OPTIMIZATION OF SPEAKER-AWARE MULTICHANNEL SPEECH EXTRACTION WITH ASR CRITERION

Katerina Zmolikova[1], *Marc Delcroix*[2], *Keisuke Kinoshita*[2], *Takuya Higuchi*[2], *Tomohiro Nakatani*[2], *Jan Černocký*[1]

[1]Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czechia
[2]NTT Communication Science Laboratories, NTT Corporation, Kyoto, Japan

ABSTRACT

This paper addresses the problem of recognizing speech corrupted by overlapping speakers in a multichannel setting. To extract a target speaker from the mixture, we use a neural network based beamformer which uses masks estimated by a neural network to compute statistically optimal spatial filters. Following our previous work, we inform the neural network about the target speaker using information extracted from an adaptation utterance, enabling the network to track the target speaker. While in the previous work, this method was used to separately extract the speaker and then pass such preprocessed speech to a speech recognition system, here we explore training both systems jointly with a common speech recognition criterion. We show that integrating the two systems and training for the final objective improves the performance. In addition, the integration enables further sharing of information between the acoustic model and the speaker extraction system, by making use of the predicted HMM-state posteriors to refine the masks used for beamforming.

Index Terms— Speaker extraction, joint training, speaker adaptive neural network, beamforming, speech recognition

1. INTRODUCTION

Far-field speech recognition with the use of microphone arrays has become a topic of high interest. Although the robustness of speech recognizers advanced greatly, the presence of interfering speakers is still significantly hurting the performance. Traditionally, this problem has been tackled by methods such as Independent Component Analysis [1] or statistical model based systems [2, 3, 4, 5], which are used to preprocess the mixture before passing it onto the speech recognition system.

Recently, combining of neural networks with statistical beamforming has shown to be efficient for extracting a target speech corrupted by background noise [6, 7]. In such methods, the neural network estimates time-frequency masks distinguishing between the target and the interfering signal, which can be then used to compute the beamforming filters. This scheme has been successfully applied for denoising [6, 7, 8], dereverberation [9] or source separation [10, 11].

[1]Katerina Zmolikova is Brno Ph.D. Talent Scholarship Holder — Funded by the Brno City Municipality.

In our previous work [12, 13], we applied the neural network based beamformer to extract a target speaker from a mixture. To encourage the neural network to estimate masks corresponding to the target speaker, we used a method inspired from speaker adaptation [14] to inform the neural network about the speaker. The speaker information was obtained from an adaptation utterance — a segment of speech containing only the target speaker. For brevity, we will use the term SpeakerBeam to call the speaker-informed neural network for speaker extraction and multi-channel SpeakerBeam (MC-SpeakerBeam) its integration with the beamformer. In our method, the neural network learns both to extract useful speaker information from the adaptation utterance and to use this information to track the target speaker in the mixture. The usage of the additional speaker information avoids the dependency of the processing on number of speakers in the mixture and enables to follow the speaker through different processing segments, which contrasts with other speech separation techniques [15, 16].

In [13] we confirmed the efficiency of MC-SpeakerBeam as a front-end for speech recognition. There, the speaker extraction and the speech recognition system were used as two separate stages with different objectives. The lack of interconnection between these two systems is, however, sub-optimal. First, the objective function of the speaker extraction is rather arbitrary and its increase may not necessarily improve the accuracy of the speech recognition. Second, the front-end processing may benefit from having a higher level information from the acoustic model.

To overcome these shortcomings, we investigate integration of the two systems by training them for a common ASR objective. Optimizing the front-end enhancement jointly with the acoustic model has been explored for a denoising scenario in [17, 18, 19]. Notably, in [20, 21], the neural network based beamforming is combined with end-to-end training, keeping the statistically optimum beamforming and feature extraction in the processing chain and propagating the errors through these stages back to the mask-estimation network. Here, we follow the same pattern for the speaker extraction task and explore how the joint training criteria influence the accuracy in this more challenging scenario. A related study of the joint optimization of speech separation and recognition was done in [22, 23] in the framework of permutation invariant training [24]. Permutation invariant training is, however, substantially different from MC-SpeakerBeam

in terms of used objective criteria. Moreover, [22, 23] do not consider the use of multichannel signals and statistically optimum beamforming as a part of the network.

The use of a tightly integrated front-end and back-end optimized with a common objective function opens possibilities for sharing higher level information from the acoustic model to the speaker extraction system in an optimal manner. As such an example, we explore feeding-back HMM state posterior information into the mask-estimation network. The idea of using ASR-level information, such as state alignments or VAD, in speech enhancement was previously explored in [8, 25]. In these works, the information feedback was performed by alternating the enhancement and recognition stages. Here, we incorporate the feedback loop directly into the model and its optimization.

The remainder of the paper is structured as follows: In Section 2, we summarize the speaker-aware neural network based beamformer introduced in our previous work. In Section 3, we overview the overall processing chain, the joint optimization criterion and introduce the posterior feedback. Section 4 then reports experimental results and Section 5 concludes the paper.

2. SPEAKER-AWARE NEURAL NETWORK BASED BEAMFORMER (MC-SPEAKERBEAM)

In this section, we first describe the neural network based beamformer scheme proposed by [6] and then summarize the speaker-aware architecture introduced in our previous work [12, 13].

2.1. Neural network based beamforming

We model the signal received at i-th microphone in the Short time Fourier transform (STFT) domain as

$$Y_i(t,f) = X_i(t,f) + N_i(t,f), \tag{1}$$

where $i = 1 \dots I$ is the microphone index, $t = 1 \dots T$ is the time frame index, $f = 1 \dots F$ is the frequency-bin index, $Y_i(t,f)$ is the observed signal at the i-th microphone, $X_i(t,f)$ is the image of the speech signal of the target speaker and $N_i(t,f)$ denotes all the undesired signal — image of speech signals from interfering speakers and possibly additional noise.

The estimated image of the target signal at the reference microphone i_{ref} is obtained by the beamforming process as $\widetilde{X}_{i_{\text{ref}}}(t,f) = \mathbf{h}^{\text{H}}(f)\mathbf{Y}(t,f)$, where $\mathbf{h}^{\text{H}}(f)$ is a vector of beamforming coefficients and $\mathbf{Y}(t,f) = [Y_1(t,f) \dots Y_I(t,f)]^{\text{T}}$. The beamforming filters are computed using the Generalized Eigenvector beamformer (GEV) [26] from the spatial covariance matrices (SCM) of the desired and undesired signal — $\mathbf{\Phi}_{XX}(f)$, $\mathbf{\Phi}_{NN}(f)$, respectively. These matrices can be obtained as

$$\mathbf{\Phi}_{rr}(f) = \sum_{t=1}^{T} M_r(t,f)\mathbf{Y}(t,f)\mathbf{Y}^{\text{H}}(t,f), \tag{2}$$

where $r \in \{X, N\}$ and $M_r(t,f)$ denotes a time-frequency mask for the desired or undesired signal.

The time-frequency masks $M_r(t,f)$ are obtained from the output of a mask-estimation DNN (Mask-DNN) which is processing magnitude spectra of the observed signal $|\mathbf{Y}_i(t)|$. The Mask-DNN processes each channel separately and the final masks used for beamforming are obtained as an average over the channels. The objective function for training the network is the cross-entropy between the estimated masks and the ideal binary masks (IBM) which are computed from parallel clean and corrupted data.

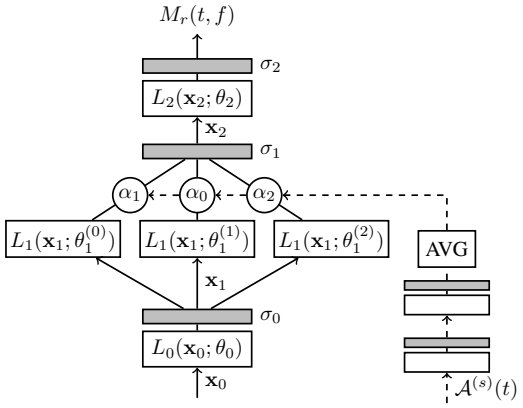

Fig. 1. Scheme of the speaker adaptive layer configuration.

2.2. Speaker-aware neural network

To enable the extraction of a target speaker, we have proposed informing the Mask-DNN with target speaker information derived from an adaptation utterance. In our previous work, we investigated different ways of incorporating the speaker information in the processing and evaluated the speaker adaptive layer approach to be the most suitable for this task [12]. In this approach, the speaker information in the form of an adaptation utterance is used to modify the behavior of one of the layers in the network so that the network can be adapted to extract speech only from the target speaker. This is achieved by factorizing the layer into several bases and combining the bases with weights computed from the speaker information.

The architecture is depicted in Figure 1. Denoting the index of the factorized layer as k, we can express the computation of the neural network as

$$\mathbf{x}_{n+1} = \begin{cases} \sigma_n(L_n(\mathbf{x}_n;\theta_n)) & \text{for } n \neq k, \\ \sigma_n(\sum_{m=0}^{M-1} \alpha_m^{(s)} L_n(\mathbf{x}_n;\theta_n^{(m)})) & \text{for } n = k, \end{cases} \tag{3}$$

where $\mathbf{x}_n$ is the input to the nth layer, $L_n(\mathbf{x},\theta)$ is the transformation computed by the nth layer parameterized by θ and σ_n is an activation function. For fully connected layers $\theta = \{\mathbf{W}, \mathbf{b}\}$ and $L(\mathbf{x},\theta) = \mathbf{W}\mathbf{x} + \mathbf{b}$, where $\mathbf{W}$ is a weight matrix and $\mathbf{b}$ is a bias vector. The output of the final layer of the network are the masks $M_r(t,f)$.

The weights $\alpha_m^{(s)}$ are computed by an auxiliary network which operates on features from an adaptation utterance from the target speaker. The auxiliary network includes an averaging operator on top of the last layer to summarize the frame-level activations into utterance-level weights that can represent the overall speaker characteristics [27]

$$\boldsymbol{\alpha} = \frac{1}{T_{\mathcal{A}}} \sum_{t=0}^{T_{\mathcal{A}}-1} z(\mathcal{A}^{(s)}(t)), \tag{4}$$

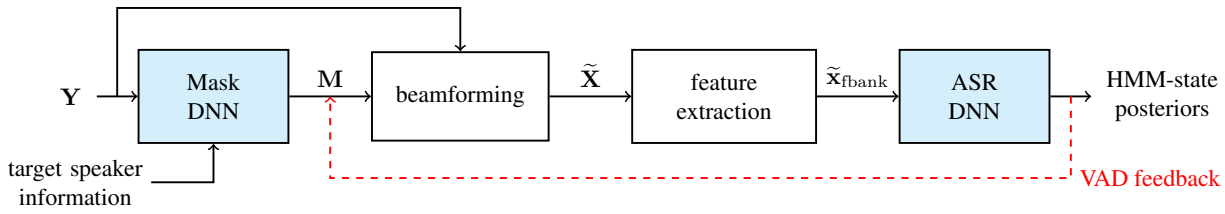

Fig. 2. The processing chain with the multichannel signal $\mathbf{Y}$ as input and the estimated HMM-state posteriors as the output. The trainable blocks are shown in blue. The red dashed line shows the VAD feedback as described in Section 3.2.

where $z(\mathcal{A}^{(s)}(t))$ are activations of the last layer of the auxiliary network, $\mathcal{A}^{(s)}(t)$ are features extracted from t-th frame of an adaptation utterance of a speaker s and $T_{\mathcal{A}}$ is the number of frames in the adaptation utterance. The auxiliary network is trained jointly with the main network. As the weights should pass information about the target speaker to the main network, the auxiliary network should learn to encapsulate speaker representation optimized directly for the target speaker extraction task without requiring direct supervision for the weights $\alpha_m^{(s)}$.

3. JOINT OPTIMIZATION

While in previous work, we treated the speaker extraction and the speech recognition as two separate stages, in this paper, we optimize both jointly with the final speech recognition level criterion. In addition, we further integrate the two stages by sharing information obtained from the acoustic model with the speaker extraction system, creating a feedback. In this section, we precise the overall processing chain and describe the posterior feedback.

3.1. Processing chain

The entire integrated system is depicted in Figure 2. It consists of four blocks — Mask-DNN (including the speaker adaptive layer and the auxiliary network), beamformer, feature extraction and the acoustic model. To optimize the front-end Mask-DNN w.r.t to the speech recognition level criterion, we need to propagate the error through the entire processing chain as follows:

$$\frac{\partial E}{\partial \boldsymbol{\theta}} = \frac{\partial E}{\partial \widetilde{\mathbf{x}}_{\text{fbank}}} \frac{\partial \widetilde{\mathbf{x}}_{\text{fbank}}}{\partial \widetilde{\mathbf{X}}} \frac{\partial \widetilde{\mathbf{X}}}{\partial \mathbf{M}} \frac{\partial \mathbf{M}}{\partial \boldsymbol{\theta}}, \tag{5}$$

where E is the cross-entropy between the estimated posteriors of HMM states and the state-alignments, $\widetilde{\mathbf{x}}_{\text{fbank}}$ are the features extracted from the estimated signal, $\widetilde{\mathbf{X}}$ is the STFT of the estimated signal, $\mathbf{M}$ are the estimated masks and $\boldsymbol{\theta}$ is the vector of the parameters of the Mask-DNN neural network.

The gradients $\partial E/\partial \widetilde{\mathbf{x}}_{\text{fbank}}$ and $\partial \mathbf{M}/\partial \boldsymbol{\theta}$ can be computed by backpropagation through standard neural network blocks. As the filterbank extraction can be formulated as a matrix multiplication, the gradient $\partial \widetilde{\mathbf{x}}_{\text{fbank}}/\partial \widetilde{\mathbf{X}}$ can be also computed by backpropagating through a neural network linear layer. For gradient $\partial \widetilde{\mathbf{X}}/\partial \mathbf{M}$ backpropagation through the GEV beamformer is needed — this step was thoroughly covered in [28, 29].

3.2. Posterior feedback

By integration of the speaker extraction with the ASR and performing their joint optimization, the speaker extraction system gets information from the acoustic model during the backpropagation step and can thus be fine-tuned to improve the speech recognition criterion. Here, we explore another form of information flow from the acoustic model back to the front-end, by using the predicted state posteriors to adjust the masks.

The option we explore is to interpret the output of the acoustic model as a simple voice activity detector (VAD) and weight the masks predicted in the front-end by the posterior probabilities of silence and non-silence states. The use of VAD information provided by ASR to fine-tune masks used for beamforming was successfully applied in [8], although there, the feedback loop was not tightly integrated into the neural network. The update of the masks happens as follows

$$M_X^{(\text{vad})}(t,f) = M_X(t,f) \sum_{\text{st}\in\text{nonsil}} p_{\text{st}}(t) \tag{6}$$

$$M_N^{(\text{vad})}(t,f) = M_N(t,f) \sum_{\text{st}\in\text{sil}} p_{\text{st}}(t), \tag{7}$$

where st labels the HMM states, $\text{sil}, \text{nonsil}$ are the sets of states corresponding to silence and non-silence phonemes respectively and $p_{st}(t)$ is the posterior probability of the state st for time frame t predicted by the acoustic model. The update of the masks forms a feedback loop which is incorporated in the chain during the training by doing two passes over the acoustic model (first pass without the VAD information and second pass with the modified masks $\mathbf{M}^{(vad)}$). During the back-propagation, we propagate the gradients from the second pass to the first pass.

4. EXPERIMENTS

4.1. Data

We evaluated the proposed methods using data created based on recordings from Wall Street Journal dataset [30]. The lists of utterances for the training, development and evaluation sets were taken from CHiME3 challenge [31]: 7138 utterances from 83 speakers in the training set, 410 utterances from 10 speakers in the development set and 330 utterances from 10 speakers in the evaluation set. For each utterance, we mixed an interference utterance from a different speaker within the same set with signal-to-interference ratio of $0\,\text{dB}$ on average.

To simulate the multichannel signals, we generated room impulse responses with the image method [32] to simulate a circular microphone array of 8 microphones, $20\,\text{cm}$ diameter in rooms with $\text{RT60}{=}0.2\,\text{s}$. The speakers were located at 1 or $1.5\,\text{m}$ meter distance from the microphone array, in angles from range 0 to 180°. For each mixture, we randomly chose an adaptation utterance from the target speaker (different than the utterance in the mixture) and used the image of this utterance at one of the microphones in the array. The length of the utterance is about $10\,\text{s}$ on average.

4.2. Settings

4.2.1. Mask estimation NN settings

The Mask-DNN consisted of 4 layers, i.e. one BLSTM layer, two fully connected layers with ReLU activation and one fully connected layer with a sigmoid activation. The number of neurons in the four layers was 512-1024-1024-512, respectively. The second layer was a speaker adaptive layer factorized into 30 sublayers. The auxiliary network predicting the weights $\boldsymbol{\alpha}$ was composed of two fully connected layers with 50 neurons and a ReLU activation and an output fully connected layer with a linear activation followed by the averaging operation.

4.2.2. Beamforming settings

For beamforming, we used GEV beamformer as described in Section 2.1 and [26]. The noise spatial covariance matrix was regularized by adding $\epsilon = 1e^{-3}$ to the diagonal to stabilize its inversion. The output signal was additionally processed by a single-channel postfilter [6, 26] to reduce the speech distortions. The beamforming was performed in the STFT domain computed with a $25\,\mathrm{ms}$ window and a $10\,\mathrm{ms}$ shift. This setting was chosen to be compatible with the ASR back-end.

4.2.3. ASR settings

The input features of the acoustic model consisted of 40 log Mel filterbank coefficients with a context extension window of 11 frames. The features were mean normalized per utterance. We used a simple DNN acoustic model, consisting of 5 fully connected hidden layers with 2048 nodes and ReLU activation functions for the ASR evaluation. The output layer had 2048 nodes corresponding to the HMM states. For training, we used HMM state alignments obtained from single channel noise-free training data using a GMM-HMM system.

4.3. Results

Table 1 shows the results of the experiments with and without the joint optimization of the *Mask-DNN* and acoustic model network (*ASR-DNN*). The Mask-DNN is initialized from a network trained for optimizing the cross entropy w.r.t. the IBMs, while the ASR-DNN from a network separately trained on clean single-speaker data. The first part of the table shows the results of recognizing clean single-speaker data, unprocessed mixtures and beamformed mixtures using oracle IBMs. These results bound the possible performance of our method.

The first row in the next part shows the WER of the separately trained system — both Mask-DNN and ASR-DNN are only initialized as described above. In the next experiment, we retrain the ASR-DNN with the enhanced training data to reduce the mismatch between the front-end and back-end. This experiment still corresponds to the separate training (the Mask-DNN is trained for mask-related criterion) and may serve as a baseline.

Retraining the joint network consisting of Mask-DNN and ASR-DNN for a common speech recognition criterion (last row of the second part of Table 1) leads to significant improvement on both development and evaluation sets. This shows that the IBMs used as a target of Mask-DNN in the separate training are quite distant objectives from the final speech recognition and optimizing for a higher-level target is beneficial.

Table 1. Results of joint training in terms of WER[%]. *Mask-DNN* refers to mask-estimation neural network in the speaker extraction front-end. *ASR-DNN* is a neural network in the acoustic model. Both networks initialized by separate training can be then retrained with ASR criterion.

	Mask-DNN retraining	**ASR-DNN retraining**	**dev**	**eval**
Single speaker	-	✗	5.00	3.92
Mixture	-	✗	75.85	75.44
IBM	-	✗	10.95	8.37
MC-SpeakerBeam	✗	✗	28.29	23.59
	✗	✓	26.26	21.76
	✓	✓	17.54	17.73
MC-SpeakerBeam +VAD feedback	✓	✓	**15.12**	**15.30**

To simplify the training process and refrain from the need of computing ideal masks from parallel data, we also experimented with jointly training the networks from random initialization. The results obtained in this case were notably worse than for joint training with mask and ASR DNNs initialized with pretrained networks (dev: 24.78, eval: 23.36). However, they are comparable to the results obtained with separate training only showing that the proposed network architecture can learn to extract the target speaker even when randomly initialized.

The last part of Table 1 shows the result of the experiment including feedback of VAD information predicted by the acoustic model back to the front-end. The experiments are done with the joint training of the Mask-DNN and ASR-DNN with both networks pretrained. We observe that scaling the masks by the voice activity factors computed from the posteriors improves the performance notably. Note that incorporating the feedback loop into the optimization is important, applying the VAD feedback only during the test-time did not bring performance improvement in our experiments.

5. CONCLUSION

In this paper, we explored integrating multichannel speaker extraction system with speech recognition and training them for a common criterion. This extends our previous work where we proposed the speaker extraction method based on neural network beamformer, additionally informed about the target speaker. We showed that training of the speaker extraction front-end together with the acoustic model improves the ASR performance. Additionally, we explored further sharing of information from the acoustic model back into the speaker extraction front-end. In future work, we plan to evaluate the robustness of this scheme to conditions with more realistic setting, such as meeting data.

6. ACKNOWLEDGMENT

The work was partly supported by Technology Agency of the Czech Republic project No. TJ01000208 "NOSICI", and by Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602".

7. REFERENCES

[1] A. Hyvärinen, J. Karhunen, and E. Oja, "Independent component analysis," 2001.

[2] 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.

[3] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 516–527, 2011.

[4] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382–394, 2010.

[5] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Acoustics, Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, 2010, pp. 241–244.

[6] 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*, 2016, pp. 196–200.

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

[8] Y.-H. Tu, J. Du, L. Sun, F. Ma, and C.-H. Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," 2017.

[9] J. Heymann, L. Drude, and R. Haeb-Umbach, "A generic neural acoustic beamforming architecture for robust multi-channel speech processing," *Computer Speech and Language*, vol. 46, no. Supplement C, pp. 374 – 385, 2017.

[10] L. Drude and R. Haeb-Umbach, "Tight integration of spatial and spectral features for bss with deep clustering embeddings," in *INTERSPEECH 2017, Stockholm, Sweden*, Aug 2017.

[11] T. Higuchi, K. Kinoshita, M. Delcroix, K. Zmolikova, and T. Nakatani, "Deep clustering-based beamforming for separation with unknown number of sources," in *INTERSPEECH 2017, Stockholm, Sweden*, 08 2017, pp. 1183–1187.

[12] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Speaker-aware neural network based beamformer for speaker extraction in speech mixtures," in *INTERSPEECH 2017, Stockholm, Sweden*, Aug 2017.

[13] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Higuchi, A. Ogawa, and T. Nakatani, "Learning speaker representation for neural network based multichannel speaker extraction," in *ASRU 2017*, Dec 2017.

[14] M. Delcroix, K. Kinoshita, C. Yu, A. Ogawa, T. Yoshioka, and T. Nakatani, "Context adaptive deep neural networks for fast acoustic model adaptation in noisy conditions," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5270–5274.

[15] 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*, 2016, pp. 31–35.

[16] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, "Deep neural networks for single-channel multi-talker speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, no. 10, pp. 1670–1679, 2015.

[17] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[18] M. L. Seltzer, B. Raj, and R. M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 5, pp. 489–498, Sept 2004.

[19] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5745–5749.

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

[21] X. Xiao, S. Zhao, D. L. Jones, E. S. Chng, and H. Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 3246–3250.

[22] D. Yu, X. Chang, and Y. Qian, "Recognizing multi-talker speech with permutation invariant training," *arXiv preprint arXiv:1704.01985*, 2017.

[23] Y. Qian, X. Chang, and D. Yu, "Single-channel multi-talker speech recognition with permutation invariant training," *CoRR*, vol. abs/1707.06527, 2017.

[24] 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*, 2017.

[25] H. Erdogan, J. R. Hershey, S. Watanabe, and J. L. Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 708–712.

[26] 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.

[27] K. Vesely, S. Watanabe, K. Zmolikova, M. Karafiat, L. Burget, and J. H. Cernocky, "Sequence summarizing neural network for speaker adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5315–5319.

[28] C. Böddeker, 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*, March 2017, pp. 171–175.

[29] C. Böddeker, P. Hanebrink, L. Drude, J. Heymann, and R. Haeb-Umbach, "On the computation of complex-valued gradients with application to statistically optimum beamforming," *CoRR*, vol. abs/1701.00392, 2017.

[30] J. Garofolo, "CSR-I (WSJ0) Complete LDC93S6A," `https://catalog.ldc.upenn.edu/ldc93s6a`, 1993.

[31] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding*. IEEE, 2015, pp. 504–511.

[32] E. A. P. Habets, "Room impulse response generator," Tech. Rep., Technische Universiteit Eindhoven, 2010.