best' and 'DABse+auto-N-best' with the PSM masks produce better STOI results than those with the IRM masks, their PESQ scores are much poorer than the latter.

$$IRM(t,f) = \frac{|X_a(t,f)|}{|X_a(t,f)| + |X_i(t,f) + H(t,f)|} \tag{24}$$

3.4.2. Effect of channel selection on performance

To demonstrate the effectiveness of the channel selection module, we visualize an example produced from different channel selection algorithms in three scenarios of the Libri-adhoc40 corpus. From Fig. 7, we see that, for 'DABse+1-best', the nearest microphone to the target speaker along the speaking direction can be accurately selected. The statistical accuracy rate of 'DABse+1-best' over the entire test set can reach as high as 80.5%. For 'DABse+fix-N-best' and 'DABse+auto-N-best', we see that a number of microphones around the target speaker are grouped together in an ad-hoc way.

In addition, to study the accuracy of the SNR^u estimation module, we calculated the estimation error between the ground-truth SNR and the estimated SNR^u on the Libri-adhoc-simu corpus. From Fig. 8, we see that the percentage of the estimation error that is smaller than 0.15 reaches 91.86%. The statistical mean and variance of the estimation error are 0.062 and 0.004 respectively. We conjecture reasonably that, if the SNR^u estimation network is further improved, the channel selection algorithms may be more effective.

3.4.3. Effect of different modules on performance

In order to study how different modules of the proposed system affects the performance, we conducted ablation experiments for 'DABse+1-best' on the Libri-adhoc-simu corpus. Specifically, for each module of the 'DABse+1-best', we used the ground-truth value instead of the estimation output of the module, which yields four experimental scenarios: 'Oracle SNR', 'Oracle channel selection', 'Oracle mask', and 'Oracle SNR+Oracle mask'.

The results are shown in Fig. 9. From the figure, it is obvious that the 'Oracle SNR+Oracle mask' variant achieves the best performance, which is drawn with a dashed line in the diagram. We also find that the results of 'Oracle SNR' are the same as those of 'Oracle channel selection', which indicates that SNR^u, which is used as the criterion for our channel selection, has strong correlation to the physical distance between the speaker and the microphones. The performance gap between 'Oracle mask' and 'DABse+1-best' is obviously larger than that between 'Oracle SNR' and 'DABse+1-best', which demonstrates that using an additional SNR^u estimation network as an independent module is relatively better than using a single mask estimation network in our experimental environment. Besides, we can see that 'Oracle mask' is very close to the best result, so we can conclude that the accuracy of mask estimation network is an important factor affecting the performance of the whole DABse system.

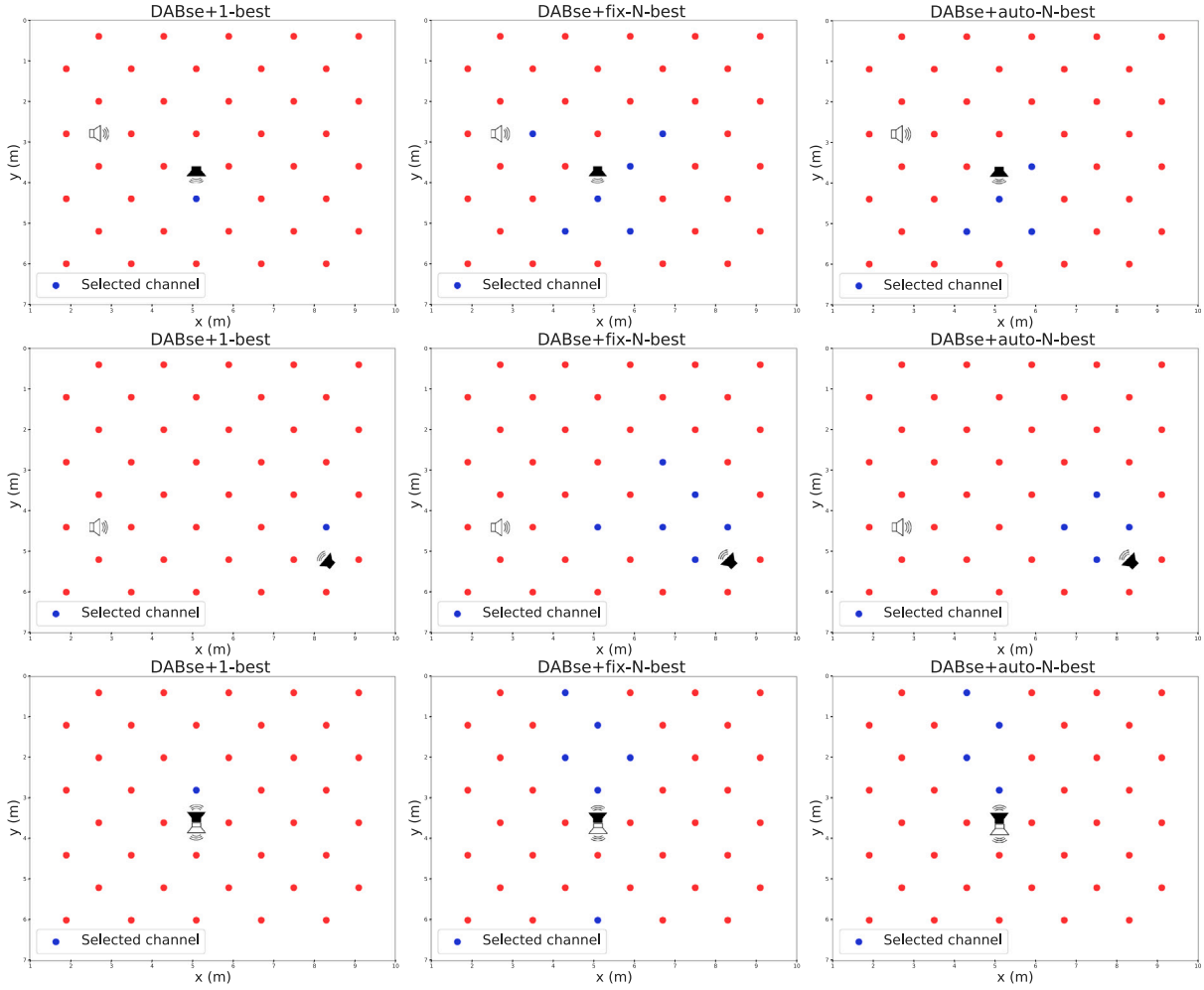

Fig. 7. Visualization results of the three channel selection algorithms in three different test scenarios of the Libri-adhoc40 corpus. The dots represent microphones, the black trumpet represents the target speaker, and the white trumpet represents the inference speaker.

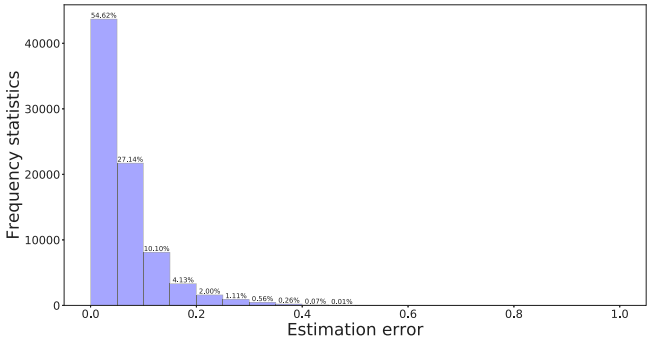

Fig. 8. Histogram of the absolute estimation error between the ground-truth SNR and the estimated SNR^u produced by the estimation network. The horizontal axis represents the value of the absolute error between the ground-truth value and the estimated SNR^u. The vertical axis represents the frequency statistics. The percentage score above the histogram represents the percentage of statistics with different errors on the entire test set.

4. Conclusions

In this paper, we have proposed deep ad-hoc beamforming based on speaker extraction, which is the first work of the target-dependent speech separation based on ad-hoc microphone arrays and deep learning. DABse uses the channel-weight estimation network based on speaker extraction to estimate the SNR^u of the target speaker, and then takes the SNR^u as the channel weight for the selection of high-quality channels, and finally takes the selected channels for the deep learning based MVDR. The deep learning based MVDR first takes the single-channel target-dependent speaker extraction network to estimate the clean spectrum of the target speech at each selected channel, and then uses the estimated spectrum to derive an MVDR filter for the final speech separation. Because the two deep models in DABse are trained in a single-channel fashion, it is able to handle any number of microphones in the test stage. Because MVDR is a linear filter, DABse does not suffer from nonlinear distortions. We have conducted extensive experiments in both simulation and real scenarios where the speech sources are located randomly in large rooms. We compared DABse with the baselines of ‘single-channel’, ‘all-channels’ and ‘linear array’. Experimental results demonstrate that the proposed DABse outperforms the baselines significantly, which illustrates the effectiveness of DABse in the adverse environments.

Our work is just the first attempt on the speaker extraction problem with ad-hoc microphone arrays. There is still some work to figure out. First, integrating spatial information into the training process of the network is able to improve the performance of DABse. Another interesting direction is to extract the target speech while the speaker is walking, which is a common situation in our real life. In addition, the speaker-independent speech separation with ad-hoc microphone arrays is an important direction.

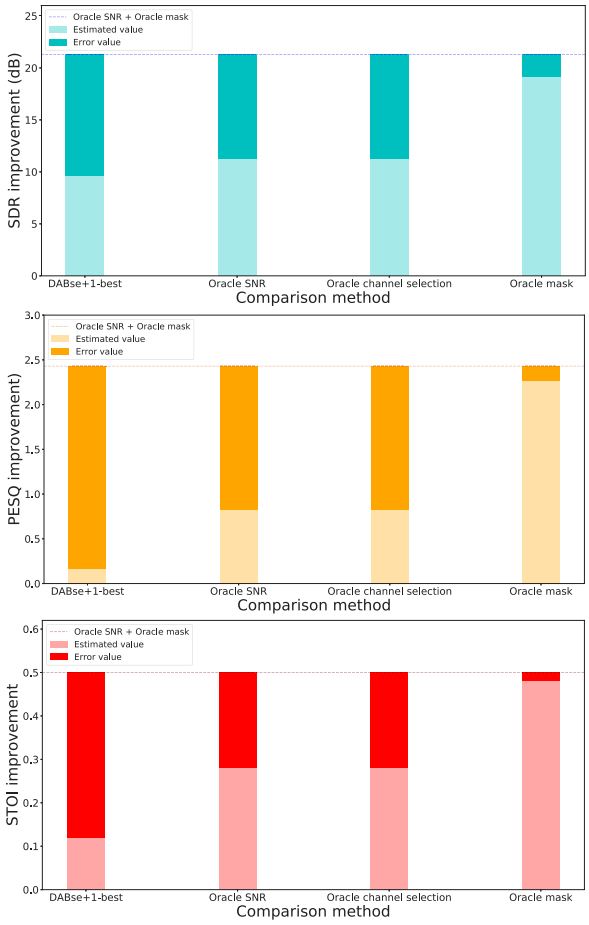

Fig. 9. Ablation experiments for 'DABse+1-best' on the Libri-adhoc-simu corpora. The term 'Oracle' means that we used ground-truth value instead of the estimated value of the corresponding module of the 'DABse+1-best' algorithm. For example, 'Oracle channel selection' means that the oracle 1-best channel is physically the closest channel to the target speaker.

CRediT authorship contribution statement

Ziye Yang: Model development, Experimental testing, Writing paper. **Shanzheng Guan:** Model development, Experimental testing, Writing paper. **Xiao-Lei Zhang:** Model development, Writing paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Science Foundation of China (NSFC) under Grant 62176211, and in part by the Project of the Science, Technology, and Innovation Commission of Shenzhen Municipality, China under grant No. JCYJ20210324143006016 and 202108023000241.

The authors read and approved the final manuscript.

References

Allen, J.B., Berkley, D.A., 1979. Image method for efficiently simulating small-room acoustics. J. Acoust. Soc. Am. 65 (4), 943–950.

Araki, S., Hayashi, T., Delcroix, M., Fujimoto, M., Takeda, K., Nakatani, T., 2015. Exploring multi-channel features for denoising-autoencoder-based speech enhancement. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 116–120.

Bregman, A.S., 1994. Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press.

Chen, J., Zhang, X.-L., 2021. Scaling sparsemax based channel selection for speech recognition with ad-hoc microphone arrays. arXiv preprint arXiv:2103.15305.

Delcroix, M., Zmolikova, K., Ochiai, T., Kinoshita, K., Araki, S., Nakatani, T., 2019. Compact network for speakerbeam target speaker extraction. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6965–6969.

Delfarah, M., Liu, Y., Wang, D., 2020. A two-stage deep learning algorithm for talker-independent speaker separation in reverberant conditions. J. Acoust. Soc. Am. 148 (3), 1157–1168.

Delfarah, M., Wang, D., 2019. Deep learning for talker-dependent reverberant speaker separation: An empirical study. IEEE/ACM Trans. Audio Speech Lang. Process. 27 (11), 1839–1848.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33 (2), 443–445.

Erdogan, H., Hershey, J.R., Watanabe, S., Le Roux, J., 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 708–712.

Furui, S., 1986. Speaker-independent isolated word recognition using dynamic features of speech spectrum. IEEE Trans. Acoust. Speech Signal Process. 34 (1), 52–59. http://dx.doi.org/10.1109/TASSP.1986.1164788.

Garofolo, J., Graff, D., Paul, D., Pallett, D., 1993. Csr-i (wsj0) Complete ldc93s6a, Web Download, Vol. 83. Linguistic Data Consortium, Philadelphia.

Gu, R., Zhang, S.-X., Xu, Y., Chen, L., Zou, Y., Yu, D., 2020. Multi-modal multi-channel target speech separation. IEEE J. Sel. Top. Sign. Proces. 14 (3), 530–541.

Guan, S., Liu, S., Chen, J., Zhu, W., Li, S., Tan, X., Yang, Z., Xu, M., Chen, Y., Wang, J., et al., 2021. Libri-adhoc40: A dataset collected from synchronized ad-hoc microphone arrays. arXiv preprint arXiv:2103.15118.

Hershey, J.R., Chen, Z., Le Roux, J., Watanabe, S., 2016. Deep clustering: Discriminative embeddings for segmentation and separation. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 31–35.

Heymann, J., Drude, L., Haeb-Umbach, R., 2016. Neural network based spectral mask estimation for acoustic beamforming. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 196–200.

Higuchi, T., Ito, N., Araki, S., Yoshioka, T., Delcroix, M., Nakatani, T., 2017. Online MVDR beamformer based on complex Gaussian mixture model with spatial prior for noise robust ASR. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (4), 780–793. http://dx.doi.org/10.1109/TASLP.2017.2665341.

Higuchi, T., Ito, N., Yoshioka, T., Nakatani, T., 2016. Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5210–5214. http://dx.doi.org/10.1109/ICASSP.2016.7472671.

Huang, P.-S., Kim, M., Hasegawa-Johnson, M., Smaragdis, P., 2014. Deep learning for monaural speech separation. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1562–1566.

Isik, Y., Roux, J.L., Chen, Z., Watanabe, S., Hershey, J.R., 2016. Single-channel multi-speaker separation using deep clustering. arXiv preprint arXiv:1607.02173.

Itakura, K., Bando, Y., Nakamura, E., Itoyama, K., Yoshii, K., Kawahara, T., 2018. Bayesian multichannel audio source separation based on integrated source and spatial models. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (4), 831–846.

Jayaprakasam, S., Rahim, S.K.A., Leow, C.Y., 2017. Distributed and collaborative beamforming in wireless sensor networks: Classifications, trends, and research directions. IEEE Commun. Surv. Tutor. 19 (4), 2092–2116.

Jiang, Y., Wang, D., Liu, R., Feng, Z., 2014. Binaural classification for reverberant speech segregation using deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 2112–2121.

Kitamura, K., Bando, Y., Itoyama, K., Yoshii, K., 2016. Student's t multichannel nonnegative matrix factorization for blind source separation. In: 2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE, pp. 1–5.

Kitamura, D., Ono, N., Sawada, H., Kameoka, H., Saruwatari, H., 2018. Determined blind source separation with independent low-rank matrix analysis. In: Audio Source Separation. Springer, pp. 125–155.

Kolbæk, M., Yu, D., Tan, Z.-H., Jensen, J., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (10), 1901–1913.

Koutrouvelis, A.I., Sherson, T.W., Heusdens, R., Hendriks, R.C., 2018. A low-cost robust distributed linearly constrained beamformer for wireless acoustic sensor networks with arbitrary topology. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (8), 1434–1448.

Liang, C., Chen, J., Guan, S., Zhang, X.-L., 2021. Attention-based multi-channel speaker verification with ad-hoc microphone arrays. arXiv preprint arXiv:2107.00178.

Luo, Y., Chen, Z., Mesgarani, N., Yoshioka, T., 2020. End-to-end microphone permutation and number invariant multi-channel speech separation. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6394–6398.

Luo, Y., Han, C., Mesgarani, N., Ceolini, E., Liu, S.-C., 2019. FaSNet: Low-latency adaptive beamforming for multi-microphone audio processing. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, pp. 260–267.

Nakatani, T., Ito, N., Higuchi, T., Araki, S., Kinoshita, K., 2017. Integrating DNN-based and spatial clustering-based mask estimation for robust MVDR beamforming. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 286–290.

Ochiai, T., Delcroix, M., Kinoshita, K., Ogawa, A., Nakatani, T., 2019. A unified framework for neural speech separation and extraction. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6975–6979.

Otsuka, T., Ishiguro, K., Sawada, H., Okuno, H.G., 2013. Bayesian nonparametrics for microphone array processing. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (2), 493–504.

Ozerov, A., Févotte, C., 2009. Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation. IEEE Trans. Audio Speech Lang. Process. 18 (3), 550–563.

Pertilä, P., Nikunen, J., 2015. Distant speech separation using predicted time–frequency masks from spatial features. Speech Commun. 68, 97–106.

Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P., 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221), Vol. 2. IEEE, pp. 749–752.

Rouat, J., 2008. Computational auditory scene analysis: Principles, algorithms, and applications (wang, d. and brown, gj, eds.; 2006)[book review]. IEEE Trans. Neural Netw. 19 (1), 199.

Sawada, H., Kameoka, H., Araki, S., Ueda, N., 2013. Multichannel extensions of non-negative matrix factorization with complex-valued data. IEEE Trans. Audio Speech Lang. Process. 21 (5), 971–982.

Schmidt, M.N., Olsson, R.K., 2006. Single-channel speech separation using sparse non-negative matrix factorization. In: Ninth International Conference on Spoken Language Processing.

Stark, M., Wohlmayr, M., Pernkopf, F., 2010. Source–filter-based single-channel speech separation using pitch information. IEEE Trans. Audio Speech Lang. Process. 19 (2), 242–255.

Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J., 2011. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. 19 (7), 2125–2136.

Taherian, H., Wang, Z.-Q., Chang, J., Wang, D., 2020. Robust speaker recognition based on single-channel and multi-channel speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. 28, 1293–1302.

Tavakoli, V.M., Jensen, J.R., Heusdens, R., Benesty, J., Christensen, M.G., 2017. Distributed max-SINR speech enhancement with ad hoc microphone arrays. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 151–155.

Vincent, E., Gribonval, R., Févotte, C., 2006. Performance measurement in blind audio source separation. IEEE Trans. Audio Speech Lang. Process. 14 (4), 1462–1469.

Virtanen, T., 2006. Speech recognition using factorial hidden Markov models for separation in the feature space. In: Ninth International Conference on Spoken Language Processing.

Virtanen, T., 2007. Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria. IEEE Trans. Audio Speech Lang. Process. 15 (3), 1066–1074.

Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (10), 1702–1726.

Wang, D., Chen, Z., Yoshioka, T., 2020. Neural speech separation using spatially distributed microphones. In: Proc. Interspeech 2020. pp. 339–343.

Wang, Z.-Q., Le Roux, J., Hershey, J.R., 2018a. Alternative objective functions for deep clustering. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 686–690.

Wang, Z.-Q., Le Roux, J., Hershey, J.R., 2018b. Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 1–5.

Wang, Q., Muckenhirn, H., Wilson, K., Sridhar, P., Wu, Z., Hershey, J., Saurous, R.A., Weiss, R.J., Jia, Y., Moreno, I.L., 2018c. Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking. arXiv preprint arXiv:1810.04826.

Wang, Y., Narayanan, A., Wang, D., 2014. On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 22 (12), 1849–1858.

Wang, Y., Wang, D., 2012. Boosting classification based speech separation using temporal dynamics. In: Thirteenth Annual Conference of the International Speech Communication Association.

Wang, Y., Wang, D., 2013. Towards scaling up classification-based speech separation. IEEE Trans. Audio Speech Lang. Process. 21 (7), 1381–1390.

Wang, D., Yoshioka, T., Chen, Z., Wang, X., Zhou, T., Meng, Z., 2021. Continuous speech separation with ad hoc microphone arrays. arXiv preprint arXiv:2103.02378.

Williamson, D.S., Wang, D., 2017. Time-frequency masking in the complex domain for speech dereverberation and denoising. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (7), 1492–1501.

Xiao, X., Chen, Z., Yoshioka, T., Erdogan, H., Liu, C., Dimitriadis, D., Droppo, J., Gong, Y., 2019. Single-channel speech extraction using speaker inventory and attention network. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 86–90.

Xu, C., Rao, W., Chng, E.S., Li, H., 2019. Optimization of speaker extraction neural network with magnitude and temporal spectrum approximation loss. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6990–6994.

Xu, C., Rao, W., Xiao, X., Chng, E.S., Li, H., 2018. Single channel speech separation with constrained utterance level permutation invariant training using grid lstm. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 6–10.

Yang, Z., Zhang, X.-L., 2019a. Boosting spatial information for deep learning based multichannel speaker-independent speech separation in reverberant environments. In: 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, pp. 1506–1510.

Yang, Z., Zhang, X.-L., 2019b. Multi-channel speech separation using deep embedding model with multilayer bootstrap networks. arXiv preprint arXiv:1910.10912.

Yin, L., Wang, Z., Xia, R., Li, J., Yan, Y., 2018. Multi-talker speech separation based on permutation invariant training and beamforming. In: INTERSPEECH. pp. 851–855.

Yoshioka, T., Erdogan, H., Chen, Z., Alleva, F., 2018. Multi-microphone neural speech separation for far-field multi-talker speech recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5739–5743.

Yu, D., Kolbæk, M., Tan, Z.-H., Jensen, J., 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 241–245.

Zhang, X.-L., 2018. Deep ad-hoc beamforming. arXiv preprint arXiv:1811.01233.

Zhang, J., Chepuri, S.P., Hendriks, R.C., Heusdens, R., 2017. Microphone subset selection for MVDR beamformer based noise reduction. IEEE/ACM Trans. Audio Speech Lang. Process. 26 (3), 550–563.

Zhang, X.-L., Wang, D., 2016. A deep ensemble learning method for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. 24 (5), 967–977.

Zhang, X., Wang, D., 2017. Deep learning based binaural speech separation in reverberant environments. IEEE/ACM Trans. Audio Speech Lang. Process. 25 (5), 1075–1084.

Žmolíková, K., Delcroix, M., Kinoshita, K., Higuchi, T., Ogawa, A., Nakatani, T., 2017. Learning speaker representation for neural network based multichannel speaker extraction. In: 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, pp. 8–15.

Žmolíková, K., Delcroix, M., Kinoshita, K., Ochiai, T., Nakatani, T., Burget, L., Černockỳ, J., 2019. SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures. IEEE J. Sel. Top. Sign. Proces. 13 (4), 800–814.