UNSUPERVISED TRAINING OF A DEEP CLUSTERING MODEL FOR MULTICHANNEL BLIND SOURCE SEPARATION

Lukas Drude, Daniel Hasenklever, Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering, Paderborn, Germany

{drude, haeb}@nt.upb.de

ABSTRACT

We propose a training scheme to train neural network-based source separation algorithms from scratch when parallel clean data is unavailable. In particular, we demonstrate that an unsupervised spatial clustering algorithm is sufficient to guide the training of a deep clustering system. We argue that previous work on deep clustering requires strong supervision and elaborate on why this is a limitation. We demonstrate that (a) the single-channel deep clustering system trained according to the proposed scheme alone is able to achieve a similar performance as the multi-channel teacher in terms of word error rates and (b) initializing the spatial clustering approach with the deep clustering result yields a relative word error rate reduction of $26\,\%$ over the unsupervised teacher.

Index Terms— blind source separation, deep learning, multi-channel, unsupervised learning, student-teacher

1. INTRODUCTION

In recent years, neural network architectures and training recipes demonstrated unprecedented performance in a wide range of applications. In particular, deep clustering (DC) and permutation invariant training (PIT) are pioneering approaches to separate unseen speakers in a single channel mixture [1, 2]. A plethora of subsequent work refined the architectures and training recipes, e.g. with respect to signal reconstruction performance by using a more direct reconstruction loss [3], using multiple complementary losses [4] or by addressing phase reconstruction [5].

In today's applications, e.g. digital home assistants/ meeting assistants, multiple microphones are the de-facto standard. Thus, it is natural to generalize or apply neural network-based source separation to multi-channel scenarios. Earlier work simply exploited spatial information for beamforming to extract the sources as a final processing step but did not make any use of spatial information to improve the clustering performance itself [6]. Subsequently, two rather different approaches emerged, which made direct use of both spatial and spectral features: On the one hand, directly feeding multi-channel features to a neural network yielded great improvements even for more than two channels [7]. On the other hand, the embedding vectors stemming from either DC or a deep attractor network (DAN) can be seen as additional spectral features for an integrated expectation maximization (EM) algorithm which jointly models spatial and spectral features in a single probabilistic model [8].

However, all of the aforementioned approaches need access to parallel clean data, i.e. recordings of the speech source or speech image at the microphones before any mixing took place. Although it may seem that mixing signals artificially is sufficient for training, recent work with the CHiME 5 challenge dataset [9] turned out to be surprisingly complicated: Quasi-parallel data from the in-ear microphones was insufficiently related to the array recordings such that directly training a source separation neural network was impractical. In particular, a correct transmission model of the room, the Lombard effect, and realistic background noise is hard to entirely simulate artificially. One way may be to train an acoustic model with a neural network source separation system end-to-end. However, this again requires transcriptions for the mixtures at hand [10, 11] and may require pretraining the individual components using again parallel data [12, 13].

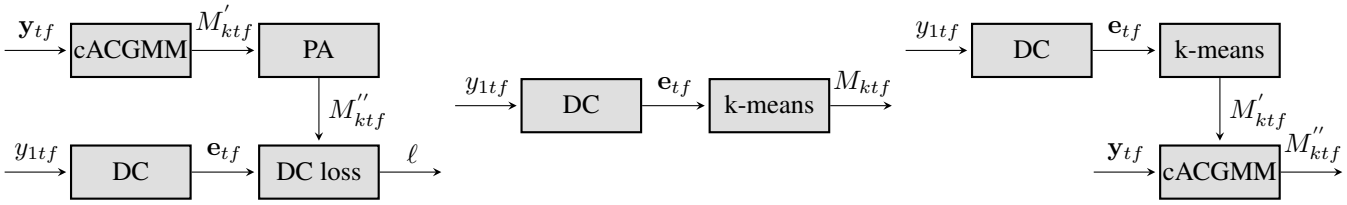
Therefore, we here propose to train a source separation neural network from scratch by leveraging only spatial cues already during training. This can be seen as a student-teacher approach [14], where the unsupervised teacher turns out to be weaker than the student. We demonstrate that a rough unsupervised spatial clustering approach is sufficient to guide the training of the neural network.

2. SIGNAL MODEL

A convolutive mixture of K independent source signals s_{tfk}, captured by D sensors is approximated in the short time Fourier transform (STFT) domain:

$$\mathbf{y}_{tf} = \sum_k \mathbf{h}_{fk}\, s_{tfk} + \mathbf{n}_{tf} = \sum_k \mathbf{x}_{tfk} + \mathbf{n}_{tf}, \qquad (1)$$

where $\mathbf{y}_{tf}$, $\mathbf{h}_{fk}$ $\mathbf{n}_{tf}$, and $\mathbf{x}_{tfk}$ are the D-dimensional observed signal vector, the unknown acoustic transfer function vector of source k, the noise vector, and the source images at the sensors, respectively. Furthermore, t and f specify the time frame index and the frequency bin index, respectively. Since speech signals are sparse in the STFT domain, we may assume that a time frequency slot is occupied either by a single source and noise or by noise only.



(a) Proposed unsupervised training procedure with no need for any clean or parallel data.

(b) Prediction is performed by clustering the embedding vectors $\mathbf{e}_{tf}$ using k-means [1].

(c) Prediction variant: spatial mixture model is initialized with the deep clustering result.

Fig. 1: A deep clustering neural network is trained without parallel data by providing sufficient guidance by an unsupervised spatial clustering algorithm. The resulting masks can then be used for beamforming or to directly mask the observed signal.

3. PROPOSED FRAMEWORK

We propose to use an unsupervised spatial clustering approach as a teacher for a neural network-based source separation student, thus rendering the whole setup to be unsupervised. Subsec. 3.1 introduces the complex angular-central Gaussian mixture model (cACGMM) as an instance of an unsupervised spatial clustering algorithm and names inherent limitations. Subsec. 3.2 explains DC as an example for a neural network-based source separation system which will then serve as the student. Lastly, Subsec. 3.3 and 3.4 detail, how training and prediction are performed.

3.1. Unsupervised spatial clustering

We here decided to use a cACGMM for unsupervised spatial clustering [15], where the normalized complex-valued observation vectors $\tilde{\mathbf{y}}_{tf} = \mathbf{y}_{tf}/\|\mathbf{y}_{tf}\|$ are modeled as follows:

$$p(\tilde{\mathbf{y}}_{tf}; \boldsymbol{\theta}) = \sum_k \pi_{kf} \mathrm{cACG}(\tilde{\mathbf{y}}_{tf}, \mathbf{B}_{kf})$$

with the model parameters $\boldsymbol{\theta} = \{\pi_{kf}, \mathbf{B}_{kf} \forall k, f\}$. The observation model is a complex angular central Gaussian:

$$\mathrm{cACG}(\tilde{\mathbf{y}}_{tf}, \mathbf{B}_{kf}) = \frac{(D-1)!}{2\pi^D \det \mathbf{B}_{kf}} \frac{1}{\left(\tilde{\mathbf{y}}_{tf}^{\mathsf{H}} \mathbf{B}_{kf}^{-1} \tilde{\mathbf{y}}_{tf}\right)^D}. \quad (2)$$

All latent variables and parameters can be estimated on each mixture signal using an EM algorithm[1]. The EM update equations coincide with the update equations of the time-variant complex Gaussian mixture model (TV-cGMM) [16] as demonstrated in the appendix of [15]. The obtained class affiliation posteriors can then be used to extract the sources either by masking or beamforming.

However, it is worth noting that mixture models, in general, tend to be very susceptible to initialization. This will be addressed in more detail in the evaluation section.

The cACGMM neglects frequency dependencies. Thus, when used without any kind of guidance, it will yield a solution where the speaker index is inconsistent over frequency bins. This issue is the so-called frequency permutation problem [17]. We address it by calculating that permutation alignment (PA) (bin by bin) which maximizes the correlation of the masks along neighboring frequencies similar to [17][1].

3.2. Deep Clustering

DC is a technique which aims at blindly separating unseen speakers in a single-channel mixture. The training procedure described in the original work [1, 18] assumes that ideal binary masks (IBMs) for each speaker are available to train a multi-layer bidirectional long short-term memory network (BLSTM) [19] to map from $T \cdot F$ spectral features (e.g. log spectrum) to the same number of E-dimensional embedding vectors $\mathbf{e}_{tf}$, where $\|\mathbf{e}_{tf}\|^2 = 1$. The objective during training is to minimize the Frobenius norm of the difference between the estimated and true affinity matrix (for a discussion of improved loss functions see [20]):

$$\ell = \left\|\hat{\mathbf{A}} - \mathbf{A}\right\|_{\mathrm{F}}^2 = \left\|\mathbf{E}\mathbf{E}^{\mathsf{T}} - \mathbf{C}\mathbf{C}^{\mathsf{T}}\right\|_{\mathrm{F}}^2, \quad (3)$$

where $\hat{\mathbf{A}}$ and $\mathbf{A}$ are the estimated and ground truth affinity matrices. The entries $A_{n,n'}$ encode, whether observation n and n' belong to the same source ($A_{n,n'} = 1$, and zero else). Correspondingly, the embeddings are stacked in a single matrix $\mathbf{E}$ with shape $(TF \times E)$ and the ground truth one-hot vectors describing which time frequency slot belongs to which source are stacked in a single matrix $\mathbf{C}$ with shape $(TF \times K)$, such that $C_{nk} = 1$, if observation n belongs to source k and $C_{nk} = 0$ otherwise.

During training, the network is encouraged to move embeddings belonging to the same source closer together while pushing embeddings which belong to different sources further apart. After training, the embeddings, which are normalized to unit-length, can be clustered to obtain time frequency masks for each source. The original work used k-means clustering. This yields masks which can be used in a subsequent source extraction scheme e.g. masking or beamforming.

[1] Our implementation of the cACGMM and permutation alignment can be found on Github: `https://github.com/fgnt/pb_bss`

3.3. Unsupervised network training

To train a DC system, we need to optimize the affinity loss in Eq. 3. Since ground truth is not available, we instead minimize the Frobenius norm of the difference between the affinity matrix of the embeddings $\mathbf{E}\mathbf{E}^{\mathsf{T}}$ and the affinity matrix according to masks predicted by an unsupervised clustering approach. Here, we use the class predictions (masks) $M^{'}_{ktf}$ of a cACGMM (Subsec. 3.1) and apply an additional frequency permutation alignment (Subsec. 3.1) to obtain $M^{''}_{ktf}$ (Fig. 1 (a)), which is then used to guide the optimization.

It is worth noting that the predicted classes (masks) from the spatial mixture model are first of all less sharp than the unavailable IBMs. Furthermore, the predicted masks often contain frequency permutations errors, when the permutation alignment step did not resolve all permutations (see Fig. 2, left). The assumption which has to be proven in the evaluation section, therefore, is that on average, the cACGMM results are sufficiently good.

Since any kind of ground-truth signal is not available in the unsupervised case, early stopping [21] can only be performed with respect to the aforementioned loss and not with respect to signal to distortion ratio (SDR) gains or similar.

3.4. Prediction

To predict masks at test time two different ways are natural to investigate. First, one can use the trained DC network to predict embeddings $\mathbf{e}_{tf}$ which are then clustered using k-means as in Fig. 1 (b). This approach is the recommended approach according to [1]. Second, one may use the k-means masks as initialization for a cACGMM as in Fig. 1 (c). This can be seen as some kind of weak integration (in contrast to [8]) between the DC model, which just uses spectral features and the cACGMM which just uses spatial features due to the normalization as mentioned in Subsec. 3.1.

4. BEAMFORMING

Both aforementioned models provide a time-frequency mask for each speaker and noise. This can be multiplied with the observed signal STFT to extract each of the sources. Alternatively, we may use it for classic statistical beamforming as a linear time-invariant way to extract each source. To do so, we first calculate covariance matrices for each target speaker:

$$\mathbf{\Phi}_{kf}^{(\text{target})} = \sum_t M_{ktf}^{(\text{target})} \mathbf{y}_{tf}\mathbf{y}_{tf}^{\mathsf{H}} \Big/ \sum_t M_{ktf}^{(\text{target})} \,. \quad (4)$$

Similarly, the covariance matrix of all interferences and noise is calculated with $M_{ktf}^{(\text{inter})} = 1 - M_{ktf}^{(\text{target})}$.

These covariance matrices can then be used to derive a statistically optimal beamformer. In this particular case, we opted for a specific formulation of the minimum variance distortionless response (MVDR) beamformer which avoids explicit knowledge of any kind of steering vector [22, Eq. 24]:

$$\mathbf{w} = \frac{1}{\lambda_{kf}} \mathbf{\Phi}_{kf}\mathbf{u}_k, \;\; \text{with} \;\; \mathbf{\Phi}_{kf} = {\mathbf{\Phi}_{kf}^{(\text{inter})}}^{-1} \mathbf{\Phi}_{kf}^{(\text{target})}, \quad (5)$$

where $\lambda_{kf} = \text{tr}\left(\mathbf{\Phi}_{kf}\right)$ and $\mathbf{u}_k = [0...1...0]^{\mathsf{T}}$ selects the reference microphone. Here, the reference microphone is selected by maximizing the expected signal to noise ratio (SNR) gain [23]. The beamforming vector is then used to linearly project the multi-channel mixture to a single-channel estimate of clean speech: $z_{ktf} = \mathbf{w}_{kf}^{\mathsf{H}}\mathbf{y}_{tf}$.

5. ACOUSTIC MODEL

For an objective comparison, we train a state of the art acoustic model (AM). The hybrid AM consists of a combination of a wide residual network to model local context and a BLSTM to model long term dependencies. The AM is thus dubbed wide bi-directional residual network [24]. To train the AM alignments are extracted with a vanilla DNN-HMM recipe from Kaldi [25]. The AM is trained on artificially reverberated Wall Street Journal (WSJ) utterances without any interfering speaker. Thus, the AM never saw mixed speech and never saw possible artifacts produced by any kind of separation system. For decoding, we use the WSJ trigram language model without additional rescoring.

6. EVALUATION

To evaluate the proposed algorithm, we artificially generated 30000, 500 and 1500 six channel mixtures with a sampling rate of 8 kHz with source signals obtained from three non-overlapping WSJ sets (`train_si284`, `cv_dev93`, `test_eval92`). We generated room impulse responses with the image method [26] with random room dimensions, a

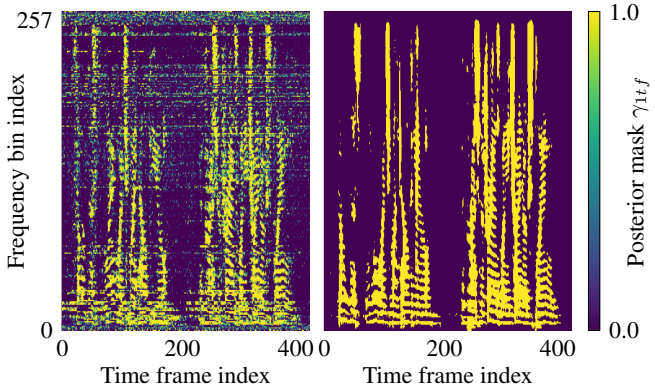


Fig. 2: Intermediate masks generated by the cACGMM (left) guide the neural network training which results in k-means masks with less artifacts (right). Especially the lower frequencies are resolved better.

Table 1: Model comparison for unsupervised systems in terms of objective quality gains and word error rates (WERs). Oracle initialized baseline and supervised baseline are denoted in gray. The best unsupervised result is set in bold.

	Model	Initialization	Extraction	BSS-Eval SDR gain		Invasive SDR gain		PESQ gain	STOI gain	WER / %
				Mean / dB	STD / dB	Mean / dB	STD / dB			
1	cACGMM	Random	Masking	7.2	2.8	10.4	3.3	0.17	0.11	38.4
2	U-DC	Random	Masking	5.5	3.9	9.4	3.7	-0.42	0.04	75.1
3	cACGMM	U-DC result	Masking	**9.5**	**2.4**	13.2	**2.9**	0.40	**0.18**	29.3
4	S-DC	Random	Masking	5.9	4.8	9.5	3.4	-0.25	0.06	75.8
5	cACGMM	S-DC result	Masking	9.1	2.4	12.6	2.8	0.37	0.16	31.0
6	cACGMM	Oracle IBM	Masking	9.7	2.3	13.3	2.6	0.48	0.14	28.9
7	cACGMM	Random	MVDR	5.1	3.2	12.7	4.0	0.37	0.09	28.0
8	U-DC	Random	MVDR	5.7	3.5	13.6	4.5	0.43	0.11	29.0
9	cACGMM	U-DC result	MVDR	6.4	3.4	**15.3**	3.5	**0.52**	0.13	**20.7**
10	S-DC	Random	MVDR	5.9	3.2	14.2	4.1	0.47	0.12	26.5
11	cACGMM	S-DC result	MVDR	6.1	3.3	14.9	3.5	0.50	0.12	21.1
12	cACGMM	Oracle IBM	MVDR	6.4	3.3	15.5	3.4	0.78	0.12	19.9

random position of the circular array and random positions of the two concurring speakers. The minimum angular distance was set to 15°. The reverberation time (T60) was uniformly sampled between 200 and $500\,\mathrm{ms}$. Additive white Gaussian noise with 20 to $30\,\mathrm{dB}$ SNR was added to the mixture. The source separation algorithms operate on STFT signals with a discrete Fourier transformation (DFT) size of 512 and a shift of 128. The DC network consists of 2 BLSTM layers with 600 forward and 600 backward units and a final linear layer which yields $E = 20$ dimensional embeddings. The AM uses 40 Mel filterbank features extracted with an STFT with a DFT size of 256, a window size of 200 and a shift of 80.

To get a good impression of the system performance, we present results in terms of mean and standard deviation (STD) of BSS-Eval SDR gain [27], mean and STD of invasive SDR gain similar as in [28], PESQ gain [29], STOI gain [30] and finally WERs. The invasive SDR is the power ratio of applying the obtained mask or beamforming vector to the target speech image and the sum of all interference images.

First of all, we evaluated unsupervised spatial clustering with a cACGMM using a random initialization. Comparing row 1 and 7, we observe that although masking yields higher BSS-Eval SDR and STOI gains, the invasive SDR gain, perceptual evaluation of speech quality (PESQ) gain and foremost the WER are better for the MVDR beamforming approach. This can be explained by the observation that masking often leads to musical tones which are avoided with a time-invariant projection approach such as beamforming.

Next, we trained the DC system according to the proposed training scheme using masks exported according to row 1 or 7. The performance of the unsupervisedly trained DC (row 8, abbreviated as U-DC) system does not quite reach the performance of the supervised DC in row 10 (S-DC). Comparing the training times, the unsupervised approach takes at least twice as many training steps (here 500k steps, batch size 4) as the supervised DC system. The sometimes misleading cACGMM masks which result in more noisy gradients are a possible explanation.

Rows 3 and 9 summarize the results obtained when initializing the cACGMM with the U-DC result of row 2 or 8, respectively. When using masking, this results in a gain of $2.6\,\mathrm{dB}$ invasive SDR over the randomly initialized cACGMM in row 1. Also, when using beamforming, a gain of $2.6\,\mathrm{dB}$ invasive SDR can be observed. Comparing this with supervisedly trained DC in row 10 and the S-DC initialized cACGMM in row 11, it can be seen that the proposed approach outperforms both oracle baselines. Interestingly, the initialization is sufficient to address the permutation problem such that an additional alignment step is not necessary anymore. It is therefore valid to conclude that the main drawback of the teacher is its initialization. However, the performance of a cACGMM initialized with oracle ideal binary masks is not quite reached. Overall, the spatial mixture model initialized with the unsupervised DC (row 9) result yields a relative WER reduction of $26\,\%$ over the unsupervised teacher (row 7) and beats the best supervised system (row 11) by $2\,\%$ relative.

7. CONCLUSION

In this work, we demonstrated that a neural network-based blind source separation system can indeed be trained from scratch without any clean or parallel data. We opted for a training scheme with an unsupervised teacher and ended up with a student outperforming the teacher. This is of particular interest when addressing real data such as the CHiME 5 challenge recordings, where regular training is impeded due to the lack of well-matching targets. We see this work as an alternative/ complementary tool to train the frond-end entirely end-to-end or to better match real recordings and plan to extend the present work to CHiME 5 challenge recordings.

8. REFERENCES

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

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

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

[4] Y. Luo, Z. Chen, J. R. Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017.

[5] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *arXiv preprint arXiv:1810.01395*, 2018.

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

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

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

[9] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Interspeech*, 2018.

[10] Dong Yu, Xuankai Chang, and Yanmin Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[11] H. Seki, T. Hori, S. Watanabe, J. Le Roux, and J. R. Hershey, "A purely end-to-end system for multi-speaker speech recognition," *arXiv preprint arXiv:1805.05826*, 2018.

[12] Z. Chen, J. Droppo, J. Li, and W. Xiong, "Progressive joint modeling in unsupervised single-channel overlapped speech recognition," *arXiv preprint arXiv:1707.07048*, 2017.

[13] S. Settle, J. Le Roux, T. Hori, S. Watanabe, and J. R. Hershey, "End-to-end multi-speaker speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[14] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *International conference on Knowledge discovery and data mining (SIGKDD)*. ACM, 2006.

[15] N. Ito, S. Araki, and T. Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *European Signal Processing Conference (EUSIPCO)*. IEEE, 2016.

[16] N. Ito, S. Araki, T. Yoshioka, and T. Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *International Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.

[17] H. Sawada, S. Araki, and S. Makino, "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2007.

[18] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using Deep Clustering," in *Interspeech*, 2016.

[19] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[20] Z. Wang, J. Le Roux, and J. R. Hershey, "Alternative objective functions for deep clustering," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[21] L. Prechelt, "Automatic early stopping using cross validation: quantifying the criteria," *Neural Networks*, vol. 11, no. 4, pp. 761–767, 1998.

[22] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Transactions on audio, speech, and language processing*, vol. 18, no. 2, pp. 260–276, 2010.

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

[24] J. Heymann, L. Drude, and R. Haeb-Umbach, "Wide residual BLSTM network with discriminative speaker adaptation for robust speech recognition," in *Workshop on Speech Processing in Everyday Environments (CHiME16)*, 2016.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, et al., "The Kaldi speech recognition toolkit," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2011.

[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, pp. 943–950, 1979.

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

[28] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010.

[29] 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 *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2001.

[30] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timefrequency weighted noisy speech," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.