ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414187

DBNET: DOA-DRIVEN BEAMFORMING NETWORK FOR END-TO-END REVERBERANT SOUND SOURCE SEPARATION

Ali Aroudi[†⋆], *Sebastian Braun*[†]

[†] Microsoft Corporation Redmond, WA, USA
[⋆] Department of Medical Physics and Acoustics, University of Oldenburg, Germany
ali.aroudi@uni-oldenburg.de, sebastian.braun@microsoft.com

ABSTRACT

Many deep learning techniques are available to perform source separation and reduce background noise. However, designing an end-to-end multi-channel source separation method using deep learning and conventional acoustic signal processing techniques still remains challenging. In this paper we propose a direction-of-arrival-driven beamforming network (DBnet) consisting of direction-of-arrival (DOA) estimation and beamforming layers for end-to-end source separation. We propose to train DBnet using loss functions that are solely based on the distances between the separated speech signals and the target speech signals, without a need for the ground-truth DOAs of speakers. To improve the source separation performance, we also propose end-to-end extensions of DBnet which incorporate post masking networks. We evaluate the proposed DBnet and its extensions on a very challenging dataset, targeting realistic far-field sound source separation in reverberant and noisy environments. The experimental results show that the proposed extended DBnet using a convolutional-recurrent post masking network outperforms state-of-the-art source separation methods.

Index Terms— sound source separation, deep learning, beamforming, direction of arrival estimation

1. INTRODUCTION

Environmental noise, reverberation and interfering sound sources negatively affect the quality of the speech signals received at the microphones and therefore degrade the performance of many speech communication systems including automatic speech recognition systems, hearing assistive devices and mobile devices. In recent years several deep learning techniques have been proposed to separate out the speakers from the microphone signals and reduce background noise based on, e.g., the frequency domain transformation [1, 2, 3, 4] or a learned latent domain transformation [4, 5, 6, 7]. The frequency-domain-based techniques perform sound source separation typically by estimating time-frequency masks corresponding to each source, while the latent-domain-based techniques aim at learning a latent space from the time-domain signals to perform source separation. Although most of these techniques are able to separate out speakers from single-channel microphone signals, designing a multi-channel source separation method that properly leverages all inter-channel information remains to be solved.

When multi-channel inputs are available or a physical interpretation of a signal is possible conventional acoustic signal processing, e.g., beamforming and direction-of-arrival estimators (DOA), have analytical solutions and reasonably good performace in many cases. This motivates to integrate conventional acoustic signal processing techniques and deep learning techniques to profit from both worlds, as has been proposed by several works [8, 9, 10]. However, the integration of techniques is typically performed in a modular way where each module is optimized individually, which may lead to non-optimal solution. More recently, a filter-and-sum network with an end-to-end training has been proposed using time-domain filtering, which does not benefit from computationally efficient filtering in the frequency domain [11]. In this paper we propose a DOA-driven beamforming network (DBnet) operating in the frequency domain, aiming at end-to-end sound source separation where the gradient is propagated in an end-to-end optimization way through the time domain. As opposed to most existing techniques in literature, we evaluate the proposed DBnet on a very challenging and realistic large-scale dataset, targeting realistic far-field sound source separation in reverberant and noisy environments.

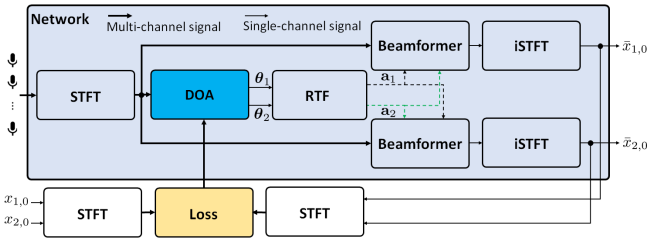


Fig. 1. Block diagram of the proposed DBnet structure

The proposed DBnet is depicted in Fig. 1. DBnet consists of layers for short-time Fourier transform (STFT), DOA estimation, beamforming and inverse STFT (iSTFT). We train DBnet using loss functions that are solely based on the distances between the separated speech signals and the target speech signals. In this way, the network is trained to inherently estimate the DOAs of speakers leading to the best possible separated speech signals while the need of the ground truth DOAs for training the network is eliminated. Although DBnet is able to separate the sources and suppress noise, the output signals may still contain residual noise. Therefore, we also propose end-to-end extensions of DBnet which incorporate a post masking network (pMnet). For pMnet we consider a convolutional-recurrent network with an encoder-decoder architecture which combines frequency and latent-domain based techniques. We train the networks using several loss functions based on complex spectra distances, magnitude distances or a combination of both distances. In addition, aiming at developing the networks to be well-generalized for unseen acoustic conditions, we generate a large-scale training dataset using several augmentation steps.

We experimentally compare the proposed DBnet and DBnet extensions with the state-of-the-art source separation methods for challenging noisy and reverberant conditions. The results show that the proposed extended DBnet outperforms state-of-the-art source separation methods.

2. END-TO-END SOUND SOURCE SEPARATION

2.1. Problem formulation

We consider an acoustic scenario comprising two competing speakers and background noise in a reverberant environment. We consider an array with M microphones. The sound captured at the m-th microphone signal can be decomposed as

$$y_m[n] = \sum_{i=1}^{2} x_{i,m}[n] + v_m[n], \tag{1}$$

where $x_{i,m}[n]$ denotes the direct speech component in the m-th microphone signal corresponding to speaker i, $v_m[n]$ denotes the noise component representing reverberation, background noise and any remaining components and n denotes the discrete time index.

In the short-time Fourier transform (STFT) domain, the M-dimensional stacked vector of all microphone signals can be given by

$$\mathbf{y}(k,f) = \sum_{i=1}^{2} \mathbf{a}_i(k,f) X_{i,0}(k,f) + \mathbf{v}(k,f), \tag{2}$$

where $\mathbf{a}_i(k,f)$ denotes the direct relative transfer function vector (DRTF) corresponding to source i, $X_{i,0}$ denotes the STFT coefficient of $x_{i,0}[n]$, $\mathbf{v}(k,f)$ denotes the noise component, and k and f are the frame index and the frequency index. Assuming the sources are in the far field of omnidirectional microphones, the DRTF $\mathbf{a}_i(k,f)$ can be given as

$$\mathbf{a}_i(k,f) = e^{j\kappa(k)\mathbf{R}_{\text{mic}}\mathbf{r}_i}, \tag{3}$$

where $\mathbf{R}_{\text{mic}}$ denote the $M \times 3$ Cartesian microphone array coordinates, $\mathbf{r}_i$ is the Cartesian i-th source position, normalized to unit distance, and $\kappa(k) = \frac{2\pi k f_s}{c N_{\text{FFT}}}$ is the wavenumber with c denoting the speed of sound, f_s denoting the sampling rate, and N_{FFT} denoting the fast Fourier transform (FFT) size. With the far-field assumption, the position of i-th source $\mathbf{r}_i$ in the far field in (3) can be assumed as distance-independent, and therefore only depends on elevation and azimuth angles, $\theta_{\text{el},i}$ and $\theta_{\text{az},i}$, corresponding to the i-th source.

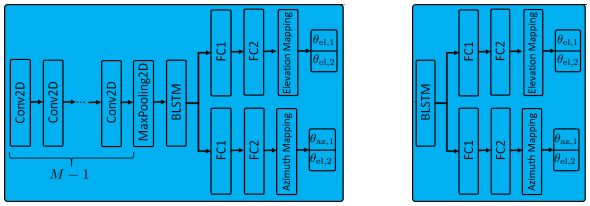
Our goal is to separate out the direct speech signal components of the speakers from the microphone signals in the time domain using a network h with linear and non-linear layers operating in the frequency and latent domains, i.e.,

$$\bar{x}_{i,0}[n] = h(y_m[n]). \tag{4}$$

2.2. DBnet

The DBnet accepts microphone signals $y_m[n]$ as input signals and computes their corresponding STFTs using the first layer (see Fig. 1). The DOA estimation layer with learnable parameters then estimates the DOAs of both sources from the phase spectrograms of input signals. Based on the estimated DOAs, the DRTFs of sources are estimated using (3) to steer two parallel beamforming layers separating out the speech sources. The output signals of the beamformers are finally transformed to the time domain using the iSTFT layer. We explore either convolutional-recurrent (DBnet(CR)), or fully recurrent structures (DBnet(R)) for DOA estimation, described in the following.

The first DOA estimation architecture is a convolutional-recurrent structure (DBnet(CR)), to learn the relevant features from phase spectrograms and to model the sequence of the learned features as shown in Fig. 2a. We use $M-1$ 2D-convolutional layers with 64 filters of size $m \times f = 2 \times 1$ to learn the phase correlations between adjacent microphones at each frequency sub-band separately as proposed in [12].



(a) convolutional-recurrent structure (b) recurrent structure

Fig. 2. Structure of the DOA estimators

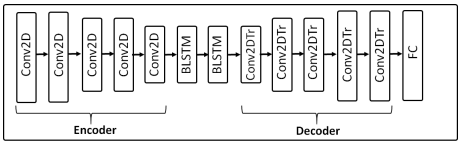


Fig. 3. Structure of the convolutional-recurrent pMnet

These learned features are then pooled using a 2D max pooling layer to increase the robustness of features against background noise. The sequence of these features is modeled using a BLSTM layer [13] with 1200 units. The outputs of the BLSTM layer then go through two parallel fully connected layers with Sigmoid functions followed by two mapping blocks to estimate the azimuth and elevation angles of the DOAs of both sources. The azimuth and elevation mapping blocks map the output of the fully connected layers to azimuth angle $\theta_{\text{el},i} \in [-175°, 185°]$ and elevation angle $\theta_{\text{az},i} \in [-175°, 185°]$ ranges, respectively.

The second DOA estimation architecture is a fully recurrent (DBnet(R)) structure, consisting only of the BLSTM, 2D max pooling, fully connected layers, and mapping blocks as shown in Fig. 2b. For both DOA estimation architectures, all 2D-convolutional layers use strides of $k \times f = 1 \times 1$ and the 2D max pooling layer uses strides of 32×32, downsampling across time and frequency dimensions. The BLSTM layers is an aggregation step returning only the last hidden state, resulting in time-and-frequency-invariant outputs.

For beamforming, we use the linearly-constrained-minimum-variance (LCMV) beamformer [14], minimizing spatially diffuse noise, while preserving the target source signal and spatially nulling the interfering source. The LCMV beamformer is computed from the DRTFs corresponding to the estimated source angles $\{\theta_{\text{el},i}, \theta_{\text{az},i}\}$ using (3), and the isotropic noise field covariance matrix [15].

2.3. DBnet extensions with post masking

Since the suppression capability of linear spatial filters such as the LCMV beamformers used in DBnet is limited, the output signals of the DBnet may still contain residual noise. Therefore, we propose an extension of DBnet, which incorporates a post masking-based network (pMnet) to suppress the residual noise. pMnet takes the log magnitude spectrum of the beamformer output signals and generates real-valued time-frequency masks corresponding to each source. Motivated by the results in [1, 4] where a masking-based network with recurrent layers was used for speech separation, we consider a recurrent pMnet consisting of four BLSTM layers followed by one fully connected layer to estimate the stacked masks corresponding to the sources. As an alternative to the recurrent pMnet, we also consider a convolutional-recurrent pMnet with an encoder-decoder network architecture (see Fig. 3), similarly as used in [16] but for speech enhancement. The 2D-convolutional layers encode the input magnitude spectrums into a higher-dimensional latent features and two BLSTM layers model the sequence of latent features. The outputs of the BLSTM

Table 1. Loss functions

Loss function	Complex ($\mathcal{L}_{\text{complex}}$)		Magnitude ($\mathcal{L}_{\text{magnitude}}$)	
MSE	$\log_{10}\sum_{k,f}\left\|X_{i,0}(k,f)-\overline{X}_{i,0}(k,f)\right\|^2$	(3)	$\log_{10}\sum_{k,f}\left\|\|X_{i,0}(k,f)\|-\left\|\overline{X}_{i,0}(k,f)\right\|\right\|^2$	(4)
cMSE	$\log_{10}\sum_{k,f}\left\|\|X_{i,0}(k,f)\|^c e^{j\varphi_X}-\left\|\overline{X}_{i,0}(k,f)\right\|^c e^{j\varphi_{\overline{X}}}\right\|^2$	(5)	$\log_{10}\sum_{k,f}\left\|\|X_{i,0}(k,f)\|^c-\left\|\overline{X}_{i,0}(k,f)\right\|^c\right\|^2$	(6)
MAE	$\log_{10}\sum_{k,f}\left\|X_{i,0}(k,f)-\overline{X}_{i,0}(k,f)\right\|$	(7)	$\log_{10}\sum_{k,f}\left\|\|X_{i,0}(k,f)\|-\left\|\overline{X}_{i,0}(k,f)\right\|\right\|$	(8)
SDR	$-\log_{10}\frac{\sum_{k,f}\left\|X_{i,0}(k,f)\right\|^2}{\sum_{k,f}\left\|X_{i,0}(k,f)-\overline{X}_{i,0}(k,f)\right\|^2}$	(9)		

layers are converted back to the original input dimension using mirrored transposed 2D-convolutional layers. The output of the encoder-decoder layers go then through a feed forward layer to estimate the masks of both sources. For both recurrent and convolutional-recurrent pMnets, the BLSTM layers have 1200 units and the fully connected layers have 514 units with Sigmoid functions. For the convolutional-recurrent pMnet, the sequence of 2D-convolutional layers has 16, 16, 32, 32, 64 filters with kernel sizes of $k\times f=6\times 6$ and strides of 1×2. The sequence of transposed 2D-convolutional layers correspondingly has 32, 32, 16, 16, 2 filters. All convolutional layers are followed by leaky ReLU functions.

To investigate the impact of the post masking on the source separation performance, we consider the following end-to-end DBnet extensions:

- **DBnet(R)-pMnet(R)**: recurrent DBnet followed by the recurrent Mnet
- **DBnet(R)-pMnet(CR)**: recurrent DBnet followed by the convolutional-recurrent pMnet
- **DBnet(CR)-pMnet(R)**: convolutional-recurrent DBnet followed by the recurrent pMnet
- **DBnet(CR)-pMnet(CR)**: convolutional-recurrent DBnet followed by the convolutional-recurrent Mnet

2.4. Baseline method: masking-based source separation

As baseline methods we consider a recurrent masking-based network (Mnet(R)) and a convolutional-recurrent masking-based network (Mnet(CR)). The Mnet(R) has a similar structure as used for pMnet(R) with only one difference that the log magnitude spectrum of a reference microphone signal is used as input. The Mnet(CR) uses the same input and has a similar structure as used for pMnet(CR). The baseline methods use only the reference microphone from the array.

3. LOSS FUNCTIONS

For training the networks we consider several loss functions using either complex spectral distance or magnitude distance between the STFT of the separated speech signals $\overline{X}_{i,0}(k,f)$ and the STFT of the target speech signals $X_{i,0}(k,f)$, given in Table 1. We consider four different loss functions in the following: i) the mean squared error (MSE) given by (3) and (4) [17]; ii) the compressed MSE (cMSE) given by (5) and (6), which is computed based on the magnitudes compressed with an exponent $c=0.3$ [18, 19], in order to deal with the large dynamic ranges of audio signals; iii) the mean absolute error (MAE) loss functions given by (7) and (8), promoting speech sparsity [17]; iv) the commonly used scale-variant signal-to-distortion ratio (SDR) (9) [20, 10].

In addition, we consider loss functions which combine the complex spectral distance and the magnitude distance per row of Table 1, i.e., $\mathcal{L}=\alpha\mathcal{L}_{\text{complex}}+(1-\alpha)\mathcal{L}_{\text{magnitude}}$ where $0\leq\alpha\leq 1$ denotes the loss combining factor. The source-to-output mapping problem is solved by using utterance permutation invariant training (uPIT) as proposed in [1] is employed.

4. DATASETS

Since the performance of neural networks are highly data dependent, we put large effort in building realistic and large enough training and test sets to draw valid conclusions. We separate training, validation and test data as much as possible using different datasets where possible. Since room impulse responses (RIRs) for our targeted microphone array are unfeasible to obtain from measurements in large quantity, we simulate RIRs for training, while we use measured RIRs for validation and testing to ensure generalization of the networks to real-world acoustic conditions.

4.1. Training, validation and test data

For training, validation and testing, we use three different speech databases, i.e. 540 h of speech data from audiobooks rated with high quality published in the Deep Noise Suppression Challenge [21], 18 h from VCTK [22], and 5 h from DAPS [23], respectively.

We consider a 7-channel microphone array with 6 microphones on a circle of 4 cm radius and one center microphone. The center mic $m=0$ is defined as reference microphone. For training, the array geometry is simulated as free-field omnidirectional microphones, while for validation and testing we use measured RIRs using an actual device. For training, we simulate RIR sets of random positions in 1000 differently sized rooms using the image method [24], while for validation and testing we use measured RIRs using the actual device in 6 different rooms. The rooms were office, meeting, living rooms, and a large entrance hall with reverberation times between 0.3 to 1.3 s. In each room, several source positions were measured at challenging distances between 2 to 10 m, resulting in direct-to-reverberation ratios (DRRs) between -10 to 0 dB.

To generate spatially realistic noise, we use an internal database of recordings made with a third-order Ambisonics array, which were rendered to the 7-channel microphone array using a full spherical set of DRTFs from all source positions to the microphone positions. For training, we use a subset of 39 h of the noise, rendered accordingly to DRTFs simulated with the image method, while for validation and testing data we use two unseen 2.5 h noise subsets, which were rendered to measured DRTFs of the actual device in an anechoic chamber.

4.2. Data generation

For data generation we use the same processing pipeline for all datasets: We create mixtures of two overlapping speech signals of 30 s length. Each source signal is generated by concatenation of recordings from the same speaker, convolved with a RIR from a randomly chosen position in the same room. The reverberant multichannel speech signals are mixed with energy ratios drawn from a normal distribution with 0 dB mean and 1 dB variance. Noise is added to the mixtures with SNRs drawn from a normal distribution with mean and variance of 8 dB and 10 dB, respectively. The microphone signals are then generated by scaling the reverberant-noisy mixtures with levels from a normal distribution with mean and variance of

-28 dB and 10 dB. The target speech signals are generated using windowed versions of reverberant RIRs, enforcing a maximum reverberation time of 200 ms, preserving the early reflections which are beneficial for speech intelligibility and naturalness. These target speech signals are also scaled jointly with the microphone signals. Using this pipeline, we generate training, validation and test sets of 1000 h, 2.5 h and 4 h, respectively. All datasets were generated at a sampling rate of $f_s = 16$ kHz.

5. NETWORK TRAINING AND STFT SETUP

All DOA estimation and beamforming layers were implemented using a weighted overlap-add framework with an STFT frame length of 512 samples, an overlap of 50% between successive frames, a Hann window and an FFT size $N_{\text{FFT}} = 512$.

All networks were trained using Adam optimizer [25]. The initial learning rate was set to an appropriate value and then decreased by a factor of 2 if the SDR validation loss does not improve for 2 consecutive epochs. In each epoch, 500 batches of 10 audio sequences were randomly selected from the training set. Each sequence was a random 10 s sub-portion from the 30 s signals. To improve the network generalization, we use gradient clipping technique with a maximum L_2 norm of 5, similarly as used in [4].

6. EXPERIMENTAL RESULTS

In this section, we evaluate the speech separation performance of the proposed networks in terms of the scale-invariant SDR and SIR of BSSEval [26] and PESQ [27]. In Section 6.1, we investigate the impact of loss functions and network structures on the performance of the masking-based networks. In Section 5.2, we investigate the impact of using post masking on the performance of DBnets.

6.1. Loss functions and source separation performance

Figure 4 depicts the SDR improvement of the masking-based networks either using the recurrent structure (Mnet(R)) or the convolutional-recurrent structure (Mnet(CR)) for all considered loss functions. We observe for all loss functions that the lowest SDR improvement is obtained when training the networks based on only the magnitude distance, i.e., $\alpha = 0$, and the largest SDR improvement is obtained when training based on the complex spectral distance, i.e., $\alpha = 1$. In addition, training the networks based on the cMSE loss function yields the highest SDR performance, showing the importance of dynamic range compression for network training. Therefore, from now on we focus only on the results obtained based on the cMSE loss function. In Table 2 the SDR, the SIR and the PESQ improvement of the masking-based networks are compared. We observe that the Mnet(R) tends to degrade the SIR improvement, while the Mnet(CR) yields a significant improvement in terms of all considered performance measures, showing the effectiveness of convolutional-recurrent structure for masking based source separation.

6.2. DBnet source separation performance

In Table 3 the source separation performance of DBnets and their extensions are compared. We observe that the DBnets either using the recurrent structure (DBnet(R)) or the convolutional-recurrent structure (DBnet(CR)) provide an SDR and an SIR improvement, however the SDR improvement is lower than the Mnets (Table 2). The lower SDR improvement of DBnets can be mainly attributed to the limited suppression capability of beamformers, resulting in output signals with residual noise. Nevertheless, both DBnet(R) and DBnet(CR) are still able to yield a larger SIR improvement of about $0.82-1.26$ dB compared to the Mnet(R) (Table 2). The SIR improvement implies that DBnet(R) and DBnet(CR) are able to estimate the DOAs of the sources to appropriately steer beamformers, although both networks were trained without using the ground-truth DOAs of sources. This confirms the possibility of training beamforming networks without having ground-truth DOAs for source separation.

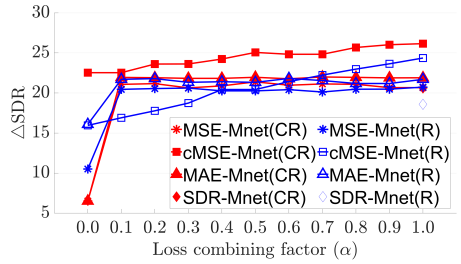


Fig. 4. SDR improvement for different loss functions when using Mnet(R) and Mnet(CR)

Table 2. Comparison of Mnet with recurrent and convolutional-recurrent structures

Method	△SDR	△SIR	△PESQ
Mnet(R)	24.36	−0.21	0.14
Mnet(CR)	**26.15**	2.54	0.20

Table 3. Comparison of DBnet and DBnet extensions with recurrent and convolutional-recurrent structures

Method	△SDR	△SIR	△PESQ
DBnet(R)	4.26	1.25	0.00
DBnet(CR)	4.29	1.26	0.00
DBnet(CR)-pMnet(R)	25.22	3.02	0.14
DBnet(CR)-pMnet(CR)	24.11	**6.86**	**0.21**
DBnet(R)-pMnet(R)	22.31	1.34	0.09
DBnet(R)-pMnet(CR)	23.53	0.29	0.03

When using the extensions of DBnets using post masking, it can be observed that all considered performance measures are significantly improved compared to the DBnets, showing the importance of post masking. When using the convolutional-recurrent DBnets followed by post masking, a considerably larger SIR improvement of 3.02 dB for DBnet(CR)-Mnet(R) and 6.86 dB for DBnet(CR)-Mnet(CR) is obtained compared to the Mnet(CR). However, when using the recurrent DBnets followed by post masking (DBnet(R)-Mnet(R) and DBnet(R)-Mnet(CR)), the improvement for all measures decreases. In general, among all considered networks the Mnet(CR) yields the highest SDR improvement but low SIR improvement and DBnet(CR)-Mnet(CR) yields the highest SIR and PESQ improvement with a considerably high SDR improvement.

7. CONCLUSION

In this paper, we proposed end-to-end source separation networks combining DOA estimation, beamforming and post masking. For DOA estimation and post masking we used recurrent and convolutional-recurrent network structures. We showed the superiority of the compressed spectral loss for source separation, and showed that this solely signal-based loss is enough to train DOA driven beamformers, without training with ground truth DOAs. Experiments in extremely challenging and realistic acoustic conditions with source distances up to 10 m in heavily reverberant and noisy environments showed that DBnet using the convolutional-recurrent structure for both DOA estimation and post masking is able to improve the SDR, SIR, and PESQ. However, further work is required to improve the separation performance even more in such adverse acoustic conditions.

8. REFERENCES

[1] M. Kolbæk, D. Yu, Z. Tan, and J. Jensen, "Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1901–1913, Oct. 2017.

[2] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 246–250.

[3] K. Kinoshita, L. Drude, M. Delcroix, and T. Nakatani, "Listening to each speaker one by one with recurrent selective hearing networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5064–5068.

[4] M. Maciejewski, G. Wichern, E. McQuinn, and J. L. Roux, "Whamr!: Noisy and reverberant single-channel speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 696–700.

[5] Y. Luo and N. Mesgarani, "Tasnet: Time-domain audio separation network for real-time, single-channel speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 696–700.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[7] Z. Shi, H. Lin, L. Liu, R. Liu, S. Hayakawa, S. Harada, and J. Han, "FurcaNet: An end-to-end deep gated convolutional, long short-term memory, deep neural networks for single channel speech separation," arXiv preprint, 2019.

[8] L. Drude and R. Haeb-Umbach, "Integration of neural networks and probabilistic spatial models for acoustic blind source separation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 815–826, Aug 2019.

[9] T. Yoshioka, Z. Chen, C. Liu, X. Xiao, H. Erdogan, and D. Dimitriadis, "Low-latency speaker-independent continuous speech separation," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 6980–6984.

[10] T. Ochiai, M. Delcroix, R. Ikeshita, K. Kinoshita, T. Nakatani, and S. Araki, "Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 6384–6388.

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

[12] S. Chakrabarty and E. A. P. Habets, "Multi-speaker doa estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 1, pp. 8–21, Feb. 2019.

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[14] B. D. van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, Apr. 1988.

[15] M. Brandstein and D. Ward, *Microphone Arrays Signal Processing Techniques and Applications*, Springer-Verlag Berlin Heidelberg, 2001.

[16] K. Tan, X. Zhang, and D. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 5751–5755.

[17] S. Braun and I. Tashev, "A consolidated view of loss functions for supervised deep learning-based speech enhancement," arXiv preprint, 2020.

[18] J. Lee, J. Skoglund, T. Shabestary, and H. Kang, "Phase-sensitive joint learning algorithms for deep learning-based speech enhancement," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1276–1280, 2018.

[19] K. Wilson, M. Chinen, J. Thorpe, B. Patton, J. Hershey, R. A. Saurous, J. Skoglund, and R. F. Lyon, "Exploring tradeoffs in models for low-latency speech enhancement," in *Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2018, pp. 366–370.

[20] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR – half-baked or well done?," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[21] "IEEE ICASSP 2021 Deep Noise Suppression (DNS) Challenge," `https://github.com/microsoft/DNS-Challenge`.

[22] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92), [sound]," in *University of Edinburgh. The Centre for Speech Technology Research*, 2019.

[23] "Device and produced speech (DAPS) dataset," `https://ccrma.stanford.edu/~gautham/Site/daps.html`.

[24] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, Apr. 1979.

[25] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint, 2014.

[26] F. R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK*, 2018, pp. 293–305.

[27] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, vol. 2, pp. 749–752.