INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Deep clustering-based beamforming for separation with unknown number of sources

Takuya Higuchi, Keisuke Kinoshita, Marc Delcroix, Kateřina Žmolíková, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan
{higuchi.takuya, kinoshita.k, nakatani.tomohiro}@lab.ntt.co.jp

Abstract

This paper extends a deep clustering algorithm for use with time-frequency masking-based beamforming and perform separation with an unknown number of sources. Deep clustering is a recently proposed single-channel source separation algorithm, which projects inputs into the embedding space and performs clustering in the embedding domain. In deep clustering, bi-directional long short-term memory (BLSTM) recurrent neural networks are trained to make embedding vectors orthogonal for different speakers and concurrent for the same speaker. Then, by clustering the embedding vectors at test time, we can estimate time-frequency masks for separation. In this paper, we extend the deep clustering algorithm to a multiple microphone setup and incorporate deep clustering-based time-frequency mask estimation into masking-based beamforming, which has been shown to be more effective than masking for automatic speech recognition. Moreover, we perform source counting by computing the rank of the covariance matrix of the embedding vectors. With our proposed approach, we can perform masking-based beamforming in a multiple-speaker case without knowing the number of speakers. Experimental results show that our proposed deep clustering-based beamformer achieves comparable source separation performance to that obtained with a complex Gaussian mixture model-based beamformer, which requires the number of sources in advance for mask estimation.

Index Terms: source separation, source counting, time-frequency masking, beamforming

1. Introduction

This paper deals with source counting and separation with a multichannel microphone setup. Time-frequency masking-based beamforming has recently been developed for speech enhancement [1–5], and is reported to be effective for noise robust automatic speech recognition (ASR) [6, 7]. These approaches utilize time-frequency masking to obtain covariance matrices of speaker(s) and noise. Several approaches have been proposed for estimating the mask [1–3, 5].

Our previous approach uses a complex Gaussian mixture model (CGMM) for time-frequency mask estimation [2, 8], where the CGMM parameters are estimated by the Expectation-Maximization algorithm exploiting spatial information extracted with a microphone array. After the parameter estimation, each of the Gaussians is assigned to one of the speaker classes or the noise class, which enables us to estimate a time-frequency mask for each speaker or noise. While the CGMM is capable of dealing with multiple speakers, this approach needs to know the number of sources in advance to set the number of Gaussians for the CGMM parameter estimation.

On the other hand, data-driven approaches for mask estimation have been developed for speech enhancement [1, 5, 9–15] and their effectiveness for noise robust ASR has recently been widely reported [1, 3–5]. Heymann et al. use a bi-directional long short-term memory (BLSTM) for their masking-based beamformer, where the BLSTM is trained to predict a mask for extracting the speaker and a mask for extracting noise. The BLSTM is typically trained in advance by using a large amount of parallel data consisting of noisy and clean speech. Although a BLSTM is capable of dealing with one target speaker, speech separation with an unknown number of sources and/or more than two sources is not a trivial extension for the following reasons.

One reason is that the dimension of BLSTM outputs, i.e., the vector dimension of an estimated mask, needs to be fixed during the training time, which makes it difficult to deal with the different number of sources in the test time. Another reason is the difficulty in defining each source class during training when the sources have similar properties, e.g. speech mixtures. Although each source signal needs to be assigned to a specific part of the BLSTM output, its arbitrary property causes a permutation problem among source classes, and leads to a training failure [16].

To address these problems, recent studies have exploited the embedding space and achieved multiple speaker separation with deep learning [17–19]. Instead of directly outputting the masks, BLSTMs in [17–19] output an embedding vector for each time-frequency point. The BLSTMs are trained to predict embedding vectors that are in the same direction for time-frequency points dominated by the same speaker, or orthogonal for those dominated by different speakers. During a test, mask estimation for multiple speakers can be performed by K-means clustering of the embedding vectors. Since the embedding vector estimation can be performed with the trained BLSTMs independently of the number of sources, we can deal with an arbitrary number of sources with deep clustering simply by setting the number of classes for K-means clustering during the test.

By incorporating deep clustering into the masking-based beamforming approach, this paper extends the application of the deep clustering algorithm to beamforming for multiple target speakers. Moreover, we propose a source counting method based on deep clustering by using the rank of the covariance matrix of the embedding vectors, which allows us to perform source separation with an unknown number of sources. Ideally, the embedding vectors are in C orthogonal directions, where C denotes the number of sources. This nature of the embedding vectors allows us to perform source counting by performing the eigenvalue decomposition of the covariance matrix of the embedding vectors. The overall proposed system allows us to separate an unknown number of sources.

Experimental results showed that our approach achieved a 67.3% source counting accuracy for 2- and 3-speaker mixtures. In terms of source separation ability, our approach achieved an 11.51 dB signal-to-distortion ratio (SDR) improve-

http://dx.doi.org/10.21437/Interspeech.2017-721

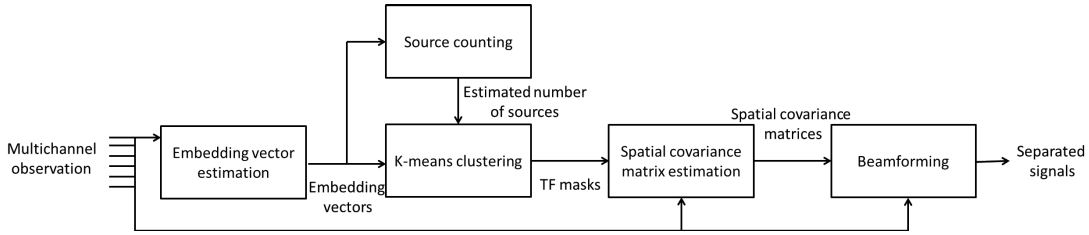

Figure 1: *Schematic diagram of our system architecture for source separation with an unknown number of sources.*

ment for 2-speaker mixtures and 9.59 dB for 3-speaker mixtures, which were comparable performances to those obtained with the CGMM-based approach that needs to know the number of sources in advance.

The remainder of this paper is organized as follows. Section 2 provides an overview of our proposed source counting and separation method. Section 3 briefly reviews deep clustering for mask estimation and describes our proposed source counting method based on the embedding vectors. Section 4 describes a masking-based beamformer. Section 5 presents an experimental evaluation of our proposed approach in terms of source counting accuracy and SDRs. Section 6 concludes the paper.

2. Overview of proposed beamformer

Figure 1 shows an overview of our proposed beamformer. The beamformer consists of embedding vector estimation, source counting, mask estimation by K-means clustering, spatial covariance matrix estimation and beamforming. We first estimate an embedding vector for each time-frequency point from a single-channel input with a BLSTM. The BLSTM is trained in advance to create embedding vectors in the same direction for the same speaker, and in an orthogonal direction for different speakers. We utilize this nature of the embedding vectors for source counting. With the estimated number of sources and the embedding vectors, we perform K-means clustering, after which each cluster corresponds to each source class. Then, based on this clustering result, we can obtain a time-frequency mask for each speaker. The masks are used to obtain the spatial covariance matrices of the speakers, and the covariance matrices are used to estimate beamformer coefficients. Finally separated signals are obtained by beamforming with observed multichannel signals.

3. Source counting and mask estimation with deep clustering

3.1. Deep clustering

Let $n \in \{1, ..., N\}$ denote an index for a time-frequency point, and $c \in \{1, ..., C\}$ denote a source index. An objective function to be minimized for BLSTM training is defined as

$$J(\Theta) = ||YY^{\mathrm{T}} - VV^{\mathrm{T}}||_F^2, \tag{1}$$

where $Y = \{y_{n,c}\}$ denotes an $N \times C$ indicator matrix and $V = \{v_{n,d}\}$ denotes an $N \times D$ matrix consisting of embedding vectors estimated with the BLSTM. Θ denotes the learnable parameters of the BLSTM. $y_{n,c} = 1$ if the c-th source dominates at time-frequency point n, otherwise, $y_{n,c} = 0$. The n-th row of V corresponds to a D dimensional embedding vector for the n-th time-frequency point, which is estimated by the BLSTM.

In Eq. (1), YY^{T} is an $N \times N$ supervision matrix, where the element at (n, n') is 1 if time-frequency points n and n' are dominated by the same speaker, otherwise the element at (n, n') is 0. To minimize the objective function, the embedding vector for the n-th time-frequency point, $\boldsymbol{v}_n = (v_{n,1}, ..., v_{n,D})^{\mathrm{T}}$, would be parallel to $\boldsymbol{v}_{n'}$ if time-frequency points n and n' are dominated by the same speaker, otherwise $\boldsymbol{v}_n^{\mathrm{T}}\boldsymbol{v}_n$ would be zero.

This training forces the embedding vectors to form clusters during the test, where one cluster corresponds to one source class. By performing the clustering, e.g. K-means clustering, of estimated embedding vectors $\boldsymbol{v}_1, ..., \boldsymbol{v}_N$, we can obtain time-frequency masks for each source. If $\boldsymbol{v}_n$ is assigned to the c-th source class, the estimated mask for the c-th source at the n-th time-frequency point is 1.

In general, we need to set the number of sources for clustering. In previous reports on deep clustering, the number of sources was assumed to be given, and K-means clustering was performed with the oracle number of sources [17–19].

3.2. Rank estimation for source counting

In this study, we estimate the number of sources with the estimated embedding vectors. Ideally, the embedding vectors are in C directions that are orthogonal to each other. This nature allows us to estimate the number of sources by estimating the rank of the covariance matrix of the embedding vectors.

We compute a covariance matrix of the embedding vectors $\boldsymbol{A}$ as

$$\boldsymbol{A} = \frac{1}{N}\sum_{N} \boldsymbol{v}_n \boldsymbol{v}_n^{\mathrm{T}}, \tag{2}$$

and extract eigenvalues $e_1, ..., e_D$ with the covariance matrix $\boldsymbol{A}$.

The rank of $\boldsymbol{A}$ corresponds to the number of sources, therefore we assume that the number of eigenvalues larger than a threshold is the number of sources as

$$\hat{C} = n(\boldsymbol{E}), \tag{3}$$

where $n(\cdot)$ is operation to compute the number of elements, and $\boldsymbol{E} = \{e_d \mid e_d > b, d = 1, ..., D\}$ is a set of eigenvalues of $\boldsymbol{A}$ larger than the threshold b. $\hat{C}$ denotes the estimated number of sources. By performing K-means clustering with the estimated number of sources $\hat{C}$, we can perform mask estimation without knowing the number of sources in advance.

4. Masking-based beamformer for source separation

This section briefly reviews a masking-based beamformer recently developed for noise robust ASR [1–5]. Although previous deep clustering approaches obtain separated signals by multiplying the estimated masks by the observed mixture in the time-frequency domain, the masking often introduces artificial noise and its effectiveness for ASR is known to be limited [20, 21]. In this paper, we assume a multiple microphone

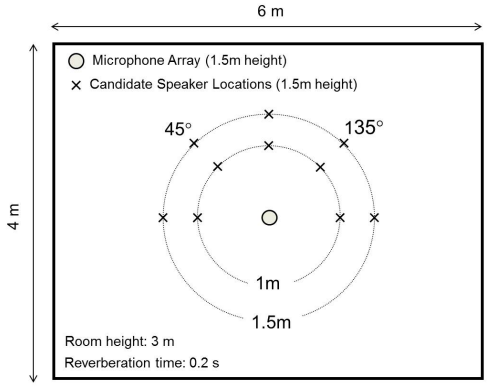

Figure 2: *Simulated room conditions. A circular microphone array with 8 microphones was located in the center of the room.*

setup, and we obtain separated signals with beamforming as

$$\hat{s}_{f,t}^{(c)} = \boldsymbol{w}_f^{(c)\mathrm{H}} \boldsymbol{x}_{f,t}, \tag{4}$$

where $\hat{s}_{f,t}^{(c)}$, $\boldsymbol{w}_f^{(c)} = (w_{f,1}^{(c)}, ..., w_{f,M}^{(c)})^{\mathrm{T}}$ and $\boldsymbol{x}_{f,t} = (x_{f,t,1}, ..., x_{f,t,M})^{\mathrm{T}}$ denote the c-th estimated signal at time-frequency point (t, f), beamformer coefficients for the c-th source at frequency f and an observed multichannel signal at (t, f), respectively. M denotes the number of microphones.

There are several approaches that can be used to obtain the beamformer coefficients based on the spatial covariance matrices of sources, e.g. a minimum variance distortion-less response beamformer [22] and a maximum signal-to-noise ratio (max-SNR) beamformer [23]. For example, the max-SNR beamformer can be defined by

$$\boldsymbol{w}_f^{(c)} = \arg\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^{\mathrm{H}} \boldsymbol{R}_f^{(c)} \boldsymbol{w}}{\boldsymbol{w}^{\mathrm{H}} \boldsymbol{R}_f^{(\neg c)} \boldsymbol{w}}, \tag{5}$$

where $\boldsymbol{R}_f^{(c)}$ denotes the spatial covariance matrix of the c-th source, and $\boldsymbol{R}_f^{(\neg c)}$ denotes the spatial covariance matrix of interference against the c-th source. The beamformer can eventually be obtained by solving the generalized eigenvalue problem

$$\boldsymbol{R}_f^{(c)} \boldsymbol{w} = e_{max} \boldsymbol{R}_f^{(\neg c)} \boldsymbol{w}, \tag{6}$$

where e_{max} denotes the maximum eigenvalues of a matrix $\boldsymbol{R}_f^{(\neg c)-1} \boldsymbol{R}_f^{(c)}$.

The spatial covariance matrices $\boldsymbol{R}_f^{(c)}$ and $\boldsymbol{R}_f^{(\neg c)}$ can be obtained with the estimated time-frequency masks as follows

$$\boldsymbol{R}_f^{(c)} = \frac{1}{\sum_t \lambda_{f,t}^{(c)}} \sum_t \lambda_{f,t}^{(c)} \boldsymbol{x}_{f,t} \boldsymbol{x}_{f,t}^{\mathrm{H}}, \tag{7}$$

$$\boldsymbol{R}_f^{(\neg c)} = \frac{1}{\sum_t (1 - \lambda_{f,t}^{(c)})} \sum_t (1 - \lambda_{f,t}^{(c)}) \boldsymbol{x}_{f,t} \boldsymbol{x}_{f,t}^{\mathrm{H}}, \tag{8}$$

where $\lambda_{f,t}^{(c)}$ denotes the time-frequency mask for the c-th source at (t, f) estimated by deep clustering.

5. Experimental evaluation

This section describes experimental evaluations of our proposed approach in terms of source counting and separation performance. We used simulated multichannel speech mixtures with 2 or 3 speakers, and evaluated source counting accuracy while varying the threshold b. Then, we evaluated source separation performance in terms of SDRs computed with [24].

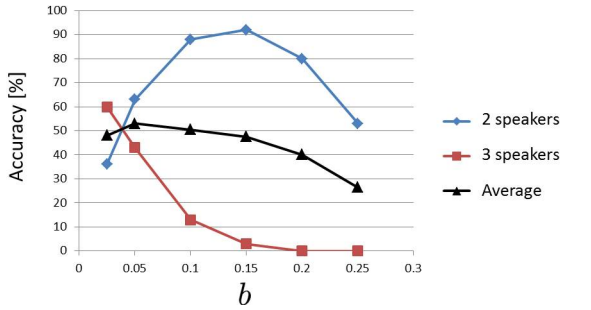

Figure 3: *Source counting accuracy obtained with BLSTM-2 for the validation set.*

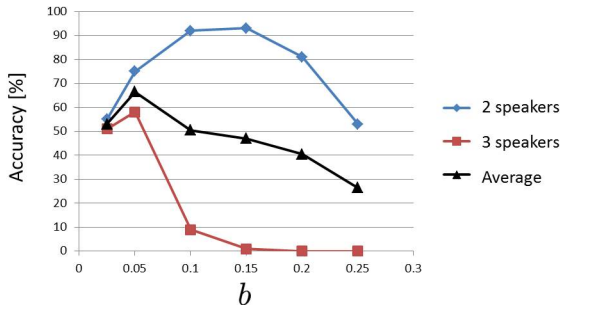

Figure 4: *Source counting accuracy obtained with BLSTM-both for the validation set.*

5.1. Data

We created data sets of multichannel speech mixtures based on the Wall Street Journal (WSJ0) corpus and the data generation code provided in [25], which was also used for the evaluation of deep clustering in previous reports [17–19]. We generated two data sets of multichannel mixtures, one of which consisted of 2-speaker mixtures and the other consisted of 3-speaker mixtures. To simulate multichannel mixtures, we convolved impulse responses with the speech signals. The impulse responses were generated with the image method [26, 27]. Figure 2 shows the details of the room conditions that were used for impulse response generation. The room size was $6m \times 4m \times 3m$ and RT_{60} was $0.2s$. We assumed a circular array of 8 microphones that was 20 cm in diameter and located at the center of the room. A source location was randomly selected from the candidate locations shown in Figure 2. A training data set was generated by randomly mixing the simulated multichannel utterances produced by different speakers from the WSJ0 training set. Mixing signal-to-noise ratios (SNRs) varied from 0 to 5 dB. The training data totaled 30 hours. A 10-hour cross validation set was generated in a similar way with closed speakers to optimize tunable parameters. 5 hours of evaluation data were generated with the different speakers from the training data set to evaluate source separation performance. The training, cross validation and evaluation data sets were prepared individually for 2-speaker mixtures and 3-speaker mixtures.The audio data were downsampled to 8 kHz.

5.2. Settings for BLSTM training

We used stacked BLSTMs for embedding vector estimation. The BLSTMs consisted of 4 BLSTM layers followed by a linear layer, where each of the BLSTM layers had 300 forward LSTM cells and 300 backward cells. Linear transformations were applied between the BLSTM layers to concatenate the for-

Table 1: *Source counting accuracies [%] and SDR improvements [dB] for the evaluation set.*

Systems	Number of sources	Source counting accuracy [%]			SDRs [dB]		
		2 speakers	3 speakers	Avg.	2 speakers	3 speakers	Avg.
CGMM	Oracle	-	-	-	11.48	**10.95**	**11.22**
BLSTM-2	Oracle	-	-	-	11.08	6.45	8.77
	Estimated (b=0.05)	61.1	44.5	52.8	11.27	6.25	8.76
BLSTM-both	Oracle	-	-	-	11.36	10.27	10.82
	Estimated (b=0.05)	74.8	59.8	67.3	**11.51**	9.59	10.55

ward and backward propagation results. The final linear layer projected the BLSTM outputs to $F \times D$ dimensions. We set D at 40. We used the log magnitude spectrum of the mixture speech as input features, where the short time Fourier transform was performed with a 32 ms window length, 8 ms window shift and a hanning window. The silence regions of the time-frequency points were ignored in the cost computation during training similarly to [17, 18]. The silence regions were defined as time-frequency points where the magnitude was smaller than -40 dB of the maximum mixture magnitude. BLSTM network propagation was performed using the whole length of an utterance, while the cost function defined in Eq. (1) was computed solely with 400 randomly chosen frames to save memory. The BLSTM training was performed with the rmsprop algorithm [28], where the learning rate was set at $l = 0.001 \times (1/2)^{\lfloor \epsilon/50 \rfloor}$. ϵ was the epoch number. The mini-batch size was set at 16. We trained the BLSTMs with the 2-speaker mixture data set (BLTSM-2) and with both the 2- and 3-speaker mixture data set (BLSTM-both). For BLSTM-both, the learning rate was set at $l = 0.001 \times (1/2)^{\lfloor \epsilon/25 \rfloor}$ because of the double size of the training data set.

5.3. Source counting performance evaluation

We tuned the threshold b with the validation set and evaluated the source counting performance of our proposed approach. Figure 3 shows source counting accuracy for the validation set obtained with BLSTM-2. The blue and red lines indicate the source counting accuracies for the 2-speaker mixtures and the 3-speaker mixtures, respectively. The black line indicates their average value. By training the BLSTM with the 2-speaker mixtures, we obtained a 91.7% source counting accuracy for the 2-speaker mixtures with $b = 0.15$, however, the best parameter for the 3-speaker mixtures was quite different from that for the 2-speaker mixtures. The best accuracy for both the 2- and 3-speaker mixtures was 53.1% with $b = 0.05$. Figure 4 shows source counting accuracy for the validation set obtained with BLSTM-both. By training with both the 2- and 3-speaker mixtures, BLSTM-both achieved a higher source counting accuracy than that obtained with BLSTM-2. The best total source counting accuracy was 66.3% with the best parameter $b = 0.05$. With the best parameter, the source counting accuracies for the evaluation set were 52.8% with BLSTM-2 and 67.3% with BLSTM-both, as shown in Table 1.

5.4. Source separation performance evaluation

We evaluated our proposed approach in terms of the SDRs [24]. For SDR computation with the estimated number of sources, we used the following procedure to align the number of processed signals to the actual number of sources. When the number of sources was underestimated, we used the mixture as the rest of the processed signals. When the number of sources was overestimated, we used C clusters, which have the largest number of cluster members, for the processed signals. We used the MVDR beamformer to obtain the separated signals. A steering vector for the MVDR beamformer was obtained by multiplying the max-SNR beamformer coefficients by the spatial covariance matrix of interference. Our beamformer is detailed in [29]. For comparison, we used a CGMM-based beamformer that has been successfully used as a front-end for noise robust ASR [2,20,30].

Table 1 shows the SDRs we obtained with the CGMM-based beamformer with the oracle number of sources and the proposed deep clustering-based beamformer with the oracle/estimated number of sources. With BLSTM-2, the proposed approach achieved comparable SDR improvements to those obtained with the CGMM-based beamformer for the 2-speaker mixtures, however, the separation performance for the 3-speaker mixtures was limited even with the oracle number of sources. With BLSTM-both and the oracle number of sources, our proposed approach achieved a comparable performance to the CGMM-based beamformer for both the 2- and 3-speaker mixtures. Although the average separation performance slightly degraded when using the estimated number of sources, the performance was still comparable to that obtained with the CGMM-based beamformer.

Even with the estimated number of sources, the separation performance of our proposed approach was not greatly degraded. One reason is that a source counting failure means that the embedding vector estimation is not very accurate, which would lead to poor mask estimation even with the oracle number of sources. Another reason is that, when evaluating the SDR in the overestimation case, clusters that have fewer members were excluded from the evaluation targets. This suggests that the number of cluster members would be useful for the refinement of source counting, which constitutes future work. Moreover, our current BLSTM used just the spectral features extracted with a single microphone, therefore our future work will also include utilizing spatial features as with [10, 12] for more accurate embedding vector estimation and source counting.

6. Conclusion

In this paper, we proposed a deep clustering-based beamformer for source counting and separation. We perform source counting by computing the rank of the covariance matrix of embedding vectors, and incorporate deep clustering-based source counting and mask estimation into masking-based beamforming. An experimental evaluation showed that our approach achieved a 67.3% source counting accuracy and a 10.55 dB SDR improvement for 2- and 3-speaker mixtures, which was a comparable separation performance to that of a CGMM-based beamformer, which needs to know the number of sources in advance. Our future work will include ASR performance evaluation with our proposed approach for unknown and time-varying numbers of sources.

7. References

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

[2] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 5210–5214.

[3] H. Erdogan, J. R. Hershey, S. Watanabe, M. Mandel, and J. Le Roux, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. INTERSPEECH*, 2016.

[4] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 276–280.

[5] J. Heymann, L. Drude, C. Boeddeker, P. Hanebrink, and R. Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel ASR system," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5325–5329.

[6] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: dataset, task and baselines," in *Proc. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 504–511.

[7] "The 4th CHiME speech separation and recognition challenge," available online: http://spandh.dcs.shef.ac.uk/chime_challenge/index.html.

[8] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Proc. Int. Worksh. Acoust. Echo, Noise Contr.*, 2014, pp. 268–272.

[9] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising." in *INTERSPEECH*, 2014, pp. 2685–2689.

[10] Y. Jiang, D. Wang, R. Liu, and Z. Feng, "Binaural classification for reverberant speech segregation using deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 2112–2121, 2014.

[11] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 91–99.

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

[13] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "DNN-based enhancement of noisy and reverberant speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6525–6529.

[14] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5220–5224.

[15] C. Boeddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, "Optimizing neural-network supported acoustic beamforming by algorithmic differentiation," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 171–175.

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

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

[18] 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.

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

[20] 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. Worksh. Automat. Speech Recognition, Understanding*, 2015, pp. 436–443.

[21] M. Delcroix and S. Watanabe, "Recent advances in distant speech recognition," tutorial, Interspeech 2016.

[22] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on antennas and propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[23] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Transactions on audio, speech, and language processing*, vol. 15, no. 5, pp. 1529–1539, 2007.

[24] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[25] Available online: http://www.merl.com/demos/deep-clustering.

[26] 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.

[27] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[28] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning*, vol. 4, no. 2, 2012.

[29] N. Ito, S. Araki, M. Delcroix, and T. Nakatani, "Probabilistic spatial dictionary based online adaptive beamforming for meeting recognition in noisy and reverberant environments," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 681–685.

[30] J. Du, Y.-H. Tu, L. Sun, F. Ma, H.-K. Wang, J. Pan, C. Liu, J.-D. Chen, and C.-H. Lee, "The USTC–iFlytek system for CHiME-4 challenge," *Proc. CHiME*, pp. 36–38, 2016.