2019 27th European Signal Processing Conference (EUSIPCO)

Multi-Microphone Speaker Separation based on Deep DOA Estimation

Shlomo E. Chazan, Hodaya Hammer, Gershon Hazan, Jacob Goldberger and Sharon Gannot
Faculty of Engineering, Bar-Ilang University, Ramal-Gan, 5290002, Israel.
Shlomi.Chazan@biu.ac.il

Abstract—**In this paper, we present a multi-microphone speech separation algorithm based on masking inferred from the speakers direction of arrival (DOA). According to the W-disjoint orthogonality property of speech signals, each time-frequency (TF) bin is dominated by a single speaker. This TF bin can therefore be associated with a single DOA. In our procedure, we apply a deep neural network (DNN) with a U-net architecture to infer the DOA of each TF bin from a concatenated set of the spectra of the microphone signals. Separation is obtained by multiplying the reference microphone by the masks associated with the different DOAs. Our proposed deep direction estimation for speech separation (DDESS) method is inspired by the recent advances in deep clustering methods. Unlike already established methods that apply the clustering in a latent embedded space, in our approach the embedding is closely associated with the spatial information, as manifested by the different speakers' directions of arrival.**

I. Introduction

Audio and speech source separation is an active research field for the past two decades. A comprehensive survey of single- and multi-microphone approaches can be found in [1]–[3] and will hence not be explored here. We rather focus on learning-based approaches, most notably those using DNN.

Most single microphone approaches are utilizing masking operation. In a nutshell, masking involves clustering of TF bins to the various speakers in the scene, and a multiplication of the noisy spectrogram by '1' in TF bins clustered to the desired speaker, and '0' otherwise. The underline assumption of these masking algorithm is the W-disjoint orthogonality principle introduced in [4], [5], stating that each TF bin is *dominated* by a single speaker, at least if the number of speakers is small enough.

Recently, deep clustering approach was introduced for single-microphone speaker separation [6], [7]. In this approach, an embedding from the high-dimensional short-time Fourier transform (STFT) representation of the speech to a low-dimensional latent space was first inferred, followed by a clustering operation in the latent space. Another approach, which uses permutation invariant training (PIT) was presented in [8]. Both these approaches had a dramatic impact on the single-microphone speech separation field. Yet, as they only exploit spectral information, their performance deteriorates in the presence of high reverberation, or when the speakers are characterized by similar spectral patterns. In many cases the outcome of these algorithms is characterized by *musical-noise* artifacts.

Spatial information, namely the attenuation and the time-delay between each of the sources' positions and a microphone pair, were utilized to estimate the separation mask in the degenerate unmixing estimation technique (DUET) approach [9]. Other multichannel separation algorithms are utilizing the single-channel deep clustering approach for estimating the building-blocks of the beamformer, specifically its steering vector [10], [11]. These approaches combine the advantages of the TF clustering operation, with the low distortion characteristics of the linear spatial processing that substitutes the masking operation. Other works train DNNs in order to estimate spectral masks. In [12] a DNN is applied to spatial features to infer a DOA-based mask, which is then used as a post-filtering stage at the output of a delay-and-sum beamformer. In [13] a group of DNNs, each applied in a different frequency band, is trained to predict a mask from spatial features. This information is then aggregated to generate a soft mask which is used for the final speech separation. In [14] an *unsupervised* deep clustering approach was applied to multiple mixtures of sources in a training stage. The trained DNN was then applied to the test mixture to predict the separating masks. In [15], a single-channel deep clustering network was trained in a supervised manner, where the supervision was obtained by a multichannel segmentation network.

Other approaches combining DNNs and beamforming are presented in [16], [17]. In these methods, a concurrent speakers detector (CSD) is implemented to distinguish between noise-only frames, single-speaker frames and concurrently active speakers frames. In the first two classes the noise spatial correlation matrix and the steering vectors are estimated, respectively. In the third class, the beamformer weights are not updated.

The deep clustering framework was extended to the multichannel setting in [18]. Spatial information was augmented with the spectral cues to form an input feature to the bidirectional long short-term memory (BLSTM) deep clustering network. The separation in this approach is still applied by single-channel masking using the clustering in the embedded latent domain.

In the current contribution, we are presenting a U-net architecture to address the speech separation task. It is assumed that the speakers are in different DOAs in the room. Consequently, rather than inferring a latent embedded domain, we utilize the DOA as the supervision of our network. Motivated by the great success of the U-net architecture in the computer vision

978-9-0827-9703-9/19/$31.00 ©2019 IEEE

field [19], and the high performance of the convolutional neural networks (CNNs) in estimation the DOA of multi speakers in noisy and reverberant environments [20], we train a U-net to classify each TF bin of the multichannel STFT image to one of the DOA candidates. The performance of the proposed schemes is demonstrated using recorded acoustic channels, while training is carried out using simulated data.

II. Deep Speech Separation

A. The separation algorithm

Consider an array of M microphones capturing a mixture of N speech sources in a reverberant enclosure. The i-th speech signal $s^i(t)$ propagates through the acoustic channel before being captured by the mth microphone:

$$z_m(t) = \sum_{i=1}^{N} s^i(t) * h_m^i(t), \qquad m \in \{1, \ldots, M\} \tag{1}$$

where, h_m^i is the room impulse response (RIR) relating the ith speaker and the mth microphone. In the STFT domain, (1) can be rewritten as:

$$z_m(l,k) = \sum_{i=1}^{N} s^i(l,k) h_m^i(l,k), \tag{2}$$

where l and k, are the time-frame and the frequency-bin (TF) indexes, respectively.

Following the W-disjoint orthogonality assumption [4], each TF bin is dominated by a single speaker. We assume that each speaker is located at a different DOA and therefore each bin is dominated by a single DOA. The crux of our speech separation method is to estimate the DOA for each TF bin by a neural network and then separate the speakers by grouping these bins according to their estimated DOA.

The main building block of the algorithm is a neural network that uses the microphone signals to infer the DOA at each TF bin of a given time-frequency image. The network input is a $L \times K$ time-frequency "image" where L is the number of time frames and K is the number of frequency bins. We have chosen to substitute the raw microphone signals with the phase of the instantaneous relative transfer function (RTF) estimate, calculated as the phase of the bin-wise ratio between the mth microphone signal and the reference microphone signal. The phase angle is encoded as a point in the unit circle. The input features to the network, therefore, is an $L \times K$ matrix $\mathcal{R}$ where each (l,k) entry has M channels each correspond to a microphone:

$$r(l,k,m) = (\cos(\angle \frac{z_m(l,k)}{z_{\text{ref}}(l,k)}), \sin(\angle \frac{z_m(l,k)}{z_{\text{ref}}(l,k)})). \tag{3}$$

Due to the W-disjoint assumption, the normalized features $r(l,k,m)$ are dominated by a single speaker and hence correspond to a specific DOA. Ideally, the speech contribution to $r(l,k,m)$ is negligible. Hence, it is expected that these are better features than the raw data for DOA estimation.

We form the DOA estimation as a classification task by discretizing the possible angles to be in the set $\Theta = \{0°, 15°, 30° \ldots, 180°\}$. Let $y_{l,k}$ be a random variable indicating the active direction at bin (l,k). The target of the network is to infer the conditional distribution of the discrete set of candidate DOAs in Θ for each TF bin, given the recorded signal:

$$p_{l,k}(\theta) = p(y_{l,k} = \theta | \mathcal{R}), \quad \theta \in \Theta. \tag{4}$$

where $\mathcal{R}$ is an $L \times K$ matrix of all the TF bins. The image-to-image DOA prediction task in (4) is implemented by a U-net, which details are given in the next section.

Next, the direction-dependent power is calculated by the instantaneous power of the reference microphone, weighted by the U-net output:

$$E(\theta) = \sum_{l,k} p_{l,k}(\theta) \cdot |z_{\text{ref}}(l,k)|^2, \qquad \theta \in \Theta. \tag{5}$$

Note that the total power is satisfying the following equation:

$$E = \sum_{\theta \in \Theta} E(\theta) = \sum_{l,k} |z_{\text{ref}}(l,k)|^2. \tag{6}$$

High power from a specific direction is an indication for an active speaker at this direction. To find all directions of the active speakers in the scene, we sort the powers according to their power level:

$$E(\theta_1) \geq E(\theta_2) \geq E(\theta_3) \ ...$$

where θ_1 corresponds to the direction with the highest power, θ_2 the second highest, etc. The speakers' directions are then determined by the N DOAs with the highest power level. If the number of speakers N is not known in advance, we can set N as the minimal value such that $\sum_{i=1}^{N} E(\theta_i) > \alpha E$, with α is a predefined threshold.

The next step is to use the estimated DOA to form a *mask* for each detected speaker in the scene. The estimated mask of the ith speaker is the U-net output:

$$\hat{M}_i(l,k) = p_{l,k}(\theta_i) \tag{7}$$

and the absolute value of the ith speaker signal is reconstructed as follows:

$$|\hat{s}^i(l,k)| = |z_{\text{ref}}(l,k)| \cdot \hat{M}_i(l,k). \tag{8}$$

The noisy phase is then used to reconstruct the separated signals in the time-domain, by the application of the inverse STFT. We dub the proposed algorithm deep direction estimation for speech separation (DDESS).

Note, that if a static acoustic scene can be assumed, namely that the sources do not significantly change their DOA during the entire utterance, permutation problems, which are typical to clustering-based approaches [7], are circumvented.

Note that estimating the DOA is modeled here as a classification problem and not as a regression task. We are not interested in finding the exact DOAs of the speakers in the scenario but rather, grouping them into distinct directions. That is, even with inaccurate DOA estimate, the speech separation can still work, provided that most TF bins are clustered to a mutually exclusive classes.

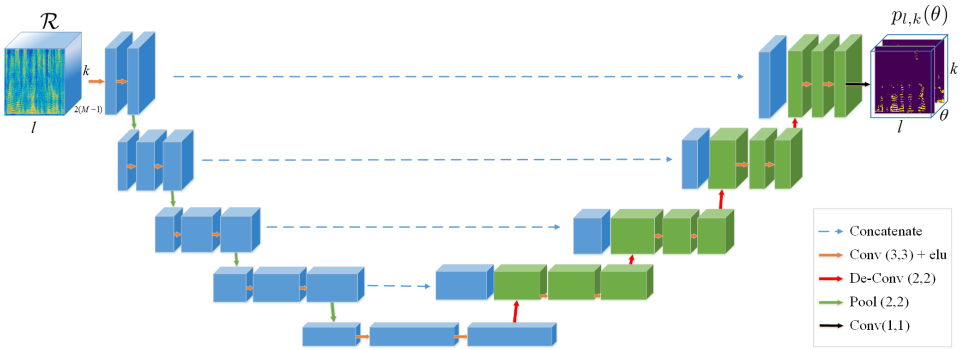


Fig. 1: U-net architecture for DOA-mask speech separation. The blue blocks depict the encoder and the green blocks depict the decoder.

B. The U-net for DOA estimation

The input to the network is the feature matrix $\mathcal{R}$. The overlap between successive STFT frames is set to 75 %. Hence, to improve the estimation accuracy of the RTFs, we have used an average of three consecutive frames both in the numerator and denominator of (3).

In our U-net architecture, the input shape is $(L, K, 2(M-1))$ where, $K = 256$ is the number of frequency bins, $L = 96$ is the number of frames, and M is the number of microphones. The output shape is $(L, K, |\Theta|)$ where $|\Theta|$ is the cardinality of the set Θ.

The U-net architecture is presented in Fig. 1. The blue boxes depict the encoder and the green boxes the decoder. In this architecture, in the encoder part, the input image is squeezed into a bottleneck using 2×2 max pooling operations (downsample), and then in the encoder part it is upsampled back to the original image shape. The main problem with this architecture is that during the pooling operation important local information is lost. To tackle this problem, a U-shape architecture was developed in [19]. The U-net connects between mirrored layers in the encoder and decoder by passing the information without going through the bottleneck and thus, alleviating the information loss problem.

Let $\text{CE}_{l,s}$ denote a 2D convolution layer with 'elu' as the activation function, where l is the number of filters and $s \times s$ is the filter size. Similarly, let $\text{DE}_{l,s}$ is the de-convolution 'elu' layer. Finally, let P_s denote the max-pooling operation with filter size $s \times s$.

The encoder down-sampling path is given by:
$\text{CE}_{16,3} \rightarrow \text{CE}_{16,3} \rightarrow \text{P}_2 \rightarrow \text{CE}_{32,3} \rightarrow \text{CE}_{32,3} \rightarrow \text{P}_2 \rightarrow \text{CE}_{64,3} \rightarrow \text{CE}_{64,3} \rightarrow \text{P}_2 \rightarrow \text{CE}_{128,3} \rightarrow \text{CE}_{128,3} \rightarrow \text{P}_2 \rightarrow \text{CE}_{256,3} \rightarrow \text{CE}_{256,3}$.

The decoder up-sampling path is given by:
$\text{DE}_{128,3} \rightarrow \text{CE}_{128,3} \rightarrow \text{CE}_{128,3} \rightarrow \text{DE}_{64,3} \rightarrow \text{CE}_{64,3} \rightarrow \text{CE}_{32,3} \rightarrow \text{DE}_{32,3} \rightarrow \text{CE}_{32,3} \rightarrow \text{CE}_{32,3} \rightarrow \text{DE}_{16,3} \rightarrow \text{CE}_{16,3} \rightarrow \text{CE}_{16,3} \rightarrow \text{CE}_{13,1}$.

The output DOA distribution is finally obtained by a softmax layer. To overcome the problem of overfitting, we add dropout layers [21] after every $\text{CE}_{l,s}$ layer. Additionally, the raw data input is normalized to zero mean and unit variance.

To train the network we use a simulated data where both the location and a clean recording of each speaker are given. We can thus easily find for each TF bin (l, k) the dominant speaker and the corresponding DOA $y_{k,l} \in \Theta$. The network is trained to minimize the cross entropy between the correct and the estimated DOA. The cross entropy cost function is summed over all the images in the training set. The network was implemented with Tensor-Flow and training was done using the ADAM optimizer [22]. The number of epochs was set to be 100, and the training stopped after validation loss was going up for 3 successive epochs. The minibatch size was 64 images.

III. Experimental Study

In this section we evaluate the proposed DDESS algorithm and compare its performance to the DUET algorithm [9].

A. Training database

To generate the training data we used the RIR generator[1] efficiently implementing the image method [23]. We simulated an eight microphone array with $(3, 3, 3, 8, 3, 3, 3)$ cm between microphones. Similar microphone inter-distance was used in the test phase. The dimensions of the room are $6 \times 6 \times 2.4$ (width, length and height), similar to the acoustic lab used in the test phase. The microphone array was positioned at $(3, 1, 1.5)$ m.

For each scenario two clean signals from the wall street journal (WSJ) database [24] were randomly selected and two different DOAs were also randomly selected from the possible values in the range $\Theta = \{0, 15, \ldots, 180\}$. The speakers were located in a radius of $r = 1.5m$ from the center of the microphone array. To increase the training diversity, the radius of the speakers was perturbed by a Gaussian noise with variance 0.3 m. The DOA of each speaker was computed with respect to the center of the array. We used $\text{T}_{60} \in \{0.2, 0.3, 0.4\}$sec. Once the scenario is set, the RIRs were generated, and the clean signals were separately convolved with them. Finally, we added the signals with signal to interference ratio (SIR) randomly chosen in the range $\text{SIR} \in [-2, 2]$. Sampling rate

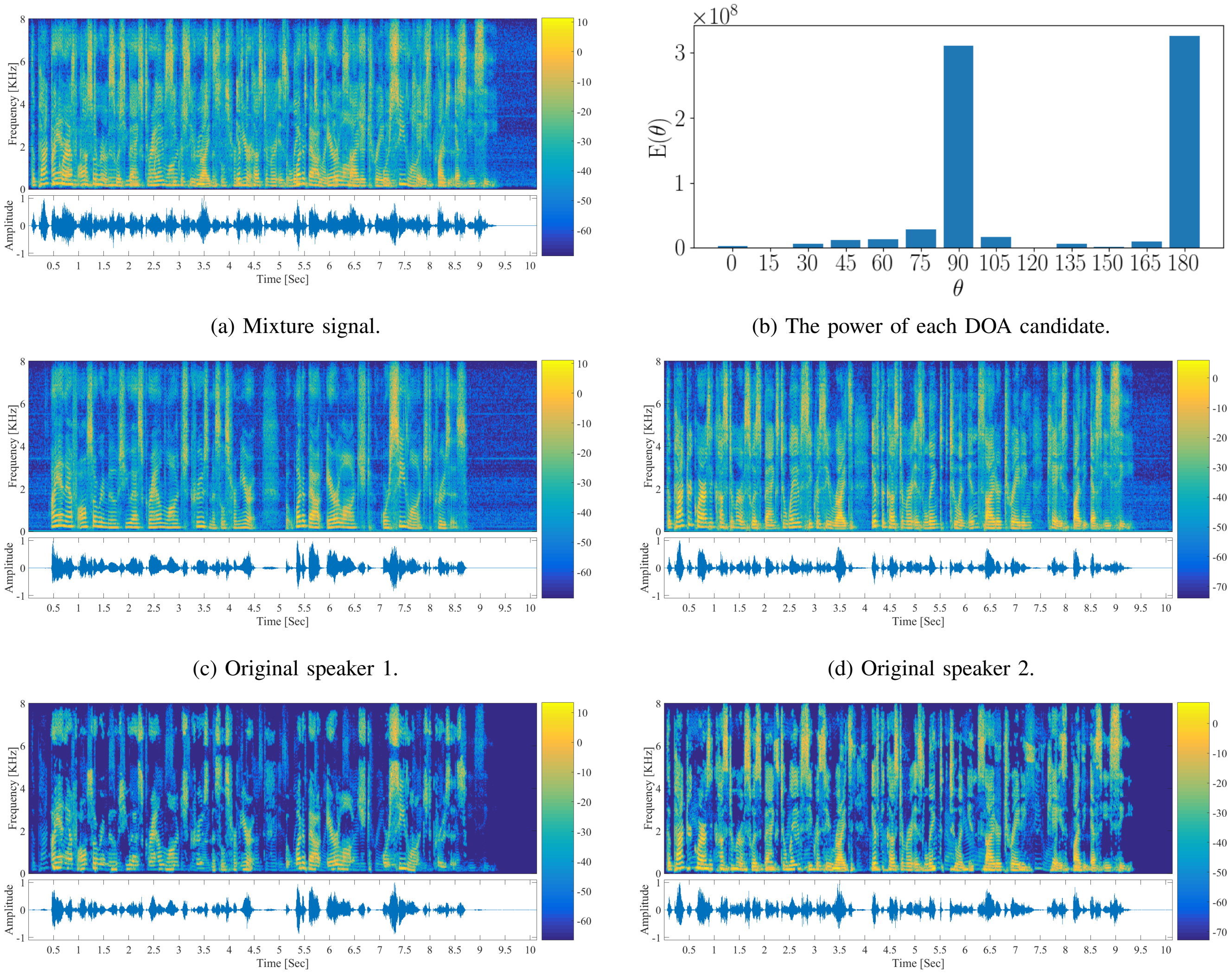
[1] Available online at `github.com/ehabets/RIR-Generator`

TABLE I: SDR and SIR results with two T_{60} and distance 1m.

	$T_{60} = 160$				$T_{60} = 360$			
	SDR		SIR		SDR		SIR	
Speaker	1	2	1	2	1	2	1	2
Noisy	-1.05	-1.41	0.23	-0.11	-0.91	-1.75	0.5	-0.41
DUET	1.3	0.7	4.24	3.38	0.87	-0.33	3.59	2.24
DDESS	**2.26**	**1.95**	**12.6**	**12.43**	**1.68**	**1.69**	**13.06**	**12.76**

TABLE II: SDR and SIR results with two T_{60} and distance 2m.

	$T_{60} = 160$				$T_{60} = 360$			
	SDR		SIR		SDR		SIR	
Speaker	1	2	1	2	1	2	1	2
Noisy	-1.22	-1.49	0.19	-0.07	-2.07	-1.07	-0.5	0.68
DUET	-0.31	-0.26	2.24	2.41	-1.79	-0.1	1.04	2.44
DDESS	**1.38**	**1.31**	**11.46**	**11.44**	**0.08**	**1.02**	**11.1**	**11.68**



(a) Mixture signal.

(b) The power of each DOA candidate.

(c) Original speaker 1.

(d) Original speaker 2.

(e) Estimated speaker 1.

(f) Estimated speaker 2.

Fig. 2: An example of the separation results of the DDESS algorithm.

was set to 16KHz and the frame length of the STFT was set to $K = 512$, with overlap of 75% between two successive frames. The training set comprises two hours of recordings with 6000 different scenarios of mixtures of two speakers.

B. Separation Results

For each test scenario, we selected two speakers (male or female) from the test set of the TIMIT database, placed them in two different angles between 0° to 180° relative to the microphone array, at the distance of either 1 m or 2 m.

Each clean speech signal was convolved with a real RIR, drawn from the multichannel impulse response database recorded in our acoustic lab [25] (similar room dimensions and microphone inter-distances to the simulated scenarios), and then mixed the with SIR=0 dB. We used $T_{60} = 160/360$ ms for the room reverberation. Overall, in the test dataset we had 30 different scenarios for each T_{60}, and the results are the averaged over all scenarios.

We used a standard blind source separation (BSS) evaluation toolbox [26] to test the separation capabilities of the DDESS algorithm and the DUET algorithm [9]. Tables I and II present the SIR and signal to distortion ratio (SDR) results for the two source distances, 1m and 2m, respectively. It is evident that the DDESS algorithm outperforms the DUET in all experiments.

Fig. 2 depicts the spectrogram of the noisy input, the clean signals and the estimates obtained by the proposed algorithm for two equi-power speakers positioned at 90° and 180° and $r = 2$ m and for $\mathrm{T}_{60} = 160$ ms. It is evident that the DDESS separates the signals. Fig. 2b depicts the power level for DOA candidates. It is clear that the DOAs were accurately classified. Sound samples can be found in the lab website.[2]

IV. CONCLUSIONS

In this study, we presented a speech separation algorithm, based on DOA classification and masking. A DNN with a U-net architecture is trained to classify TF bins to DOAs. The association of each TF bin to specific DOA is used to construct spectral masks, which when applied to the spectrogram of the reference microphone obtain spectral source separation. The U-net was trained in a simulated room and tested with real RIR recordings, demonstrating the proposed algorithm capabilities in the task of blindly separating the sources. In the future we plan to increase the robustness of the proposed algorithm to mismatch between train and test conditions. Another possible extension is to address dynamic scenarios and to provide a trajectory estimate for the speakers.

REFERENCES

[1] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 25, no. 4, pp. 692–730, 2017.

[2] E. Vincent, T. Virtanen, and S. Gannot, Eds., *Audio Source Separation and Speech Enhancement*. Wiley, Sep. 2018.

[3] S. Makino, *Audio Source Separation*. Springer, 2018.

[4] S. Rickard and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.

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

[7] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[8] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[9] S. Rickard, *Blind speech separation*. Springer, 2007, vol. 615, ch. The DUET blind source separation algorithm, pp. 217–241.

[10] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolkova, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *Proc. Interspeech*, 2017.

[11] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for BSS with deep clustering embeddings," in *Proc. Interspeech*, 2017.

[12] P. Pertilä and J. Nikunen, "Distant speech separation using predicted time–frequency masks from spatial features," *Speech Communication*, vol. 68, pp. 97–106, 2015.

[13] Y. Yu, W. Wang, and P. Han, "Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2016, no. 1, p. 7, 2016.

[14] E. Tzinis, S. Venkataramani, and P. Smaragdis, "Unsupervised deep clustering for source separation: Direct learning from mixtures using spatial information," *arXiv preprint arXiv:1811.01531*, 2018.

[15] P. Seetharaman, G. Wichern, J. L. Roux, and B. Pardo, "Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures," *arXiv preprint arXiv:1811.02130*, 2018.

[16] S. E. Chazan, J. Goldberger, and S. Gannot, "DNN-based concurrent speakers detector and its application to speaker extraction with LCMV beamforming," in *IEEE International Conference on Audio and Acoustic Signal Processing (ICASSP)*, 2018.

[17] S. E. Chazan, S. Gannot, and J. Goldberger, "Attention-based neural network for joint diarization and speaker extraction," in *IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2018.

[18] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, "Multi-channel deep clustering: Discriminative spectral and spatial embeddings for speaker-independent speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[20] S. Chakrabarty and E. A. Habets, "Multi-speaker DOA estimation using deep convolutional networks trained with noise signals," *IEEE Journal of Selected Topics in Signal Processing*, 2019.

[21] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] 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, 1979.

[24] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91. Stroudsburg, PA, USA: Association for Computational Linguistics, 1992, pp. 357–362.

[25] E. Hadad, F. Heese, P. Vary, and S. Gannot, "Multichannel audio database in various acoustic environments," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[26] C. Févotte, R. Gribonval, and E. Vincent, "BSS_EVAL toolbox user guide–revision 2.0," online, 2005.

[2] www.eng.biu.ac.il/gannot/speech-enhancement/