

RESEARCH

Open Access



# A multichannel learning-based approach for sound source separation in reverberant environments

You-Siang Chen, Zi-Jie Lin and Mingsian R. Bai\* 

## Abstract

In this paper, a multichannel learning-based network is proposed for sound source separation in reverberant field. The network can be divided into two parts according to the training strategies. In the first stage, time-dilated convolutional blocks are trained to estimate the array weights for beamforming the multichannel microphone signals. Next, the output of the network is processed by a weight-and-sum operation that is reformulated to handle real-valued data in the frequency domain. In the second stage, a U-net model is concatenated to the beamforming network to serve as a non-linear mapping filter for joint separation and dereverberation. The scale invariant mean square error (SI-MSE) that is a frequency-domain modification from the scale invariant signal-to-noise ratio (SI-SNR) is used as the objective function for training. Furthermore, the combined network is also trained with the speech segments filtered by a great variety of room impulse responses. Simulations are conducted for comprehensive multisource scenarios of various subtending angles of sources and reverberation times. The proposed network is compared with several baseline approaches in terms of objective evaluation matrices. The results have demonstrated the excellent performance of the proposed network in dereverberation and separation, as compared to baseline methods.

**Keywords:** Source separation and dereverberation, Multichannel learning-based network, Time-dilated convolution network, U-net, Beamforming

## 1 Introduction

As an important problem in speech enhancement, source separation seeks to separate independent source signals from mixture signals, based on the spatial cue, the temporal-spectral cue, or statistical characteristics of sources. For semi-blind source separation, the free-field wave propagation model is assumed to facilitate a two-stage procedure of source localization and separation by using an array. Beamforming (BF) [1], time difference of arrival (TDOA) [2], and multiple signal classification (MUSIC) [3] are generally used source localization methods. In the separation stage, BF methods such as minimum power distortionless response (MPDR) can be

employed to extract source signals, based on the direction of arrivals estimated in the localization stage [4, 5]. In addition to BF methods, Tikhonov regularization (TIKR) [6] which treats the separation problem as a linear inverse problem can also be used.

On the other hand, blind source separation (BSS) approaches do not rely on a wave propagation model and exploits mainly the time-frequency (T-F) or statistical characteristics of mixture signals. Independent component analysis (ICA) is a well-known BSS algorithm that separates the signals into statistically independent components [7–11]. ICA was further extended to deal with convolutive processes such as acoustic propagation, e.g., triple-N ICA for convolutive mixtures (TRINICON) [12]. An alternative separation algorithm, independent vector analysis (IVA) [13], cleverly circumvents the

\* Correspondence: [msbai@pme.nthu.edu.tw](mailto:msbai@pme.nthu.edu.tw)

Department of Power Mechanical Engineering/Electrical Engineering,  
National Tsing Hua University, Hsinchu, Taiwan

permutation issue in ICA by modeling the statistical interdependency between frequency components.

In this paper, we shall explore the possibility of addressing source separation problems using a learning-based approach, namely, deep neural networks (DNNs). Wang et al. approached source separation by using DNNs in which spectrogram was used as the input features [14]. Promising results were obtained in light of various network structures, including convolutional neural network (CNN) [15], recurrent neural network (RNN) [16], and the deep clustering (DC) method [17], etc. Furthermore, utterance-level permutation invariant training (uPIT) was introduced to resolve the label permutation problem [18]. Recently, fully convolutional time-domain audio separation network (Conv-TasNet) was proposed [19] to separate source signals in the time domain in a computationally efficient way.

Reverberation is detrimental to speech quality, which leads to degradation in speech intelligibility. Multichannel inverse filtering (MINT) was developed to achieve nearly perfect dereverberation [20]. Multi-channel linear prediction (MCLP) [21] based on a time-domain linear prediction model in the T-F domain was reported effective. As a refined version of MCLP, the weighted prediction error (WPE) algorithm was developed in the short-time Fourier transform (STFT) domain via a long-term linear prediction [22]. A multi-channel generalization can be found in [23–25]. DNN approaches have also become promising techniques for dereverberation. Mapping-based approaches [26] attempt to enhance directly the reverberated signals, whereas masking-based approaches [27] attempt to learn a “mask” for anechoic signals. In addition, combined systems of a DNN and the WPE unit were also suggested [28, 29].

Source separation in a reverberant field is particularly challenging. This problem was tackled by cascading a WPE unit and a MPDR beamformer [30, 31]. Several systems have been proposed in light of the joint optimization of the preceding two units [32, 33]. In a very recent work, the weighted power minimization distortionless response (WPD) [34] beamformer was developed by integrating optimally the WPE and MPDR units into a single convolutional beamformer. DNN-based approaches have also been reported recently. An end-to-end learning model was trained to establish a mapping from the reverberant mixture input to anechoic separated speech outputs [35]. Cascade systems [36, 37] were also investigated. Multichannel networks [38, 39] were proposed to exploit the spatial cue of microphone signals. In addition, integrated DNN and conventional beamformers are suggested in recent years [40–42].

Most approaches employ a cascaded structure in which a DNN is trained for the prior information required by the subsequent beamforming algorithm, e.g., a

post-enhancement mask for the beamforming output, masking-based spatial cue estimation, and estimation of the spatial covariance matrix, etc. In practice, DNNs could have some limitations in obtaining the required information for array beamforming where the magnitude of target signals be held fixed in the training stage. Under this circumstance, there is no guarantee that fixed loss functions such as mean-square-error (MSE) or signal-to-noise ratio (SNR) will lead to an optimal estimate [43]. The proposed method seeks to achieve a synergetic integration of arrays and DNN to reformulate and implement the real-valued weight-and-sum operation in a multichannel DNN through a learning-based training for optimal weights. In addition, a new scale-independent MSE loss is derived for optimal estimation in the frequency domain. The proposed network is shown to be resilient to various reverberation conditions and subtending angles, as compared to the cascaded DNN-array network.

Known for its efficacy on the separation task, Conv-TasNet [19] uses the time-domain learnable analysis and synthesis transformation and time-dilated convolutional blocks as the separation module. Moreover, U-net [44] which constitutes of multiple convolutional layers on the basis of encoder-decoder structure was recently applied and proved its effectiveness on the dereverberation task [45, 46]. In this paper, we build upon Conv-TasNet and U-net to develop a two-stage dereverberation-separation end-to-end system. The proposed network consists of two parts according to the training strategies. In the first part, the network is trained for beamforming network (BF-net), whereas in the second part, a U-net follows as a non-linear postfilter of the BF-net whose parameters are imported from the first part. The experiments are conducted using the proposed network for the spatialized Voice Cloning Toolkit (VCTK) corpus [47]. The results are evaluated in terms of SI-SNR [43], Perceptual Evaluation of Speech Quality (PESQ) [48], and Short-Time Objective Intelligibility (STOI) [49].

## 2 Conventional approaches on separation and dereverberation

Several conventional methods to be used as the baseline approaches are reviewed in this section. The typical processing flow of these methods has a dereverberation unit as the front end, e.g., WPE [50] and a separation unit as the back end, e.g., MPDR [5], TIKR [6], or IVA [13]. The cascaded structure of the DNN method, Beam-TasNet [42], is also considered as the baseline to illustrate the benefit of end-to-end training with SI-SNR.

### 2.1 Dereverberation using the WPE

To account for the prolonged effects of reverberation, a multichannel convolutional signal model [50] for a

single-source scenario is generally formulated in the T-F domain as

$$\mathbf{x}(t, f) = \sum_{l=0}^{L-1} \mathbf{h}(l, f) s(t-l, f), \quad (1)$$

where  $\mathbf{x}(t, f) = [x_1(t, f) \ x_2(t, f) \ \dots \ x_M(t, f)]^T$  is the microphone signal vector and  $\mathbf{h}(l, f) = [h_1(l, f) \ h_2(l, f) \ \dots \ h_M(l, f)]^T$  with  $l = 0, 1, \dots, L$  is the convolutional acoustic transfer functions from the source to the array microphones. A delayed autoregressive linear prediction model can be utilized to estimate recursively the late re-verberation [23].

## 2.2 Dereverberation and separation systems

Three conventional methods and a DNN approach to be used as the baselines are summarized next.

### 2.2.1 The baseline method 1: WPE-MPDR approach

The first baseline method is depicted in Fig. 1. The reverberated mixture signals  $\mathbf{x}(t, f)$  are de-reverberated by the WPE unit and then filtered by the MPDR beamformer. After the de-reverberated signals  $\tilde{\mathbf{x}}(t, f)$  are acquired through WPE, the weight vector of MPDR [5]  $\mathbf{w}_{MPDR}$  can be obtained as

$$\mathbf{w}_{MPDR} = \frac{\mathbf{R}_{xx}^{-1} \mathbf{a}(\theta_n, f)}{\mathbf{a}^H(\theta_n, f) \mathbf{R}_{xx}^{-1} \mathbf{a}(\theta_n, f)}, \quad (2)$$

where  $\mathbf{a}(\theta_n, f) \in \mathbb{C}^M$  is the steering vector associated with the  $n$ th source at the direction  $\theta_n$  and  $\mathbf{R}_{xx} = E\{\tilde{\mathbf{x}}(t, f) \tilde{\mathbf{x}}^H(t, f)\}$  is the spatial covariance matrix with  $E\{\cdot\}$  being the expectation operator with respect to the time frames and can be estimated using recursive averaging. In this paper, the steering vector is modeled with the acoustic transfer function of the free-field plane-wave propagation. We investigate the scenario of the fixed source locations for which the direction of arrivals of source speakers are known.

### 2.2.2 The baseline method 2: WPE-TIKR approach

The baseline method 2 is illustrated in Fig. 2. The microphone signals are de-reverberated by using WPE,

followed by the source signal extraction using TIKR. With the steering matrix  $\mathbf{A}(f) = [\mathbf{a}(\theta_1, f) \ \dots \ \mathbf{a}(\theta_n, f)]$  established with the known source locations, the source signals can be extracted by solving a linear inverse problem for the source signal vector  $\mathbf{s}(t, f)$  in terms of TIKR [6]. That is,

$$\mathbf{s}(t, f) = [\mathbf{A}^H(f) \mathbf{A}(f) + \rho^2 \mathbf{I}]^{-1} \mathbf{A}^H(f) \tilde{\mathbf{x}}(t, f), \quad (3)$$

where  $\rho$  is the regularization parameter that trades off the separability and audio quality of the extracted signals and  $\mathbf{I}$  denotes the identity matrix.

### 2.2.3 The baseline method 3: WPE-IVA approach

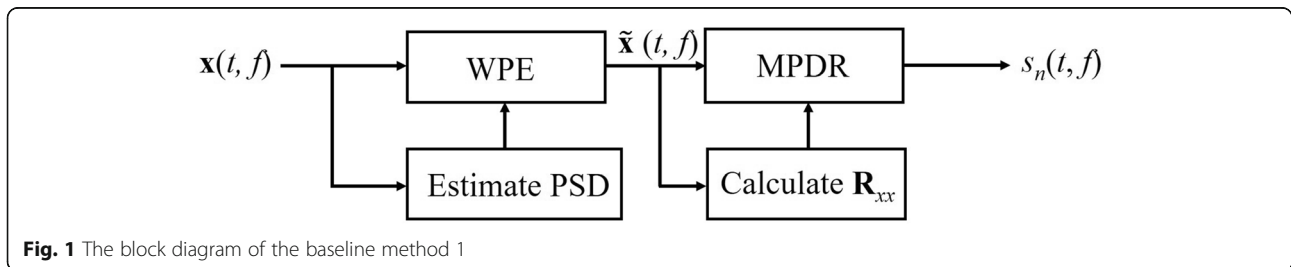
The baseline method 3 is illustrated in Fig. 3. The mixture signals are de-reverberated by WPE, followed by the source signal extraction using IVA. The IVA algorithm resolves the permutation ambiguity in ICA by exploiting the interdependence of frequency components of a particular source. A de-mixing matrix  $\mathbf{W}$  can be calculated using natural gradient method [51]. It follows that the independent source vector  $\hat{\mathbf{s}}$  in the T-F domain can be separated as [13]

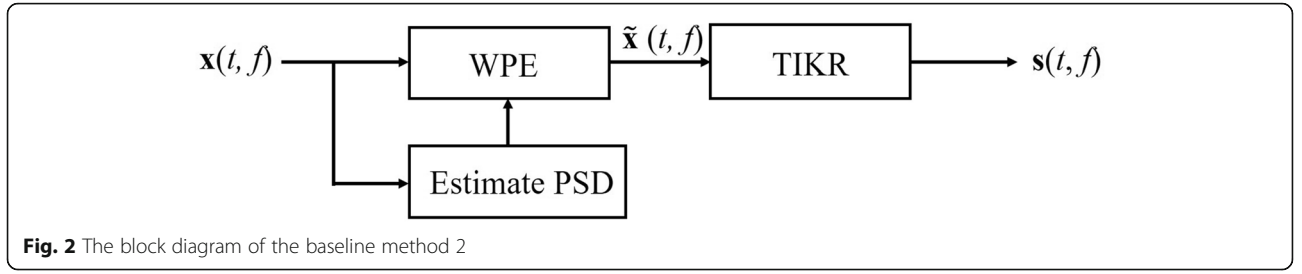
$$\hat{\mathbf{s}}(t, f) = \mathbf{W}(t, f) \tilde{\mathbf{x}}(t, f), \quad (4)$$

To reduce the dimension of the de-reverberated signals when there are more microphones than sources, principle component analysis (PCA) [52] can be used.

### 2.2.4 The baseline method 4: Beam-TasNet approach

In Beam-TasNet, the front-end multichannel TasNet (MC-TasNet) [53] is trained with scale-dependent SNR to estimate the spatial covariance matrix for MVDR that serves as a back-end separator. MC-TasNet consists of a parallel encoder with unconstrained learnable kernels. Once the separated signals are obtained using MC-TasNet, the signal and noise spatial covariance matrices associated with some target source can be estimated. Next, an MVDR beamformer can be implemented with weights:





$$\mathbf{w}_{MVDR} = \frac{\left(\Phi_f^{N_n}\right)^{-1} \Phi_f^{S_n}}{\text{Tr}\left(\left(\Phi_f^{N_n}\right)^{-1} \Phi_f^{S_n}\right)} \mathbf{u}, \quad (5)$$

where  $\Phi_{f_n}^S$  and  $\Phi_{f_n}^N$  denote the signal and noise covariance matrices of the  $n$ th source signals,  $\text{Tr}(\cdot)$  denotes the trace operation, and  $\mathbf{u} = [1 \ 0 \ \cdots \ 0]^T$  is an  $M$ -dimensional vector with one element representing the reference microphone. In this evaluation, the refinement using voice activity detection is not used.

### 3 The proposed multichannel end-to-end NN

In this contribution, an end-to-end multichannel learning-based approach is proposed to separate source signals in reverberant rooms. The network performs joint dereverberation and separation on the basis of Conv-TasNet. Unlike original Conv-TasNet that uses the time-domain learnable transformation to generate features, we use instead STFT and inverse STFT to reduce the computational complexity for our BF-net. In addition, the masks in Conv-TasNet can be reformulated into a learning-based beamformer. Moreover, a U-net is concatenated to the output layer of the BF-net to serve as a postfilter of the beamformer.

#### 3.1 Neural network-based beamforming

In array signal processing, an array aims to recover the source signals via the optimal beamforming weights  $\mathbf{w} \in \mathbb{C}^M$ :

$$\tilde{s}_n(t, f) = \mathbf{w}^H \mathbf{x}(t, f). \quad (6)$$

The learning approach of T-F masks can be applied to the training of the beamforming weights. By converting

the complex representation to the real-valued representation that is amenable to NN platforms, Eq. (6) can be rewritten as

$$\begin{bmatrix} \text{Re}\{\mathbf{x}\}^T & \text{Im}\{\mathbf{x}\}^T \\ \text{Im}\{\mathbf{x}\}^T & -\text{Re}\{\mathbf{x}\}^T \end{bmatrix} \begin{bmatrix} \text{Re}\{\mathbf{w}\} \\ \text{Im}\{\mathbf{w}\} \end{bmatrix} = \begin{bmatrix} \text{Re}\{\tilde{s}_n\} \\ \text{Im}\{\tilde{s}_n\} \end{bmatrix}, \quad (7)$$

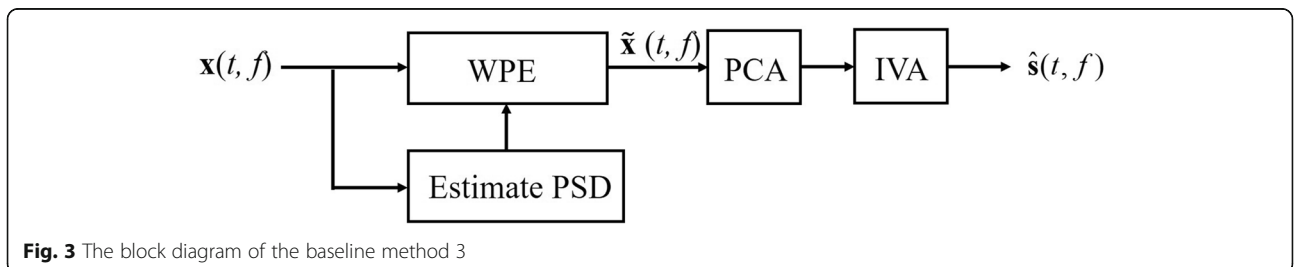
where  $\text{Re}\{\cdot\}$  and  $\text{Im}\{\cdot\}$  denote the real part and imaginary part operations. The goal of the NN training is to obtain the beamforming weights such that the masked signal well approximates the target signal

$$\tilde{\mathbf{S}}_n = \sum_{m=1}^M \text{conj}(\mathbf{W}_m) \circ \mathbf{X}_m, \quad (8)$$

where the  $\{\tilde{\mathbf{S}}_n, \mathbf{W}_m, \mathbf{X}_m\} \in \mathbb{C}^{F \times T}$  denote as the STFT of the  $n$ th target signals, the  $m$ th beamforming weights, and the  $m$ th microphone signal. The symbol “ $\circ$ ” represents element-wise multiplication,  $\text{conj}(\cdot)$  is the conjugate operation element-wisely applied on matrix  $\mathbf{W}_m$ , and  $\{F, T\}$  denote the dimension of T-F bins. The preceding complex STFT representation of the  $n$ th target signal can be converted to its corresponding real part and imaginary part as follows:

$$\begin{aligned} \tilde{\mathbf{S}}_n^r &= \sum_{m=1}^M \mathbf{W}_m^r \circ \mathbf{X}_m^r + \sum_{m=1}^M \mathbf{W}_m^i \circ \mathbf{X}_m^i, \\ \tilde{\mathbf{S}}_n^i &= \sum_{m=1}^M \mathbf{W}_m^r \circ \mathbf{X}_m^i - \sum_{m=1}^M \mathbf{W}_m^i \circ \mathbf{X}_m^r, \end{aligned} \quad (9)$$

where the superscripts  $\{r, i\}$  indicate the real and imaginary part.



### 3.2 Dereverberation via spectral mapping

The reverberated speech signal is pre-processed by the NN-based beamforming to give the  $n$ th enhanced signal  $\tilde{s}_n(t, f)$ . As indicated in the literature [54], the spectral mapping approach is in general more effective than the T-F masking approach for dereverberation problems. Therefore, an additional DNN is employed as a postfilter to learn the non-linear spectral mapping function  $\mathcal{H}(\cdot)$ . The speech signals can be de-reverberated by using this mapping function

$$\hat{s}_n(t, f) = \mathcal{H}(\tilde{s}_n(t, f)). \quad (10)$$

The mapping network  $\mathcal{H}$  is based on a U-net model.

### 3.3 Multichannel network structure

The proposed network depicted in Fig. 4 is comprised of two parts according to the training strategy. At the first stage, the BF-net learns to separate the independent reverberated source signal from the mixture signals received at microphones. At the second stage, the BF-net in conjunction with the U-net postfilter attempts to learn the spectral mapping between the reverberated signal and the anechoic signal of independent sources. To initialize the training, the parameters of the BF-net trained in the first stage are transferred to that in the second stage. In both stages, uPIT [18] is used to avoid permutation ambiguity. The network architectures are detailed next.

#### 3.3.1 The first stage: the weight-and-sum beamforming network

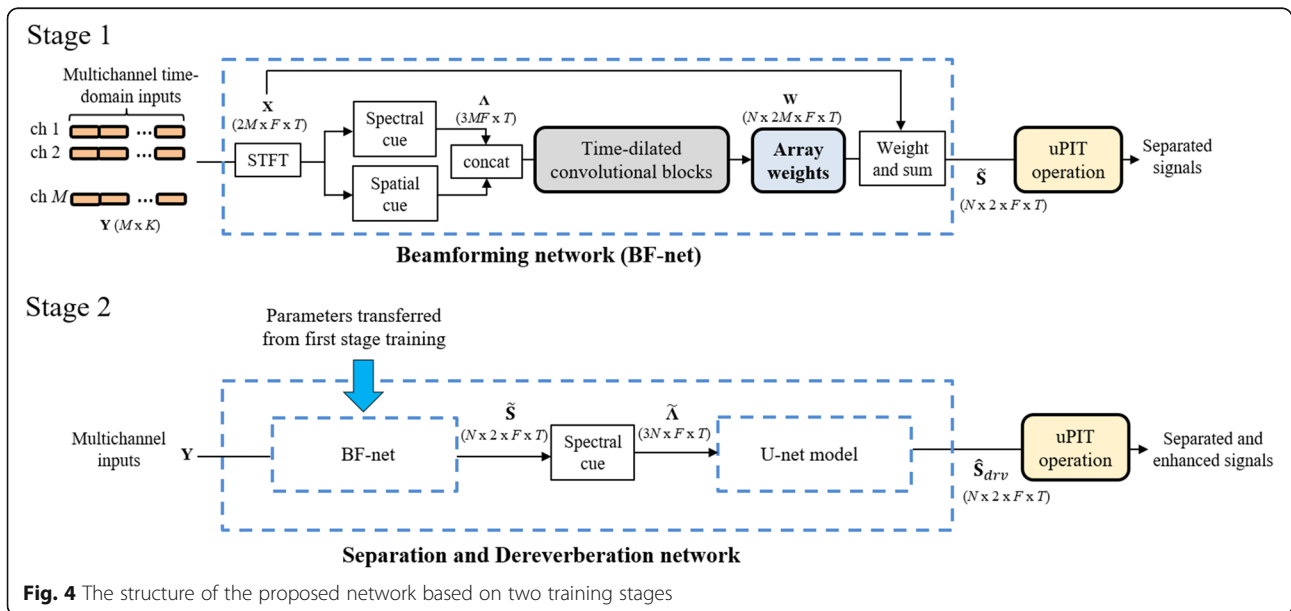
The aim of this network is to generate  $N$  sets of optimal beamforming weights  $\{W_m^r, W_m^i\}_{m=1}^M \in \mathbb{R}^{F \times T}$  for the

weight-and-sum operation in Eq. (9). STFT is utilized to produce the input acoustic features. Inter-channel time, phase, and level differences (ITD, IPD, and ILD) [38] that are commonly used spatial cues can be estimated from the STFT data. In this contribution, we adopt ILD, cosine IPD, and sine IPD defined as

$$\begin{aligned} \text{ILD} &= 10 \log \frac{|x_m(t, f)|}{|x_1(t, f)|}, \\ \cos \text{IPD} &= \cos [\angle x_m(t, f) - \angle x_1(t, f)], \\ \text{and sin IPD} &= \sin [\angle x_m(t, f) - \angle x_1(t, f)], \end{aligned} \quad (11)$$

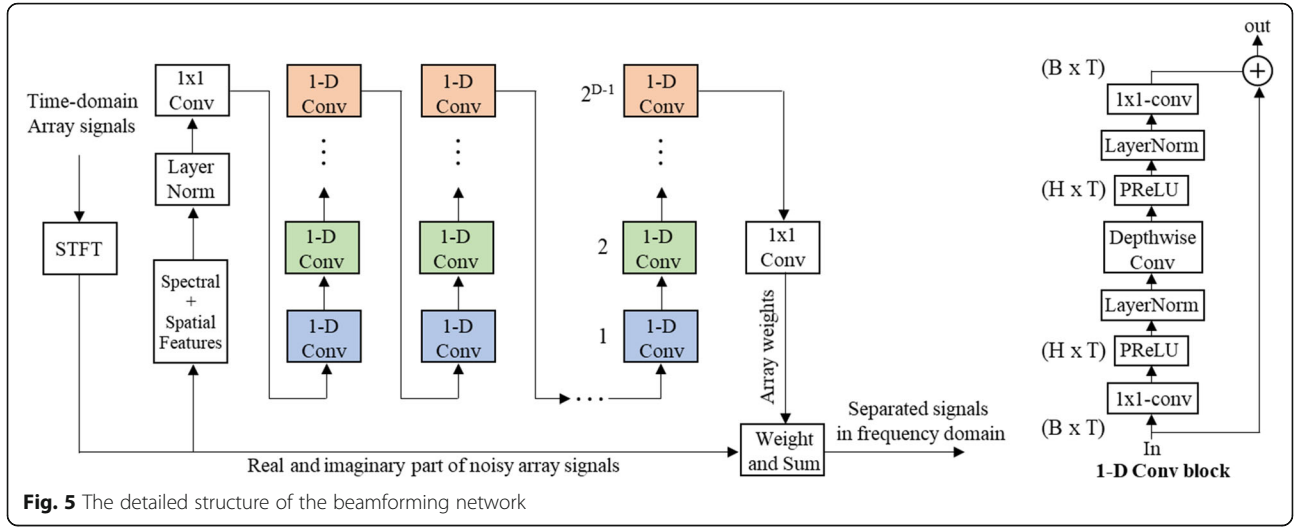
where the first microphone is used as the reference sensor and  $x_m(t, f)$ ,  $m = 2, \dots, M$ , is the STFT of the  $m$ th microphone signal. In addition, the spectral features such as log power spectral density (LPSD), cosine, and sine phase of the first microphone are combined with the spatial features. That is, we concatenate spatial features,  $\{X_{ILD}, X_{\cos IPD}, X_{\sin IPD}\}_{m=1}^M \in \mathbb{R}^{F \times T}$ , and spectral features of the first microphone,  $\{X_{LPSD}, X_{\cos \angle x_1}, X_{\sin \angle x_1}\} \in \mathbb{R}^{F \times T}$  to form the complete features,  $\Lambda \in \mathbb{R}^{3MF \times T}$ , as the input to the BF-net.

The BF-net leverages the main architecture of Conv-TasNet [19] which consists of multiple time-dilated convolutional blocks, as illustrated in Fig. 5. Each layer of the time-dilated blocks contains dilated factors of the number in two's powers ( $2^{D-1}$ ). The input data is zero padded to keep the output dimension for each convolutional block. The increasingly dilated kernel of a block repeats itself  $R$  times. The array weights are estimated through the  $1 \times 1$  pointwise convolutional layer ( $1 \times 1$ -Conv) with no activation function. The network is modified from Conv-TasNet by retaining only the residual



**Fig. 4** The structure of the proposed network based on two training stages





path of the time-dilated CNN blocks. That is, every output of the convolutional block sums with its input to become the input of the next block. The detailed design of the convolution block is shown on the right-hand side of Fig. 5. Before the data is passed to the convolutional block, the input size is adjusted to  $B$  by using a bottleneck layer that is essentially a  $1 \times 1$ -Conv layer. In the convolutional block, the feature is adjusted to larger size  $H > B$  also through a  $1 \times 1$ -Conv layer. Followed by the depthwise separable convolution [55], the separated one-dimensional CNN with kernel size  $P$  convolves with the corresponding input vectors. Next, with the  $1 \times 1$ -Conv, the output size returns to  $B$  in order to merge with the input data to the next layer of the convolutional block. Parametric rectified linear unit (PReLU) is used as the activation function [56], with the aid of the global layer normalization [19].

The curriculum learning [57] is employed in the training stage. The training starts with using the reverberant utterances as the training target, followed by switching the targets to the anechoic utterances when the convergence condition of loss function is met. Finally, the  $N$  sets of separated signals,  $\tilde{\mathbf{S}} \in \mathbb{R}^{N \times 2 \times F \times T}$ , are obtained as described in Fig. 4. The hyperparameters of the non-causal time-dilated convolutional blocks employed in the BF-net are summarized in Table 1. Adam [58] is used as the optimizer with the learning rate  $10^{-3}$ .

### 3.3.2 The second stage: separation and dereverberation network

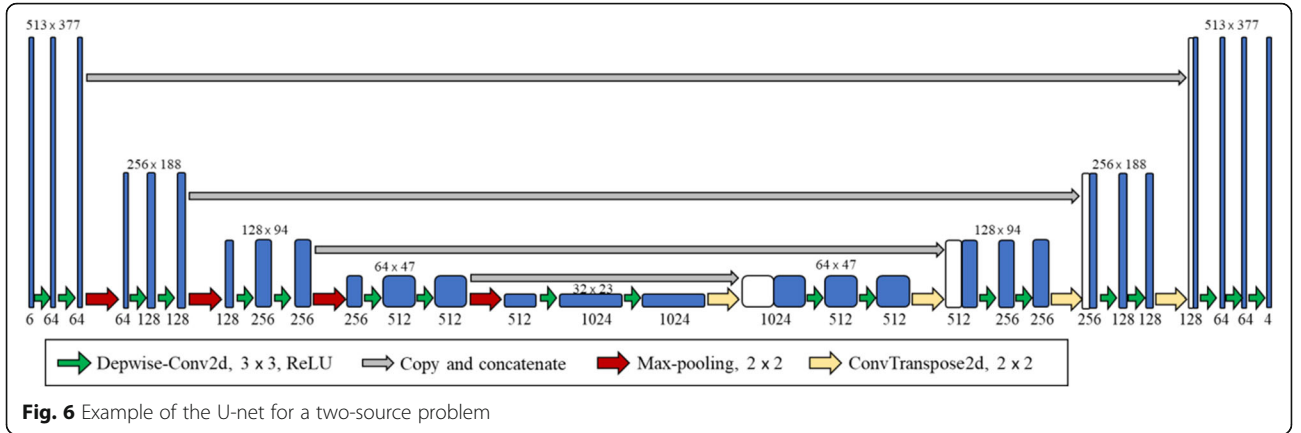
As illustrated in Fig. 4, the BF-net in conjunction with a U-net postfilter is employed in the second stage of joint network training. The U-net postfilter is intended for dereverberation. The parameters trained in the first stage are transferred to the BF-net in the second stage.

The outcome of the training is the direct mapping between the  $N$  sets of the de-reverberated signals,  $\hat{\mathbf{S}}_{drv} \in \mathbb{R}^{N \times 2 \times F \times T}$ , and the anechoic speech signals,  $\mathbf{S} \in \mathbb{R}^{N \times 2 \times F \times T}$ . Before the estimated output of the BF-net,  $\tilde{\mathbf{S}} \in \mathbb{R}^{N \times 2 \times F \times T}$ , is passed to the U-net, the signals in STFT domain are pre-processed to obtain the spectral cues, including LPSD and its corresponding sine and cosine phases,  $\{\tilde{\mathbf{S}}_{LPSD}^n, \tilde{\mathbf{S}}_{\cos Lx}^n, \tilde{\mathbf{S}}_{\sin Lx}^n\}_1^N \in \mathbb{R}^{F \times T}$ . This feature set serves as the input to the U-net model with an appropriate input channel number. For example, if the output number in the first stage is  $N$  separated sources, the pre-processing channel number will be  $3N$ . Hence, the feature size passed to the U-net is  $\tilde{\mathbf{A}} \in \mathbb{R}^{3N \times F \times T}$ .

The U-net model for a two-source problem is depicted in the Fig. 6. The encoder structure consists of two  $3 \times 3$  two-dimensional CNN, where the output is zero-padded to keep the size of the data, followed by a rectified linear unit (ReLU) and a  $2 \times 2$  max-pooling layer with a stride size equal to 2. In a down-sampling step, the number of input channels is doubled and the output features serve as the shared information for the decoder.

**Table 1** Hyper-parameters used in the first stage of the BF-net

Symbol	Parameter	Description
$N_{FFT}$	1024	Number of FFT
$W_{len}$	512	Window size of FFT
$Hop$	128	Hopsize of the window
$B$	256	Channels in bottleneck
$H$	512	Channels in convolutional blocks
$P$	3	Kernel size in convolutional blocks
$D$	6	Convolutional blocks in each repeat
$R$	4	Number of repeats



The decoder up-samples the data through the  $2 \times 2$  transpose convolutional network along with halved feature maps of the input channels, where each is followed by the concatenation of the corresponding maps from the encoder and a repeated  $3 \times 3$  CNN layers with ReLU activation. To accelerate the training process, we also perform the depthwise separable convolution [55] in the consecutive CNN layers. The output layer produces the  $n$ th real and imaginary parts of the enhanced signal in STFT domain,  $\hat{\mathbf{S}}_{drv,n} = \{\hat{\mathbf{S}}_n^r, \hat{\mathbf{S}}_n^i\} \in \mathbb{R}^{F \times T}$ , through a  $1 \times 1$  CNN layer.

The estimated signals can be recovered to the time-domain with the ISTFT process, where the overlap-and-add operation is applied. The network parameters are summarized in Fig. 6, with the channel number indicated and the kernel size of the associated layer labeled at the bottom. During training, Adam [58] is used as the optimizer with the learning rate of  $10^{-4}$ .

### 3.4 The objective function

The time-domain SI-SNR [43] is widely used as the objective function in separation tasks [19, 59]. However, if the system is designed in frequency domain, the direct minimization of the mean square error (MSE) is usually adopted as the objective function, which is not directly related to the separation criterion. Furthermore, because the target signals are usually the T-F spectrogram with a fixed magnitude, the estimated output is basically limited to a certain level. Therefore, the performance of the network will be intrinsically restricted by the definition of the MSE loss function. In order to improve the flexibility of the network output which is trained in the frequency-domain, the scale-invariant MSE (SI-MSE) is formulated by introducing a scaling factor  $\gamma$ :

$$\mathcal{L} = \|\hat{\mathbf{S}}_n - \gamma \mathbf{S}_n\|_F^2, \quad (12)$$

where  $\hat{\mathbf{S}}_n$  and  $\mathbf{S}_n$  are the  $n$ th estimated signal and the target signal in the STFT domain. By minimizing the

objective function with respect to  $\gamma$ , the optimal scaling value  $\gamma$  can be obtained as

$$\gamma = \frac{\sum_{t,f} \hat{\mathbf{S}}_n^r(t,f) \mathbf{S}_n^r(t,f) + \hat{\mathbf{S}}_n^i(t,f) \mathbf{S}_n^i(t,f)}{\sum_{t,f} \mathbf{S}_n^r(t,f)^2 + \mathbf{S}_n^i(t,f)^2}, \quad (13)$$

where the  $\{\hat{\mathbf{S}}_n^r(t,f), \hat{\mathbf{S}}_n^i(t,f)\}$  denote the real and imaginary part of the  $n$ th estimated signal,  $\hat{\mathbf{S}}_n$  in Eq. (12) and so on for the target signal,  $\mathbf{S}_n$ . Therefore, the MSE loss can be rewritten in the form of SI-SNR as

$$SI-SNR(\hat{\mathbf{S}}_n, \gamma \mathbf{S}_n) = 10 \log_{10} \frac{\|\gamma \mathbf{S}_n\|_F^2}{\|\hat{\mathbf{S}}_n - \gamma \mathbf{S}_n\|_F^2}, \quad (14)$$

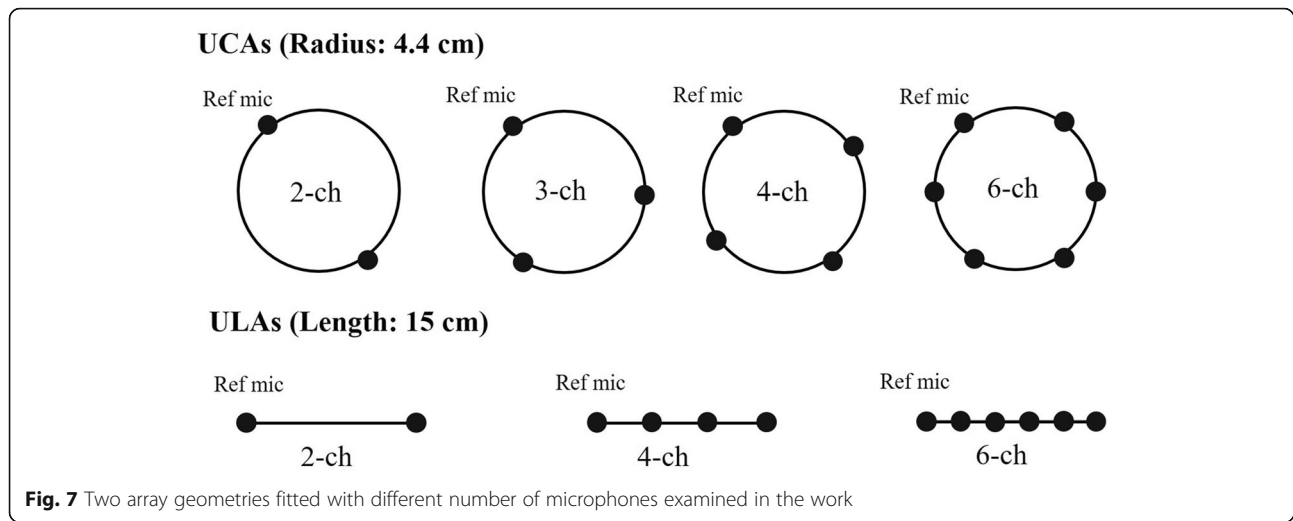
which can be optimized in the frequency domain with a scalable the network output. We adopt this objective function in both training stages and, meanwhile, the uPIT [18] is also employed to prevent the network outputs from the permutation ambiguity error. When the value of SI-SNR in the validation set is no longer decreasing after 10 consecutive epochs, the convergence criterion is said to be met and the training stages will be stopped.

## 4 Results and discussions

### 4.1 Dataset generation

Two array geometries fitted with different number of microphones examined including uniform circular arrays (UCAs) and uniform linear arrays (ULAs). As illustrated in Fig. 7, UCAs of 4.4 cm radius fitted with 2, 3, 4, and 6 microphones are illustrated at the upper row. ULAs of 15 cm fitted with 2, 4, and 6 microphones are illustrated at the lower row.

The dataset generation is considered in a Monte Carlo simulation. Two independent speakers are randomly positioned in rooms with five different sizes. The microphone array is also randomly placed in the same room at half of the room height. The sources are kept at least 0.5 m away from the wall. The two sources are kept at



least 1 m apart, while the distance between the source and the array center is at least 0.7 m. The ranges of the azimuth angles,  $0^\circ$  to  $360^\circ$  and elevation angles,  $0^\circ$  to  $70^\circ$ , are examined. The dataset is remixed from the VCTK corpus [47] where the speech recordings are down-sampled to 16 kHz for our use. Speech segments of 92 speakers are randomly selected for training and validation, whereas 15 unseen speakers are selected for testing. The image source method (ISM) [60] is employed to generate room impulse responses (RIRs) with various reverberation times (T60) ranging from 200 ms to 900 ms. The anechoic signal received at the reference microphone is adopted as the training target. Mixture signals are generated by mixing four-second RIR-filtered utterance segments of two randomly selected speakers. Speech mixture with signal-to-interference-ratio ranging from  $-5$  dB to  $5$  dB used in the training and testing. The simulation settings are summarized in Table 2 and the resulting data size are 30000, 3000 for the training and testing set. The additional 5000 data for the validation are created with the same manner of the training set in order to determine the convergence of the network. To further improve the performance of the network, we also use the dynamic mixing (DM) approach [61] to augment the dataset. The training set is changed to the online data generation, where two randomly selected speech segments are convolved with the pre-generated RIRs and mixed together during the training phase.

#### 4.2 Evaluation of the proposed network

The separation performance of the proposed network is assessed according to the testing set in Table 2. The processed data are evaluated and averaged in terms of the improvement of time-domain SI-SNR [43] ( $\Delta$ SI-SNR), the improvement of PESQ [48] ( $\Delta$ PESQ), and the improvement of STOI [49] ( $\Delta$ STOI) with respect to the unprocessed signal received at the first microphone. In this section, the evaluation is based on the six-element UCA. The models to evaluate are BF-net (the first stage), BF-net with LSTM, BF-net with U-net, and BF-net with U-net and DM. The BF-net (the first stage) refers to the half-trained network where the training is only performed for the first stage. BF-net with LSTM is an alternative network where four layers of the deep long short-term memory (LSTM) with 1024 neurons are adopted as the non-linear postfilter. The BF-net with U-net is the complete model of the proposed network. Moreover, the performance can be further improved by utilizing the DM approach. Two sources with subtending angles within  $0^\circ$ – $15^\circ$ ,  $15^\circ$ – $45^\circ$ ,  $45^\circ$ – $90^\circ$ , and  $90^\circ$ – $180^\circ$  are investigated. The results summarized in Table 3 suggest that separation performance can be improved by the nonlinear postfilter network and adopting DM during training. It can be seen from the  $\Delta$ SI-SNR results, the subtending angle of the two sources has little effect on the performance. However, the  $\Delta$ PESQ score varies significantly with subtending angle.  $\Delta$ PESQ increases for subtending angles less than  $90^\circ$ , slightly decreases for subtending

**Table 2** Data settings of the training and testing set

	Training set					Testing set			
Room height (m)	2.7					3			
Room width $\times$ length (m)	5 $\times$ 4	6 $\times$ 6	8 $\times$ 3	8 $\times$ 5	10 $\times$ 6	4 $\times$ 4	5 $\times$ 7	9 $\times$ 4	12 $\times$ 4
T60 (s)	0.2	0.3	0.4	0.6	0.8	0.16	0.36	0.61	0.9



**Table 3** Performance improvement of the proposed network evaluated with the six-channel UCA for different subtending angles

Model	$\Delta$ SI-SNR (dB)					$\Delta$ PESQ					$\Delta$ STOI
	0°–15°	15°–45°	45°–90°	90°–180°	Avg.	0°–15°	15°–45°	45°–90°	90°–180°	Avg.	
BF-net (the first stage)	7.25	7.62	7.92	8.16	7.78	0.22	0.37	0.52	0.52	0.43	0.14
BF-net + LSTM	7.62	8.73	9.30	9.04	8.86	0.17	0.33	0.52	0.49	0.41	0.15
BF-net + U-net	12.29	12.87	12.51	12.48	12.62	0.32	0.63	0.84	0.79	0.70	0.20
BF-net + U-net + DM	<b>14.32</b>	<b>14.27</b>	<b>13.69</b>	<b>13.81</b>	<b>13.99</b>	<b>0.51</b>	<b>0.78</b>	<b>0.95</b>	<b>0.89</b>	<b>0.83</b>	<b>0.22</b>
Score of unprocessed signal	– 16.72	– 13.15	– 10.93	– 12.95	– 12.73	1.46	1.56	1.60	1.51	1.55	0.59

The best scores for the respective subset are boldfaced

angles larger than 90°. In addition, room responses with different reverberation times,  $T_{60} = 0.16$  s, 0.36 s, 0.61 s, and 0.9 s are also investigated. In Table 4,  $\Delta$ SI-SNR appears to be independent of the reverberation time. We can expect that the proposed network performs better when  $T_{60}$  is low than that of high  $T_{60}$  because the unprocessed signal is not significantly corrupted.  $\Delta$ PESQ also follows the similar trend. The average scores of the performance indices including  $\Delta$ STOI indicate that the six-channel BF-net with U-net and DM turns out to be the best model.

#### 4.3 Comparison with the baseline approaches

In this section, we compare our best model with the traditional BF, BSS, and DNN approaches introduced in the Section 2 where WPE with MPDR and WPE with TIKR are the BF approaches, WPE with IVA is the BSS approach, while the Beam-TasNet approach is the DNN method. The test cases are identical to that discussed in the Section 4.2. The separation performance is summarized in Tables 5 and 6. The results indicate that the proposed network outperforms the baseline methods in three performance metrics. To be specific,  $\Delta$ SI-SNR in Table 5 reveals that the performance of the BF approaches is highly dependent on the subtending angles. For closely spaced sources with the subtending angle within 0°–15°, WPE + TIKR performs poorly. In contrast, the BSS and the proposed learning-based approaches are more robust than the BF approach for

separating closely spaced sources. Furthermore,  $\Delta$ SI-SNR and  $\Delta$ PESQ of the BSS approach and the proposed DNN-based approach exhibit little variation for different subtending angles and reverberation times. Although Beam-TasNet that performs well in  $\Delta$ SI-SNR, enhancement is not satisfactory in terms of  $\Delta$ PESQ and  $\Delta$ STOI in particular when the subtending angle is small or when the reverberation time is large. Because the estimation of the spatial covariance matrix for the MVDR beamformer relies heavily on MC-TasNet, the estimation error has significant impact on the performance of MVDR, especially in adverse acoustic conditions.

#### 4.4 Genericity to different array geometry

To further assess the applicability of the proposed pipeline to different array geometries, two kinds of array geometries fitted with different number of microphones examined in the work. Tables 7 and 8 summarize the performance improvement for both UCAs and ULAs when applied in rooms with different reverberation times. The results in both tables indicate that the proposed network performs well for various numbers of microphones. Furthermore, the performance of the proposed network is increased with number of microphones in both UCAs and ULAs. The results also show that ULA can perform better than UCA when only two microphones are adopted, owing to larger aperture. In summary, the proposed network is applicable to different array geometries if the dataset is properly generated

**Table 4** Performance improvement of the proposed network evaluated with the six-channel UCA for different reverberation time

Model	$\Delta$ SI-SNR (dB)				$\Delta$ PESQ				$\Delta$ STOI
	0.16 s	0.36 s	0.61 s	0.9 s	0.16 s	0.36 s	0.61 s	0.9 s	
BF-net (the first stage)	7.65	7.69	8.10	7.69	0.62	0.46	0.36	0.27	0.14
BF-net + LSTM	9.33	9.70	8.90	7.41	0.63	0.48	0.31	0.18	0.15
BF-net + U-net	11.27	12.48	13.44	13.38	0.95	0.77	0.58	0.48	0.20
BF-net + U-net + DM	<b>12.33</b>	<b>13.78</b>	<b>15.02</b>	<b>14.98</b>	<b>1.10</b>	<b>0.90</b>	<b>0.70</b>	<b>0.59</b>	<b>0.22</b>
Score of unprocessed signal	– 5.67	– 11.45	– 16.17	– 18.27	1.80	1.59	1.43	1.36	0.59

The best scores for the respective subset are boldfaced

**Table 5** Comparison of the separation approaches based on the six-channel UCA for different subtending angles

Model	$\Delta$ SI-SNR (dB)					$\Delta$ PESQ					$\Delta$ STOI
	0°–15°	15°–45°	45°–90°	90°–180°	Avg.	0°–15°	15°–45°	45°–90°	90°–180°	Avg.	
WPE + MPDR	3.27	3.05	3.98	6.17	3.94	– 1.22	– 1.29	– 1.31	– 1.18	– 1.27	0.09
WPE + TIKR	– 5.68	– 0.92	4.89	7.61	2.05	– 0.30	– 1.03	– 1.25	– 1.20	– 1.06	0.11
WPE + IVA	3.80	4.10	4.94	5.98	4.69	<b>0.58</b>	0.64	0.77	0.81	0.71	0.17
Beam-TasNet	7.03	7.55	7.79	7.48	7.57	– 0.01	0.01	0.13	0.08	0.05	0.06
BF-net + U-net + DM	<b>14.32</b>	<b>14.27</b>	<b>13.69</b>	<b>13.81</b>	<b>13.99</b>	0.51	<b>0.78</b>	<b>0.95</b>	<b>0.89</b>	<b>0.83</b>	<b>0.22</b>

The best scores for the respective subset are boldfaced

**Table 6** Comparison of the separation approaches based on the six-channel UCA for different reverberation time

Model	$\Delta$ SI-SNR (dB)				$\Delta$ PESQ				$\Delta$ STOI
	0.16 s	0.36 s	0.61 s	0.9 s	0.16 s	0.36 s	0.61 s	0.9 s	
WPE + MPDR	0.00	4.16	5.69	6.25	– 1.42	– 1.33	– 1.21	– 1.10	0.09
WPE + TIKR	– 0.41	2.44	3.26	3.10	– 1.28	– 1.14	– 0.96	– 0.83	0.11
WPE + IVA	3.88	4.86	5.18	4.88	0.81	0.80	<b>0.70</b>	0.51	0.17
Beam-TasNet	7.37	7.68	7.79	7.42	0.25	0.10	– 0.05	– 0.10	0.06
BF-net + U-net + DM	<b>12.33</b>	<b>13.78</b>	<b>15.02</b>	<b>14.98</b>	<b>1.10</b>	<b>0.90</b>	<b>0.70</b>	<b>0.59</b>	<b>0.22</b>

The best scores for the respective subset are boldfaced

**Table 7** Performance improvement for UCAs with different number of microphones when applied in rooms with different reverberation times

Channel	Model	$\Delta$ SI-SNR (dB)					$\Delta$ PESQ					$\Delta$ STOI
		0.16 s	0.36 s	0.61 s	0.9 s	Avg.	0.16 s	0.36 s	0.61 s	0.9 s	Avg.	
2-ch	BF-net + U-net + DM	6.92	8.25	8.75	8.54	8.10	0.24	0.26	0.25	0.19	0.24	0.10
3-ch		8.41	9.72	10.84	10.63	9.88	0.42	0.39	0.33	0.28	0.36	0.14
4-ch		9.55	11.06	11.69	11.7	11.01	0.49	0.50	0.37	0.34	0.43	0.16
6-ch		<b>12.33</b>	<b>13.78</b>	<b>15.02</b>	<b>14.98</b>	<b>13.99</b>	<b>1.10</b>	<b>0.90</b>	<b>0.70</b>	<b>0.59</b>	<b>0.83</b>	<b>0.22</b>
Score of unprocessed signals		– 5.67	– 11.45	– 16.17	– 18.27	– 12.73	1.80	1.59	1.43	1.36	1.55	0.59

The best scores for the respective subset are boldfaced

**Table 8** Performance improvement for ULAs with different number of microphones when applied in rooms with different reverberation times

Channel	Model	$\Delta$ SI-SNR (dB)					$\Delta$ PESQ					$\Delta$ STOI
		0.16 s	0.36 s	0.61 s	0.9 s	Avg.	0.16 s	0.36 s	0.61 s	0.9 s	Avg.	
2-ch	BF-net + U-net + DM	9.43	10.60	12.04	11.94	11.02	0.55	0.48	0.42	0.32	0.44	0.15
4-ch		11.31	12.52	14.57	14.69	13.29	0.91	0.74	0.59	<b>0.49</b>	0.68	<b>0.20</b>
6-ch		<b>11.37</b>	<b>12.51</b>	<b>14.61</b>	<b>14.72</b>	<b>13.32</b>	<b>0.92</b>	<b>0.76</b>	<b>0.60</b>	0.47	<b>0.69</b>	<b>0.20</b>
Score of unprocessed signals		– 6.10	– 11.94	– 16.24	– 18.50	– 13.25	1.82	1.60	1.44	1.38	1.56	0.59

The best scores for the respective subset are boldfaced

for the corresponding geometries. Nevertheless, the network trained on a UCA cannot be directly utilized on a ULA and re-training is required.

## 5 Conclusions

In this paper, we have proposed a multichannel learning-based DNN and demonstrated its efficacy in source separation in reverberant environments. The end-to-end system relies on a joint training of a BF-net and a U-net. In light of the two-stage training strategy and the DM approach, the proposed six-channel network proves effective in dereverberation and separation. The proposed network has demonstrated superior performance in terms of SI-SNR, PESQ, and STOI, as compared with several baseline methods. The proposed network remains effective, even for closely spaced sources and high reverberation scenarios. Also, the applicability to different array geometries is validated if the dataset is properly generated for the corresponding geometries. However, the network trained on a UCA cannot be utilized directly on a ULA, and vice versa.

Despite the excellent performance of the DNN-based approach, it is noteworthy to mention some of its limitations. It is a “black box” approach in which physical insights play little role. Big data are required for training the network, which is difficult if not impossible in applications. Generalization may be limited if the dataset is not sufficiently comprehensive. These limitations to the DNNs turn out to be the strengths of the BF and BSS approaches. Network integration to create the synergy of these techniques is on the future research agenda.

## Abbreviations

SI-MSE: Scale invariant mean square error; SI-SNR: Scale invariant signal-to-noise ratio; BSS: Blind source separation; BF: Beamforming; BF-net: Beamforming network; MPDR: Minimum power distortionless response; TIKR: Tikhonov regularization; T-F: Time-frequency; IVA: Independent vector analysis; DNN: Deep neural network; CNN: Convolutional neural network; uPIT: Utterance-level permutation invariant training; Conv-TasNet: Fully convolutional time-domain audio separation network; WPE: Weighted prediction error; STFT: Short-time Fourier transform; PESQ: Perceptual evaluation of speech quality; STOI: Short-time objective intelligibility;  $1 \times 1$ -Conv:  $1 \times 1$  pointwise convolutional layer; IPD: Inter-channel phase differences; ILD: Inter-channel level differences; LPSD: Log power spectral density; UCA: Uniform circular array; ULA: Uniform linear array; DM: Dynamic mixing

## Acknowledgements

Thanks to Dr. Mingsian Bai for his three-month visit to the LMS, FAU, Erlangen-Nuremberg, which made this research work possible.

## Authors' contributions

Model development: Y.S. Chen, Z.J. Lin, M. R. Bai. Design of the dataset and test cases: Y.S. Chen and Z.J. Lin. Experimental testing: Y.S. Chen and Z.J. Lin. Writing paper: Y.S. Chen. All the authors review and approved the final manuscript.

## Funding

The work was supported by the Add-on Grant for International Cooperation (MAGIC) of the Ministry of Science and Technology (MOST) in Taiwan, under the project number 107-2221-E-007-039-MY3.

## Availability of data and materials

The demonstration of the processed audio samples can be found at: <https://Siang-Chen.github.io/>

## Declarations

### Competing interests

The authors declare that they have no competing interests.

Received: 14 June 2021 Accepted: 13 October 2021

Published online: 20 November 2021

## References

1. I. McCowan, *Microphone arrays: a tutorial* (Queensland University, Australia, 2001), p. 1
2. F. Gustafsson, F. Gunnarsson, in *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)*. Positioning using time-difference of arrival measurements, vol 6 (2003), pp. VI-553
3. Z. Khan, M.M. Kamal, N. Hamzah, K. Othman, N. Khan, in *2008 IEEE International RF and Microwave Conference*. Analysis of performance for multiple signal classification (MUSIC) in estimating direction of arrival (2008), pp. 524–529
4. K. Nakadai, K. Nakamura, in *Wiley Encyclopedia of Electrical and Electronics Engineering*. Sound source localization and separation, (New York: John Wiley & Sons, 2015), pp. 1–18
5. S.A. Vorobyov, Principles of minimum variance robust adaptive beamforming design. *Signal Process.* **93**, 3264 (2013)
6. M. Fuhry, L. Reichel, A new Tikhonov regularization method. *Numerical Algorithms* **59**, 433 (2012)
7. S. Amari, S.C. Douglas, A. Cichocki, H.H. Yang, in *First IEEE Signal Processing Workshop on Signal Processing Advances in Wireless Communications*. Multichannel blind deconvolution and equalization using the natural gradient (1997), pp. 101–104
8. M. Kawamoto, K. Matsuoka, N. Ohnishi, A method of blind separation for convolved non-stationary signals. *Neurocomputing* **22**, 157 (1998)
9. T. Takatani, T. Nishikawa, H. Saruwatari, K. Shikano, in *Seventh International Symposium on Signal Processing and Its Applications, 2003. Proceedings*. High-fidelity blind separation for convolutive mixture of acoustic signals using simo-model-based in-dependent component analysis, vol 2 (2003), pp. 77–80
10. D.W. Schobben, P. Sommen, A frequency domain blind signal separation method based on decorrelation. *IEEE Trans. Signal Process.* **50**, 1855 (2002)
11. S. Makino, H. Sawada, S. Araki, in *Blind Speech Separation*. Frequency-domain blind source separation (Dordrecht: Springer, 2007), pp. 47–78
12. H. Buchner, R. Aichner, W. Kellermann, A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics. *IEEE Trans. Speech Audio Process.* **13**, 120 (2004)
13. T. Kim, I. Lee, T.-W. Lee, in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*. Independent vector analysis: definition and algorithms (2006), pp. 1393–1396
14. Y. Wang, D. Wang, Towards scaling up classification-based speech separation. *IEEE Trans. Audio Speech Lang. Process.* **21**, 1381 (2013)
15. S. Mobin, B. Cheung, B. Olshausen, *Generalization challenges for neural architectures in audio source separation*, *arXiv preprint arXiv:1803.08629* (2018)
16. P.-S. Huang, M. Kim, M. Hasegawa-Johnson, P. Smaragdis, Joint optimization of masks and deep re-current neural networks for monaural source separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **23**, 2136 (2015)
17. J.R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep clustering: discriminative embeddings for segmentation and separation (2016), pp. 31–35
18. M. Kolbæk, D. Yu, Z.-H. Tan, J. Jensen, Mul-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Trans. Audio Speech Lang. Process.* **25**, 1901 (2017)
19. Y. Luo, N. Mesgarani, Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **27**, 1256 (2019)
20. K. Furuya, S. Sakauchi, A. Kataoka, in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. Speech dereverberation

- by combining MINT-based blind deconvolution and modified spectral subtraction, vol 1 (2006), p. I–I
21. T. Nakatani, B.-H. Juang, T. Yoshioka, K. Kinoshita, M. Miyoshi, in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. Importance of energy and spectral features in gaussian source model for speech dereverberation (New Paltz: IEEE, 2007), pp. 299–302
  22. T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, B.-H. Juang, in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation (2008), pp. 85–88
  23. T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, B.-H. Juang, Speech dereverberation based on variance-normalized delayed linear prediction. *IEEE Trans. Audio Speech Lang. Process.* **18**(1717) (2010)
  24. T. Yoshioka, T. Nakatani, M. Miyoshi, H.G. Okuno, Blind separation and dereverberation of speech mixtures by joint optimization. *IEEE Trans. Audio Speech Lang. Process.* **19**(69) (2010)
  25. A. Jukić, N. Mohammadiha, T. van Waterschoot, T. Gerkmann, S. Doclo, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Multi-channel linear prediction-based speech dereverberation with low-rank power spectrogram approximation (2015), pp. 96–100
  26. F. Weninger, S. Watanabe, Y. Tachioka, B. Schuller, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep recurrent de-noising auto-encoder and blind dereverberation for reverberated speech recognition (2014), pp. 4623–4627
  27. D.S. Williamson, D. Wang, in *2017 IEEE International conference on acoustics, speech and signal processing (ICASSP)*. Speech dereverberation and denoising using complex ratio masks (2017), pp. 5590–5594
  28. J. Heymann, L. Drude, R. Haeb-Umbach, K. Kinoshita, T. Nakatani, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Joint optimization of neural network-based WPE dereverberation and acoustic model for robust online ASR (2019), pp. 6655–6659
  29. K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, T. Nakatani, in *Interspeech*. Neural network-based spectrum estimation for online wpe dereverberation (2017), pp. 384–388
  30. M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, S. Araki, T. Hori, et al., Strategies for distant speech recognition in reverberant environments. *EURASIP J. Adv. Signal Process.* **2015**, 1 (2015)
  31. W. Yang, G. Huang, W. Zhang, J. Chen, J. Benesty, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Dereverberation with differential microphone arrays and the weighted-prediction-error method (2018), pp. 376–380
  32. M. Togami, in *2015 23rd European Signal Processing Conference (EUSIPCO)*. Multichannel online speech dereverberation under noisy environments (2015), pp. 1078–1082
  33. L. Drude, C. Boeddeker, J. Heymann, R. Haeb-Umbach, K. Kinoshita, M. Delcroix, T. Nakatani, in *Interspeech*. Integrating neural network based beamforming and weighted prediction error dereverberation (2018), pp. 043–047
  34. T. Nakatani, K. Kinoshita, A unified convolutional beamformer for simultaneous denoising and dereverberation. *IEEE Signal Process. Lett.* **26**, 903 (2019)
  35. G. Wichern, J. Antognini, M. Flynn, L.R. Zhu, E. McQuinn, D. Crow, E. Manilow, J.L. Roux, *Wham!: Extending speech separation to noisy environments*, *arXiv preprint arXiv:1907.01160* (2019)
  36. C. Ma, D. Li, X. Jia, *Two-stage model and optimal si-snr for monaural multi-speaker speech separation in noisy environment*, *arXiv preprint arXiv:2004.06332* (2020)
  37. T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, D. Dimitriadis, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Low-latency speaker-independent continuous speech separation (2019), pp. 6980–6984
  38. Z.-Q. Wang, D. Wang, in *Interspeech*. Integrating spectral and spatial features for multi-channel speaker separation (2018), pp. 2718–2722
  39. J. Wu, Z. Chen, J. Li, T. Yoshioka, Z. Tan, E. Lin, Y. Luo, L. Xie, *An end-to-end architecture of online multi-channel speech separation*, *arXiv preprint arXiv:2009.03141* (2020)
  40. T. Nakatani, R. Takahashi, T. Ochiai, K. Kinoshita, R. Ikeshita, M. Delcroix, S. Araki, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. DNN-supported mask-based convolutional beamforming for simultaneous denoising, dereverberation, and source separation (2020), pp. 6399–6403
  41. Y. Fu, J. Wu, Y. Hu, M. Xing, L. Xie, in *2021 IEEE Spoken Language Technology Workshop (SLT)*. DESNET: A multi-channel network for simultaneous speech dereverberation, enhancement and separation (2021), pp. 857–864
  42. T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, S. Araki, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Beam-Tasnet: Time-domain audio separation network meets frequency-domain beam-former (2020), pp. 6384–6388
  43. J. Le Roux, S. Wisdom, H. Erdogan, J.R. Hershey, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. SDR-half-baked or well done? (2019), pp. 626–630
  44. O. Ronneberger, P. Fischer, T. Brox, in *International Conference on Medical image computing and computer-assisted intervention*. U-net: Convolutional networks for biomedical image segmentation (Cham: Springer, 2015), pp. 234–241
  45. O. Ernst, S.E. Chazan, S. Gannot, J. Goldberger, in *2018 26th European Signal Processing Conference (EUSIPCO)*. Speech dereverberation using fully convolutional networks (2018), pp. 390–394
  46. V. Kothapally, W. Xia, S. Ghorbani, J.H. Hansen, W. Xue, J. Huang, *Skipconvnet: Skip convolutional neural network for speech dereverberation using optimally smoothed spectral mapping*, *arXiv preprint arXiv:2007.09131* (2020)
  47. J. Yamagishi, C. Veaux, K. MacDonald, et al., *Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92)* (2019).
  48. A.W. Rix, J.G. Beerends, M.P. Hollier, A.P. Hekstra, in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs, vol 2 (2001), pp. 749–752
  49. C.H. Taal, R.C. Hendriks, R. Heusdens, J. Jensen, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. A short-time objective intelligibility measure for time-frequency weighted noisy speech (2010), pp. 4214–4217
  50. K. Kinoshita, M. Delcroix, T. Nakatani, M. Miyoshi, Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction. *IEEE Trans. Audio Speech Lang. Process.* **17**, 534 (2009)
  51. S.-I. Amari, A. Cichocki, H.H. Yang, et al., in *Advances in neural information processing systems*. A new learning algorithm for blind signal separation (1996), pp. 757–763
  52. S. Wold, K. Esbensen, P. Geladi, Principal component analysis. *Chemom. Intell. Lab. Syst.* **2**, 37 (1987)
  53. R. Gu, J. Wu, S. Zhang, L. Chen, Y. Xu, M. Yu, D. Su, Y. Zou, D. Yu, *End-to-end multi-channel speech separation*, *arXiv preprint arXiv:1905.06286* (2019)
  54. Y. Zhao, Z.-Q. Wang, D. Wang, in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A two-stage algorithm for noisy and reverberant speech enhancement (2017), pp. 5580–5584
  55. F. Chollet, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Xception: Deep learning with depthwise separable convolutions (2017)
  56. K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE international conference on computer vision*. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification (2015), pp. 1026–1034
  57. Y. Bengio, J. Louradour, R. Collobert, J. Weston, in *Proceedings of the 26th annual international conference on machine learning*. Curriculum learning (2009), pp. 41–48
  58. D.P. Kingma, J. Ba, *Adam: A method for stochastic optimization*, *arXiv preprint arXiv:1412.6980* (2014)
  59. F. Bahmaninezhad, J. Wu, R. Gu, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, *A comprehensive study of speech separation: spectrogram vs waveform separation*, *arXiv preprint arXiv:1905.07497* (2019)
  60. J.B. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics. *J. Acoust. Soc. Am.* **65**, 943 (1979)
  61. N. Zeghidour, D. Grangier, *Wavesplit: End-to-end speech separation by speaker clustering*, *arXiv preprint arXiv:2002.08933* (2020)

## Publisher's Note

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