MULTI-CHANNEL DEEP CLUSTERING: DISCRIMINATIVE SPECTRAL AND SPATIAL EMBEDDINGS FOR SPEAKER-INDEPENDENT SPEECH SEPARATION

Zhong-Qiu Wang[1,2], *Jonathan Le Roux*[1], *John R. Hershey*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), USA
[2]Department of Computer Science and Engineering, The Ohio State University, USA

ABSTRACT

The recently-proposed deep clustering algorithm represents a fundamental advance towards solving the cocktail party problem in the single-channel case. When multiple microphones are available, spatial information can be leveraged to differentiate signals from different directions. This study combines spectral and spatial features in a deep clustering framework so that the complementary spectral and spatial information can be simultaneously exploited to improve speech separation. We find that simply encoding inter-microphone phase patterns as additional input features during deep clustering provides a significant improvement in separation performance, even with random microphone array geometry. Experiments on a spatialized version of the wsj0-2mix dataset show the strong potential of the proposed algorithm for speech separation in reverberant environments.

Index Terms— deep clustering, spatial clustering, deep learning, cocktail party problem, speaker-independent speech separation

1. INTRODUCTION

Dramatic advances have been made in monaural speaker-independent multi-speaker speech separation since the introduction of the deep clustering algorithm [1, 2]. However, there has been little work extending this framework to the multi-channel setting. Deep clustering addresses the *cocktail party problem* by training a deep neural network (DNN) to project each time-frequency (T-F) unit to a high-dimensional embedding vector such that the embeddings for the T-F unit pairs dominated by the same speaker are close, while those for pairs dominated by different speakers are farther away from each other. This way, the speaker assignment of each T-F unit can be determined at run time by applying a simple clustering algorithm to the embeddings. Deep clustering was the first modern approach to perform speaker-independent separation, and demonstrated superior performance over previous attempts at speech separation, including graphical modeling approaches [3], spectral clustering approaches [4], and computational auditory scene analysis based methods [5].

When multiple microphones are available, the directional information associated with each source can be exploited for separation, as sound sources are often spatially separated in real-world environments. To utilize this information, conventional wisdom focuses on clustering the individual T-F units into different sources according to their spatial origins by assuming that each T-F unit is dominated by only one source across all the microphone channels [6–9]. Interchannel time/phase differences (ITDs/IPDs), interchannel level differences (ILDs), and directional statistics [8] are the most commonly used spatial cues for spatial clustering [5]. However, when the sound sources are spatially close, when room reverberation is present, or when the sound sources are moving, the ITDs, ILDs, and directional statistics are typically not good enough to achieve sufficient source separation. In such cases, spectral information can complement the insufficient spatial information, as sound sources such as speech exhibit characteristic spectral patterns that can be learned, as demonstrated by single-channel deep clustering [1, 2].

Here we propose to improve deep clustering by incorporating spatial information into the input features, along with the usual spectral information, in order to provide a stronger set of separation cues. Other recent studies have combined spatial information with deep clustering in a different way. The systems by [10] and [11] use the outputs of single-channel deep clustering to derive a beamformer for each source. Their beamforming approaches follow the success of T-F masking and deep learning based beamforming for speech enhancement [12] in the recent CHiME-3 and CHiME-4 challenges [13], where the estimated single-channel masks are combined to compute speech or noise covariance matrices. However, the power of their algorithms is largely limited by beamforming, which is only a linear spatial filter per frequency. Many factors can reduce the effectiveness of such beamforming: room reverberation, moving sources, diffuse noise, and conditions with more sources than microphones, all can significantly degrade the resulting separation. In addition, when sound sources arrive from the same general direction, beamforming may fail to resolve them. In contrast, by jointly training on spectral and spatial features, our approach can learn to balance the two types of information. The estimated embeddings and masks from the proposed algorithm may also serve as a better initialization for the T-F masking based beamforming approaches. However, in our experiments the joint deep clustering was often able to outperform the mask-based beamforming approach even when using the ideal oracle masks to obtain the beamforming parameters.

There are previous studies using spatial features for DNN training [14–17] for speech enhancement tasks (i.e., only one target speaker with background noise) and they typically assume that the target speaker is in a fixed direction, typically the front direction in the binaural setup. However, in the more general setting we pursue, the target speakers may be in any combination of positions. To the best of our knowledge, this study is the first attempt at applying spatial features to train deep networks for speaker separation. For evaluation, we create a spatialized and reverberant version of the wsj0-2mix dataset [1]. Our method is shown to give much better separation performance in the stereo case in terms of signal-to-distortion ratio (SDR) than the oracle multi-channel Wiener filter (MCWF) [9], model-based EM source separation and localization (MESSL) [18], and GCC-NMF [19] algorithms.

This work was done while Z.-Q. Wang was an intern at MERL.

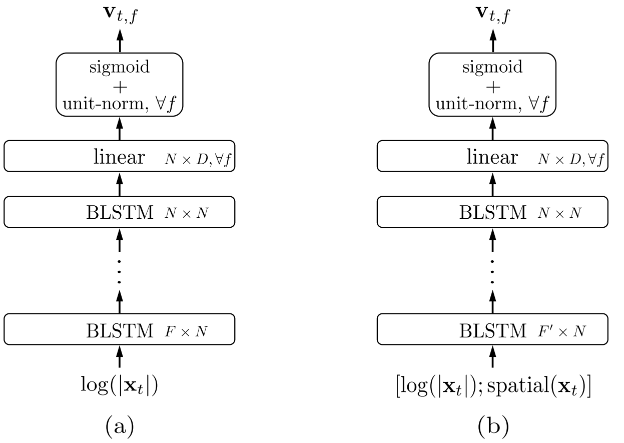
$\mathbf{v}_{t,f}$
sigmoid + unit-norm, $\forall f$
linear $N \times D, \forall f$
BLSTM $N \times N$
BLSTM $F \times N$
$\log(|\mathbf{x}_t|)$
(a)
$\mathbf{v}_{t,f}$
sigmoid + unit-norm, $\forall f$
linear $N \times D, \forall f$
BLSTM $N \times N$
BLSTM $F' \times N$
$[\log(|\mathbf{x}_t|); \mathrm{spatial}(\mathbf{x}_t)]$
(b)

Fig. 1. Network architecture of (a) single-channel deep clustering, (b) multi-channel deep clustering.

2. SYSTEM DESCRIPTION

2.1. Single-Channel Deep Clustering

The key idea of deep clustering [1, 2] is to learn a high-dimensional embedding for each T-F unit using a neural network such that the embeddings for the T-F unit pairs dominated by the same speaker are close while farther away otherwise. At test time, the speaker assignment of each T-F unit can then be simply determined using a clustering algorithm, such as k-means, on the learned embeddings. More specifically, let $y_i \in \mathbb{R}^{1\times C}$ be a one-hot label vector denoting which of the C sources in a mixture dominates the i-th T-F unit. The network learns to produce a D-dimensional unit-length vector v_i such that the affinity matrix computed from all the y_i can be approximated using the learned embeddings. Mathematically, the objective function optimized in [1, 2] is as follows:

$$\mathcal{L}_{\mathrm{DC}}(V, Y) = \|VV^{\mathsf{T}} - YY^{\mathsf{T}}\|_{\mathrm{F}}^2 \qquad (1)$$

where the embedding matrix $V \in \mathbb{R}^{TF\times D}$ and the label matrix $Y \in \mathbb{R}^{TF\times C}$ are respectively obtained by vertically stacking all the embedding vectors v_i and all the one hot vectors y_i in an utterance. The network architecture is shown in Fig. 1(a). This framework has produced remarkable improvements over conventional methods on single-channel speech separation tasks [1, 2].

Our recent study [20] found that further improvements could be obtained using an alternative cost function based on whitening the embeddings in a k-means objective:

$$\begin{aligned}\mathcal{L}_{\mathrm{DC,W}}(V, Y) &= \|V(V^{\mathsf{T}}V)^{-\frac{1}{2}} - Y(Y^{\mathsf{T}}Y)^{-1}Y^{\mathsf{T}}V(V^{\mathsf{T}}V)^{-\frac{1}{2}}\|_{\mathrm{F}}^2 \\ &= D - \operatorname{tr}\big((V^{\mathsf{T}}V)^{-1}V^{\mathsf{T}}Y(Y^{\mathsf{T}}Y)^{-1}Y^{\mathsf{T}}V\big). \qquad (2)\end{aligned}$$

One caveat to the labels, Y, is that for quiet T-F regions, it becomes arbitrary which source dominates, and the resulting noisy labels can provide an inconsistent training signal for deep clustering. In addition, these T-F units do not carry directional phase information regarding the underlying source directions. Following [1], we filter out the T-F bins where none of the sources is significantly active. In particular, a source is considered active in a T-F bin if its magnitude is within some threshold from its largest magnitude in the mixture: $w_i = \max_k \big[10\log_{10}(|s_{k,i}|^2/\max_j |s_{k,j}|^2) > \beta\big]$, where $[\cdot]$ is the Iverson bracket, $|s_{k,i}|$ represents the magnitude of the i-th T-F unit of the clean source k, and β is empirically set to -40 dB.

2.2. Two-Channel Deep Clustering

In the proposed method we encode not only spectral but also spatial information into the embedding of each T-F unit by including spatial features as additional inputs, as illustrated in Fig. 1(b). Since spectral and spatial features can be complementary in terms of their sources of uncertainty and failure modes, we expect their combination to show improved robustness relative to each type of feature in isolation. We consider two types of additional features, each inspired by one of the two popular approaches for conventional spatial clustering, the *narrowband approach* and the *wideband approach*.

The narrowband approach performs clustering within each frequency band using spatial cues such as IPDs or ILDs. The DUET algorithm [6, 7] assumes that the microphone pairs are placed sufficiently close to each other so that phase-wrapping effects can be neglected. It estimates the ITD of each T-F unit by directly dividing the phase difference by the angular frequency, and then performs clustering on the estimated ITDs and ILDs of all the T-F units. Unfortunately, with narrowly separated microphones, the ITDs could be too small to be useful for separation. Moreover room reverberation can substantially deteriorate the ITDs and ILDs. In contrast Sawada et al. [8] handles phase wrapping by clustering based ultimately on IPDs intead of ITDs. This results in an unknown permutation of the clusters across frequencies, which is handled by further clustering across frequency. Using the STFT vectors for spatial clustering has also been explored in acoustic beamforming [21–23]. However, those approaches only perform clustering based on spatial cues and there are no clear ways to combine them with spectral features. It would be beneficial to perform clustering across all the frequencies and over all the T-F units, as some frequency bands may be particularly worse than others and most importantly, more T-F units would form stronger cluster patterns that could elicit better clustering results.

Here, we use the following IPDs as additional features for model training:

$$\mathrm{cosIPD}(t, f, p, q) = \cos(\theta_{t,f,p,q}) \qquad (3)$$

$$\mathrm{sinIPD}(t, f, p, q) = \sin(\theta_{t,f,p,q}) \qquad (4)$$

where $\theta_{t,f,p,q} = \angle x_{t,f,p} - \angle x_{t,f,q}$ is the phase difference between the STFT coefficients $x_{t,f,p}$ and $x_{t,f,q}$ at time t and frequency f of the signals at microphones p and q. The rationale is that for spatially-separated sources with different time delays, $\frac{x_{t,f,p}}{x_{t,f,q}} = \frac{|x_{t,f,p}|}{|x_{t,f,q}|}e^{\theta_{t,f,p,q}}$ should naturally form clusters within each frequency band due to the speech sparsity property [6]. As the gains at different microphones are usually very similar under the far-field assumption, our study only uses the phase term for deep clustering, in the form of its real and imaginary parts. For a given source, the cosIPD and sinIPD features at different frequency bands are very different, so we combine them with spectral features that can help resolve the ambiguity. Note that the loss function is always computed from all the T-F units rather than independently within each frequency band.

The wideband approach avoids the IPD ambiguity by enumerating a set of potential time delays. The key insight [18] is that, given a time delay, the phase differences at all the frequency bands can be unambiguously determined in anechoic environments. The MESSL [18] algorithm therefore performs spatial clustering according to the time delays by checking whether the hypothesized time delay matches the observed phase differences at different frequencies. Motivated by [18] and the GCC-PHAT algorithm [24], we derive the following spatial feature for model training:

$$\mathrm{GCC}(t,f,p,q,\tau)=\cos\left(\theta_{t,f,p,q}-\frac{2\pi f}{N}\tau\right) \tag{5}$$

where N is the frame length, $\theta_{t,f,p,q}$ is the observed phase difference, τ is the hypothesized time delay in samples, and $\frac{2\pi f}{N}\tau$ is the hypothesized phase difference. The 2π-periodic cosine operation here can deal with potential phase wrapping effects.

The rationale behind this feature is that each of the underlying sources in a mixture could come from any direction. Our approach avoids a separate sound localization module [19, 25] by enumerating a set of potential time delays. When a hypothesized time delay matches the observed phase difference, it appears as a peak in the derived spatial feature. The entire set of GCC coefficients encodes all the direction information of each source and hence could be useful for deep clustering. Although this feature exhibits strong spatial aliasing effects in high-frequency bands, we hand it over to a neural network which may learn to deal with the spatial aliasing effects automatically.

The GCC features have a much higher dimension than the spectral features. If each dimension of these two features is normalized to unit variance, more importance will be implicitly placed on the GCC features. However, spectral features are also very important for deep clustering. Our system places equal importance on them by normalizing each dimension of the spatial features to have $1/K$ variance, where K is the number of the time delays of interest, and each dimension of the spectral features to have unit variance. This simple strategy leads to faster convergence and better performance compared with normalizing all the dimensions to unit variance in our experiments.

2.3. Multi-Channel Deep Clustering

We propose a simple yet effective algorithm to extend our system to arrays with more than two microphones. At run time, we first choose a reference microphone, and for each pair constituting of a reference microphone and a non-reference microphone, we get an embedding for each T-F unit using the two-channel deep clustering model. Then we stack the embeddings of all the pairs of reference and non-reference microphones at each T-F unit, and perform k-means clustering on the stacked embeddings. The resulting binary mask is applied to the reference microphone signal for separation. This way, our algorithm is readily applicable to microphone arrays with diverse microphone geometries and with any number of microphones.

3. EXPERIMENTAL SETUP

We use a room impulse response (RIR) generator[1] to spatialize the wsj0-2mix [1] dataset. This dataset has been widely used in many single-channel speech separation studies [1, 2, 26–28] since the debut of deep clustering. The training and validation sets are generated by mixing two randomly-selected utterances from two randomly-selected speakers in the WSJ0 training set after re-scaling them by randomly-selected coefficients such that the SNR of one source with respect to the other in the mixture is uniformly distributed between -5 dB and +5 dB. The test set is similarly generated using the speakers in the development and evaluation sets of the WSJ0 corpus. Note that the test speakers in wsj0-2mix are unseen in the training and validation set. There are 20,000 (∼30h), 5,000 (∼10h), and 3,000 (∼5h) utterances in the training, validation, and test set, respectively. The RIR generator uses the image method [29] to generate simulated room impulse responses.The general guideline is to make the setup as random as possible within realistic constraints. For each two-speaker mixture, we randomly generate a room with random room characteristics, speaker locations and microphone geometry. The aperture sizes are randomly sampled from 15 cm to 25 cm. The T60s are randomly drawn from 0.2 s to 0.6 s, as this range of T60s is most common in domestic environments [9]. All microphones are omni-directional. The average distance between speaker and array center is 1.3 m with 0.4 m standard deviation. The average direct-to-reverberant energy ratio is 2.5 dB with 3.8 dB standard deviation. The code to generate the data is available online[2].

We only use the signals at the first channel for model training and evaluation. During training, we extract the spectral feature from the first-channel signals and compute Y from the clean reverberant speech of each source captured at the first microphone. The spatial feature is extracted using the first-channel signal together with one of the other signals. For evaluation, we extract the spectral features from the first-channel signal and use the clean reverberant speech of each source at the first channel as the reference for SDR computation. The τ in Eq. (5) is enumerated from -6 to 6 samples[3] in steps of 0.25 samples. We found that this step size leads to a good enough spatial resolution (340/8,000*0.25=0.01 m). The dimension of the GCC feature at each T-F unit is therefore 49 (K=2*6/0.25+1). Although the dimension may sound high, it is still acceptable as, in deep clustering, the dimension D of the embeddings is typically set to 20 per T-F unit [1, 2], which is of the same order of magnitude as that of the GCC feature dimension.

Our BLSTM consists of four hidden layers, each with 600 units in each direction. It is trained from scratch on 400-frame segments for a maximum of 200 epochs using the Adam algorithm. The window size is 32 ms and the hop size is 8 ms. The sampling rate is 8 kHz. We perform 256-point FFT to get 129-dimensional log magnitude features for each frame for BLSTM training. At run time, k-means clustering is always performed on the entire utterance to get a binary mask for separation.

4. EVALUATION RESULTS

To check the validity and correctness of using the proposed features for deep clustering, we first evaluate them on the spatialized anechoic wsj0-2mix data. We emphasize that it is known that this anechoic setup using two microphones and two speakers is not challenging at all for beamforming algorithms, as they can achieve almost perfect separation [9]. Evaluating on this basic setup however can prove the correctness and show the potential of our algorithm. Note that this spatialized anechoic wsj0-2mix data is almost the same as the original wsj-2mix data. The only difference is that the signals are delayed and decayed slightly due to sound propagation in the air. The results are presented in Table 1. Using only the log magnitude features as input for deep clustering, we obtain 10.3 dB, which matches the 10.3 dB SDR result obtained on the original wsj0-2mix in [2], using a similar architecture. After further incorporating the cosIPD, cosIPD+sinIPD, and GCC features for model training, we achieve 12.5 dB, 12.9 dB, and 12.9 dB, respectively. Surprisingly, these results are comparable to or even better than using the ideal ratio mask (IRM), an oracle mask defined as the magnitude of each source over the sum of all the magnitudes. It is also close to the

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

[2] http://www.merl.com/demos/deep-clustering

[3] The maximum time delay is 0.25/340*8,000=5.88 samples.

result obtained using the ideal binary mask (IBM) defined based on which source dominates each T-F unit.

Next, we evaluate our algorithms on the spatialized reverberant wsj0-2mix data. As shown in Table 2, reverberation substantially deteriorates the performance of single-channel deep clustering to 6.9 dB SDR. This is likely because reverberation blurs the spectral features and breaks the assumption of sparsity in the speech spectrogram, making mask-based separation more difficult. Using the GCC feature for training improves the performance to 8.8 dB, indicating its effectiveness for encoding spatial information. We can see that using variance normalization to equalize the variance of the spectral and spatial features leads to significant improvement, from 7.5 dB to 8.8 dB. Interestingly, replacing the GCC feature by the much simpler cosIPD feature (which matches a single dimension of GCC: $\text{cosIPD}(t, f, p, q) = \text{GCC}(t, f, p, q, 0)$) leads to almost the same performance, at 8.6 dB. Further appending the sinIPD feature pushes the performance to 8.9 dB. This is likely because the spectral feature can help to resolve the IPD ambiguity, and the information in the GCC feature is actually a set of linear combinations of the cosIPD and sinIPD features: $\text{GCC}(t, f, p, q, \tau)$ can indeed be developed as $\cos\left(\frac{2\pi f}{N}\tau\right)\text{cosIPD}(t, f, p, q) - \sin\left(\frac{2\pi f}{N}\tau\right)\text{sinIPD}(t, f, p, q)$, so the network may implicitly recreate some of this information as needed.

The systems by [10] and [11] derive a beamformer for each source based on single-channel deep clustering to obtain the final separation results. Their systems rely heavily on the performance of beamforming, as only slight reverberation is considered and a large number of microphones are simulated in their study (eight microphones in [10] and six in [11]). It is well-known that beamforming methods [30–32] perform well in anechoic environments. However, they are less effective when room reverberation is present, when the sources are close to each other, and when the number of microphones is limited. To compare our approach with theirs, and following the recent development of T-F masking based beamforming [32,33], we use the IRM of each source directly to compute the oracle spatial covariance matrices and report oracle multi-channel Wiener filter results in Table 2. We also performed oracle minimum variance distortionless response (MVDR) beamforming as in [12,34], but results were significantly worse than the MCWF. The MCWF is computed in the following way:

$$\mathbf{w}(f) = \mathbf{\Phi}_x^{-1}(f)\mathbf{\Phi}_s(f)\mathbf{u} \tag{6}$$

$$\mathbf{\Phi}_s(f) = \frac{1}{\sum_t M_{t,f}^{(s)}} \sum_t M_{t,f}^{(s)} \mathbf{x}_{t,f}\mathbf{x}_{t,f}^H \tag{7}$$

where $\mathbf{\Phi}_s$ and $\mathbf{\Phi}_x$ represent the spatial covariance matrices of the target and mixture speech, $M_{t,f}^{(s)}$ is the median IRM of source s across all the microphone channels [32], and $\mathbf{u}$ is a one-hot vector denoting the reference microphone. In our experiments, the first element of $\mathbf{u}$ is set to one and the rest is set to zero, as we always perform separation on the first channel signal and consider the reverberant speech of each source at the first channel as the reference to compute the SDR. The oracle MCWF performance for two microphones in this setting is only 4.9 dB SDR, while increasing the number of microphones leads to consistent improvements. Nonetheless, it is interesting to see that our algorithm, only using two microphones, is able to compete with the oracle MCWF beamforming using up to five microphones. We also emphasize that the proposed algorithm may benefit the systems in [10, 11], because the use of spatial information in our approach leads to better estimated masks and more discriminative embeddings, which is likely to in turn lead to better beamforming performance.

Table 1. SDR (dB) results on spatialized anechoic wsj0-2mix data

Approaches	Features	SDR
1ch Deep Clustering	Log mag.	10.3
2ch Deep Clustering	Log mag. + cosIPD	12.5
2ch Deep Clustering	Log mag. + cosIPD + sinIPD	12.9
2ch Deep Clustering	Log mag. + GCC	12.9
IRM/IBM	-	12.7/13.5

Table 2. SDR (dB) results on spatialized reverberant wsj0-2mix data

Approaches	Features	SDR
1ch Deep Clustering	Log mag.	6.9
2ch Deep Clustering	Log mag., GCC	7.5
	+ variance normalization	8.8
2ch Deep Clustering	Log mag., cosIPD	8.6
	Log mag., cosIPD, sinIPD	8.9
3ch Deep Clustering	Log mag., cosIPD, sinIPD	9.3
4ch Deep Clustering	Log mag., cosIPD, sinIPD	9.4
Oracle MCWF (2ch)	-	4.9
Oracle MCWF (3ch)	-	7.0
Oracle MCWF (4ch)	-	8.3
Oracle MCWF (5ch)	-	9.2
Oracle MCWF (6ch)	-	9.9
Oracle MCWF (7ch)	-	10.5
Oracle MCWF (8ch)	-	10.9
MESSL [18]	see [18]	3.3
GCC-NMF [19]	see [19]	2.7
IRM/IBM	-	11.9/12.7

We also compare our algorithm with the MESSL[4] algorithm [18] proposed for two-microphone arrays. We use for MESSL the same set of potential time delays as in the GCC feature. Performance was 3.3 dB SDR. The GCC-NMF[5] [19] is a recently proposed blind source separation algorithm for two-microphone arrays, which combines non-negative matrix factorization (NMF) based unsupervised dictionary learning with GCC based spatial localization to estimate a binary mask for each source for separation. Performance was 2.7 dB SDR in our experiments.

Applying the two-channel deep clustering model directly to arrays with three and four microphones by concatenating the embeddings improves the performance slightly from 8.9 dB to 9.3 dB and 9.4 dB, respectively. These results outperform the oracle MCWF results using up to five microphones.

5. CONCLUSION

This paper proposed a novel approach to combine deep clustering with spatial clustering for blind source separation. By including phase difference features in the input to a deep clustering network, we can encode both spatial and spectral information in the embeddings it creates, leading to better estimated time-frequency masks. While we considered here the case of two-speaker separation, our algorithm can be readily extended to address under-determined cases where more than two speakers are presented, simply by modifying the number of clusters in the final k-means clustering step. Future work will consider combining the proposed approach with beamforming algorithms.

[4] Available online at https://github.com/mim/messl.

[5] Available online at https://github.com/seanwood/GCC-nmf.

6. REFERENCES

[1] J. R. Hershey, Z. Chen, and J. Le Roux, “Deep Clustering: Discriminative Embeddings for Segmentation and Separation,” in *Proc. ICASSP*, Mar. 2016.

[2] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, “Single-Channel Multi-Speaker Separation using Deep Clustering,” in *Proc. Interspeech*, Sep. 2016.

[3] J. Hershey, S. Rennie, P. A. Olsen, and T. T. Kristjansson, “Super-Human Multi-Talker Speech Recognition: A Graphical Modeling Approach,” *Computer Speech & Language*, vol. 24, no. 1, 2010.

[4] F. Bach and M. Jordan, “Learning Spectral Clustering, with Application to Speech Separation,” *JMLR*, vol. 7, 2006.

[5] D. Wang and G. J. Brown, *Eds., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. Hoboken, NJ: Wiley-IEEE Press, Sep. 2006.

[6] S. Rickard and O. Yilmaz, “On the Approximate W-disjoint Orthogonality of Speech,” *Proc. ICASSP*, 2002.

[7] S. Rickard, “The DUET Blind Source Separation Algorithm,” in *Blind Speech Separation*, 2007.

[8] H. Sawada, S. Araki, and S. Makino, “A Two-Stage Frequency-Domain Blind Source Separation Method for Underdetermined Convolutive Mixtures,” in *Proc. WASPAA*, Oct. 2007.

[9] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, “A Consolidated Perspective on Multi-Microphone Speech Enhancement and Source Separation,” *IEEE/ACM Trans. ASLP*, vol. 25, 2017.

[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*, Aug. 2017.

[11] L. Drude and R. Haeb-Umbach, “Tight Integration of Spatial and Spectral Features for BSS with Deep Clustering Embeddings,” in *Proc. Interspeech*, Aug. 2017.

[12] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, “Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks,” in *Proc. Interspeech*, Sep. 2016.

[13] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The Third ‘CHiME’ Speech Separation and Recognition Challenge: Analysis and Outcomes,” *Computer Speech & Language*, vol. 46, 2017.

[14] Y. Jiang, D. L. Wang, R. S. Liu, and Z. M. Feng, “Binaural Classification for Reverberant Speech Segregation using Deep Neural Networks,” *IEEE/ACM Trans. ASLP*, vol. 22, no. 12, 2014.

[15] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, and T. Nakatani, “Exploring Multi-channel Features for Denoising-autoencoder-based Speech Enhancement,” in *Proc. ICASSP*, Apr. 2015.

[16] P. Pertilä and J. Nikunen, “Distant Speech Separation using Predicted Time-Frequency Masks from Spatial Features,” *Speech Communication*, vol. 68, 2015.

[17] X. Zhang and D. Wang, “Deep Learning Based Binaural Speech Separation in Reverberant Environments,” *IEEE/ACM Trans. ASLP*, vol. 25, no. 5, 2017.

[18] M. I. Mandel, R. J. Weiss, and D. P. Ellis, “Model-Based Expectation-Maximization Source Separation and Localization,” *IEEE Trans. ASLP*, vol. 18, no. 2, 2010.

[19] S. U. Wood, J. Rouat, S. Dupont, and G. Pironkov, “Blind Speech Separation and Enhancement with GCC-NMF,” *IEEE/ACM Trans. ASLP*, vol. 25, no. 4, 2017.

[20] Z.-Q. Wang, J. Le Roux, and J. R. Hershey, “Alternative Objective Functions for Deep Clustering,” in *Proc. ICASSP*, Apr. 2018.

[21] T. Yoshioka, N. Ito, M. Delcroix, A. Ogawa, K. Kinoshita, M. Fujimoto, C. Yu, W. J. Fabian, M. Espi, T. Higuchi, S. Araki, and T. Nakatani, “The NTT CHiME-3 System: Advances in Speech Enhancement and Recognition for Mobile Multi-microphone Devices,” in *Proc. ASRU*, Dec. 2015.

[22] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, “Robust MVDR Beamforming using Time-frequency Masks for Online/Offline ASR in Noise,” in *Proc. ICASSP*, Mar. 2016.

[23] T. Nakatani, N. Ito, T. Higuchi, S. Araki, and K. Kinoshita, “Integrating DNN-Based and Spatial Clustering-Based Mask Estimation for Robust MVDR Beamforming,” in *Proc. ICASSP*, Mar. 2017.

[24] J. DiBiase, H. Silverman, and M. Brandstein, “Robust Localization in Reverberant Rooms,” in *Microphone Arrays*. Berlin Heidelberg: Springer, 2001.

[25] Z.-Q. Wang and D. Wang, “On Spatial Features for Supervised Speech Separation and its Application to Beamforming and Robust ASR,” in *Proc. ICASSP*, Apr. 2018.

[26] Z. Chen, Y. Luo, and N. Mesgarani, “Deep Attractor Network for Single-Microphone Speaker Separation,” in *Proc. ICASSP*, Mar. 2017.

[27] Y. Luo, Z. Chen, and N. Mesgarani, “Speaker-Independent Speech Separation with Deep Attractor Network,” *IEEE/ACM Trans. ASLP*, Apr. 2018.

[28] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multi-Talker Speech Separation and Tracing with Permutation Invariant Training of Deep Recurrent Neural Networks,” in *arXiv:1703.06284*, Mar. 2017.

[29] J. B. Allen and D. A. Berkley, “Image Method for Efficiently Simulating Small-Room Acoustics,” *JASA*, vol. 65, 1979.

[30] E. Habets, J. Benesty, S. Gannot, and I. Cohen, “The MVDR Beamformer for Speech Enhancement,” in *Speech Processing in Modern Communication*. Springer, 2010.

[31] E. Warsitz and R. Haeb-Umbach, “Blind Acoustic Beamforming Based on Generalized Eigenvalue Decomposition,” *IEEE Trans. ASLP*, vol. 15, 2007.

[32] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, “BLSTM Supported GEV Beamformer Front-end for the 3rd CHiME Challenge,” in *Proc. ASRU*, Dec. 2015.

[33] X. Zhang, Z.-Q. Wang, and D. Wang, “A Speech Enhancement Algorithm by Iterating Single- and Multi-microphone Processing and its Application to Robust ASR,” in *Proc. ICASSP*, Mar. 2017.

[34] M. Souden, J. Benesty, and S. Affes, “On Optimal Frequency-domain Multichannel Linear Filtering for Noise Reduction,” *IEEE Trans. ASLP*, vol. 18, 2010.