ROBUST MASK ESTIMATION BY INTEGRATING NEURAL NETWORK-BASED AND CLUSTERING-BASED APPROACHES FOR ADAPTIVE ACOUSTIC BEAMFORMING

Ying Zhou, Yanmin Qian[†]

Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
{zhouy49@sjtu.edu.cn, yanminqian@tencent.com}

ABSTRACT

Recently the mask-based beamforming approach received tremendous interest and is widely studied for multi-channel noise robust automatic speech recognition (ASR). Among the known mask estimation models, the neural network based mask estimation approach has received the most attention, resulting in a competitive performance. However this approach still suffers from training-testing mismatch between the simulated training and real test data. This paper proposes a new unsupervised scheme that can utilize the real data during NN-based mask estimator training. The clustering-based approach is applied on the real data first to generate the soft masks, which are then taken as the labels for NN-mask modeling. Moreover, acoustic adaptation technologies are borrowed from usual back-end acoustic modeling to the front-end NN-mask based beamformer, further reducing the training-testing acoustic mismatch. The proposed methods are evaluated on the CHIME-4 dataset. Experimental results show that the mismatch can be reduced significantly by the proposed strategies, leading to relative $\sim$15.0% WER reduction compared to the conventional NN-mask beamforming for the real data under noisy conditions.

Index Terms— acoustic beamforming, time-frequency mask, deep neural network, adaptation

1. INTRODUCTION

In recent years, significant progress has been achieved in automatic speech recognition (ASR) due to the introduction of deep neural networks to acoustic modeling [1, 2]. The ASR systems based on deep neural networks, still perform poorly in many real-world far-field microphone scenarios. The main reasons for the poor performance are background interferences, e.g. additive noise, channel distortion and reverberation, which lower the SNR and degrade ASR performance. Acoustic beamforming [3, 4, 5, 6] has been shown as a helpful front-end approach to improve the system performance under these conditions. While conventional beamforming approaches usually rely on inaccurate prior knowledge, such as an array geometry or a plane wave assumption, time-frequency mask-based beamforming approaches do not need such extra knowledge, thus have been widely studied in recent years [7, 8, 9, 10, 11, 12, 13, 14, 15].

An accurate estimation of time-frequency masks is important in order to perform beamforming effectively. Most approaches for mask estimation can be divided into two categories: (1) the clustering-based approach [7, 9, 12] estimating masks in an unsupervised mode, and (2) the neural network-based (NN-based) approach [10, 11, 13, 14, 15] estimating masks with a trained neural network in advance. Both methods have achieved competitive results on some tasks, e.g. CHiME-4 dataset [16]. The NN-based method [10], although seemingly outperforming the clustering-based method [9], may easily cause a mismatch between the training and test conditions. The mismatch appears because a NN-based mask estimator can only be trained with the parallel simulated data while it is implemented for the real data. Therefore, the degree of similarity between the simulated and real data as well as the differences in acoustic backgrounds between the training and testing will have an impact on the performance of the NN-based approach for real applications. In contrast, the clustering-based mask estimator can be trained unsupervised and doesn't need simulated data for model construction, therefore no mismatch occurs.

Many attempts that take advantage of both NN-based and clustering-based approaches for multi-channel noise reduction have been proposed [17, 18]. [19] utilizes this integration idea to address the mismatch problem. In that method, initial masks, first estimated based on the NN-based approach, are utilized as the weight initializations for each cluster for the clustering-based mask estimator training. This work aims to reduce the training-testing mismatch, while focusing on optimizing the NN-based mask estimator. Two main strategies are proposed. First, to make the NN-mask model training applicable to real data, the soft mask labels of real data are generated by the clustering-based mask estimator, so that both simulated and real data can be utilized for NN-mask model optimization. The unsupervised mode of the clustering-based mask estimator enables us to generate additional real data for model training. Second, these NN-based acoustic beamformers can further be improved by commonly used speech recognition adaptation [20, 21, 22]. Acoustic adaptation can further reduce the mismatch significantly. The proposed approaches are evaluated on CHiME-4 dataset [16], leading to promising results.

The rest of the paper is organized as follows. Section 2 briefly reviews mask-based beamformers and two main techniques for mask estimation. Section 3 and Section 4 describe the proposed approaches in detail, including real data augmentation for NN-mask training and adaptive NN-based acoustic beamformer. The

[†]Yanmin Qian is the corresponding author and now he is with Tencent AI Lab, Tencent, Bellevue, WA, USA.

This work has been supported by the China NSFC projects (No. U1736202 and No. 61603252), and the Shanghai Sailing Program No. 16YF1405300. Experiments have been carried out on the PI supercomputer at Shanghai Jiao Tong University.

experimental results are discussed in Section 5 and conclusions are summarized in Section 6.

2. MASK-BASED BEAMFORMER

2.1. Mask-based beamforming

A noisy signal received from an array of M microphones is

$$\mathbf{Y}_{f,t} = \mathbf{X}_{f,t} + \mathbf{N}_{f,t}, \tag{1}$$

where $\mathbf{Y}_{f,t}$, $\mathbf{X}_{f,t}$ and $\mathbf{N}_{f,t}$ represent the short-time Fourier transforms (STFT) of the noisy signal, clean speech and noise respectively. f and t denote frequency bin and time frame index.

A beamformer is designed to recover the clean speech by applying a linear filter $\mathbf{w}_f^H$ to the observed noisy signal $\mathbf{Y}_{f,t}$. The enhanced signal is given by

$$\hat{s}_{f,t} = \mathbf{w}_f^H \mathbf{Y}_{f,t}, \tag{2}$$

where superscript H denotes conjugate transpose.

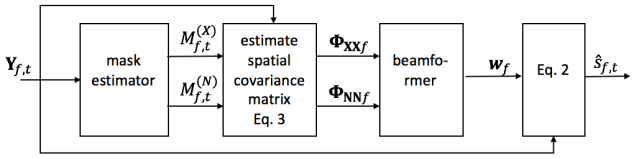

Fig. 1. Overview of mask-based beamforming.

Fig. 1 gives an overview of the mask-based beamformer. Firstly, a speech mask $M_{f,t}^{(X)}$ and a noise mask $M_{f,t}^{(N)}$ are estimated and then are used to calculate spatial covariance matrices of speech and noise respectively.

$$\mathbf{\Phi}_{\nu\nu f} = \sum_{t=1}^{T} M_{f,t}^{(\nu)} \mathbf{Y}_{f,t} \mathbf{Y}_{f,t}^H, \qquad \nu \in \{\mathbf{X}, \mathbf{N}\}. \tag{3}$$

These spatial covariance matrices are used to compute beamforming coefficients $\mathbf{w}_f$. In this paper, we consider the generalized eigenvalue (GEV) beamformer, which beamforming coefficients are given by:

$$\mathbf{w}_{gev\,f} = \arg\max_{\mathbf{w}_f} \frac{\mathbf{w}_f^H \mathbf{\Phi}_{\mathbf{XX}f} \mathbf{w}_f}{\mathbf{w}_f^H \mathbf{\Phi}_{\mathbf{NN}f} \mathbf{w}_f}. \tag{4}$$

Since the magnitude of each beamforming vector can be chosen arbitrary, the Blind Analytic Normalization (BAN) [23] technique can be used as a post-filter to reduce arbitrary distortions of the GEV beamformer.

2.2. Clustering based mask estimation

Complex Gaussian mixture model (CGMM) proposed in [9] has been show to be useful for mask based beamforming. This model assumes that each frequency can be clustered into two categories with different distributions, i.e. the noisy speech class and the noise-only class. A two Gaussian components CGMM is built to model each class's frequency distribution (one Gaussian represents one class). The estimated masks are the posteriors of each cluster at the corresponding time-frequency points. The parameters of CGMM are estimated in an unsupervised way.

2.3. Neural network based mask estimation

The neural network for mask estimation proposed in [10] is composed of a bidirectional long short-term memory (BLSTM) layer and three feed-forward layers. The training target is ideal binary masks (IBM) for speech and noise. In order to achieve the best result, manual optimization of the target masks is required , such as the two threshold $\text{th}_{\mathbf{X}}$ and $\text{th}_{\mathbf{N}}$ usage in [10]. In the test phase, the masks for each channel are estimated by the trained neural network separately and then combined to a single mask using a median operation.

3. REAL TRAINING DATA AUGMENTATION

This work focuses on the mismatch problem in NN-based mask estimation. Considering target mask labels are needed to train the mask estimation network, parallel speech data, i.e. original clean and simulated noisy speech, is required to prepare the needed mask labels. The problem is that the parallel speech data can be obtained by artificially generated data, but may not be available for real data applications. This also means, that the NN-based mask estimator can only be trained using simulated data, which may lead to a mismatch when facing real test data. In order to reduce this kind of mismatch, suitable methods for utilizing real training data are investigated. Therefore, the unsupervised CGMM-based mask estimator [9] is introduced, which processes the original real data in order to generate mask label estimates.

The left part of Fig.2 shows the scheme of real training data augmentation. For the simulated data, the mask labels are calculated from the corresponding clean speech and simulated noise. For the real data, soft masks $M_{f,t}^{(X)}$ and $M_{f,t}^{(N)}$ for speech and noise, respectively, are estimated by the CGMM on the real noisy data. Then, these soft masks are used to form mask labels for NN training. Note that the value of soft masks are real numbers within the range [0,1] and $M_{f,t}^{(X)} + M_{f,t}^{(N)} = 1$. Thus when training the NN-based mask estimator, two kinds of mask labels can be used: ideal binary mask (IBM) and ideal ratio mask (IRM).

The soft masks $M_{f,t}^{(X)}$ and $M_{f,t}^{(N)}$ can be directly used as IRM of the real data for speech and noise respectively. The IBM of the real data is defined as:

$$\text{IBM}_{\mathbf{X}} = \begin{cases} 1, & \frac{M_{f,t}{}^{(X)}}{M_{f,t}{}^{(N)}} > 10^{\text{th}_{\mathbf{X}}}, \\ 0, & \text{else.} \end{cases} \tag{5}$$

$$\text{IBM}_{\mathbf{N}} = \begin{cases} 1, & \frac{M_{f,t}{}^{(X)}}{M_{f,t}{}^{(N)}} < 10^{\text{th}_{\mathbf{N}}}, \\ 0, & \text{else.} \end{cases} \tag{6}$$

To achieve the best results, the two thresholds $\text{th}_{\mathbf{X}}$ and $\text{th}_{\mathbf{N}}$ are manually chosen to be different from each other. This procedure is similar to the IBM usage in [10].

For IBM, we use the binary cross-entropy cost described in [10] to train the network, and for IRM, the mean squared error (MSE) between the inferred mask prediction and the target IRM label is used as the loss function. Note that the structure of the neural network based mask estimator in this work is the same as the one in [10], including a BLSTM layer followed with three feed-forward layers.

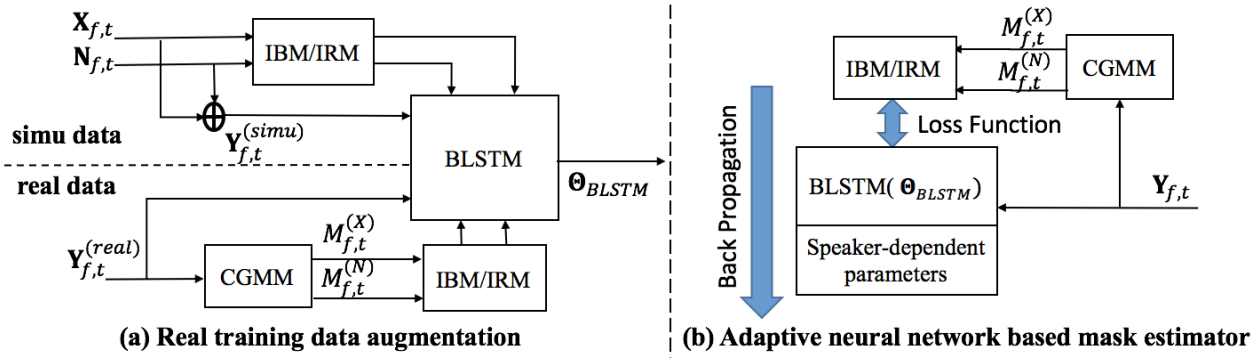

Fig. 2. Overview of the proposed schemes.

4. ADAPTIVE NEURAL NETWORK BASED ACOUSTIC BEAMFORMER

To deal with the mismatch arisen from acoustic backgrounds, e.g. variation in speakers and environments, adaptation techniques are usually applied on the acoustic model to improve the ASR performance. Motivated by the success from acoustic modeling, this work tries to implement adaptation on the mask estimators to obtain an adaptive acoustic beamformer.

The right part of Fig.2 shows the flow chart of our proposed adaptive neural network based beamformer. In the first pass of unsupervised adaptation, CGMM-based mask estimator is utilized to generate the soft masks ${M_{f,t}}^{(X)}$ and ${M_{f,t}}^{(N)}$ on the test data. Then, these soft masks can be used as labels for either IBM or IRM mode. After that, the original speaker-independent (or environment-independent) NN-mask estimator can be adapted on the speaker-dependent (or environment-dependent) test data in the second pass. Three types of NN-based adaptation approaches are investigated for the NN-mask based acoustic beamformer on speaker level in this work.

4.1. BLSTM re-training approach

Re-training [20] is a simple and intuitive adaptation technique that uses the predicted mask labels to re-train the entire neural network for each speaker. In this work, the entire network is re-trained, including a BLSTM layer and three feed-forward layers on speaker-level.

4.2. Linear input network (LIN) for BLSTM

Linear input network (LIN) [21] applies a linear transformation specified by the weight matrix $\mathbf{W}^{\text{LIN}} \in \mathbb{R}^{N_0 \times N_0}$ and the bias vector $\mathbf{b}^{\text{LIN}} \in \mathbb{R}^{N_0}$ to the input features vector $\mathbf{x} \in \mathbb{R}^{N_0}$ as

$$\mathbf{x}^{\text{LIN}} = \mathbf{W}^{\text{LIN}}\mathbf{x} + \mathbf{b}^{\text{LIN}}, \tag{7}$$

where N_0 is the size of input feature. In this work, the individual LIN linear layer for each speaker is added before the BLSTM network. The parameters of LIN layer are initialized with identity weight matrix and zero bias. LIN are updated in adaptation stage while the rest of the model is fixed.

4.3. Learning hidden unit contributions (LHUC) for BLSTM

For speaker m, a set of speaker-dependent parameters $\mathbf{r}_m$ are defined for the first feed-forward hidden layer [22] of the original trained BLSTM, where $\mathbf{r}_m \in \mathbb{R}^M$, M is the size of the hidden layer. Then the output of the first feed-forward layer $\mathbf{h}_m$ can be modified as

$$\mathbf{h}_m^{\text{LHUC}} = a(\mathbf{r}_m) \circ \mathbf{h}_m, \tag{8}$$

where $\circ$ is an element-wise multiplication, and $a(\cdot)$ is a sigmoid with amplitude 2 that constrains the range of $\mathbf{r}_m$ within [0,2]. $\mathbf{r}_m$ is initialized with zero value. In this case, $a(\mathbf{r}_m)$ is set to 1.0, and the modified model is equivalent to original trained model. $\mathbf{r}_m$ is updated discriminatively by back propagating the error with the original BLSTM parameters fixed.

5. EXPERIMENTS

The CHIME-4 challenge dataset [16] is used to evaluate the proposed approaches. The dataset consists of real and simulated audio data of prompts taken from the 5k WSJ0-Corpus [24] with four different environmental noise recordings. In this work we only consider the 6ch-track. The training set consists of 1600 real and 7138 simulated utterances. The development and evaluation sets comprise 3280 and 2640 utterances, respectively, both including simulated and real data.

5.1. Experimental Setups

In our experiments, different mask estimation approaches are compared using the same GEV+BAN beamformer (see Section 2.1). The same standard back-end ASR provided by the 4th CHiME challenge is directly used, which is composed of a DNN-HMM acoustic model trained with sMBR and a combination of a 5-gram and recurrent neural network language model rescoring [16]. Three traditional mask estimators served as baselines are built first, shown in Table 1, including a CGMM estimated from each test utterance, and two BLSTM trained on all simulated training data with IBM or IRM labels. The toolkit used in [10] is utilized to train the BLSTM mask estimator. It is noted that all BLSTM networks are initialized with same parameters for a cogent comparison of our proposed approaches. The results show that the baseline NN-mask based beamforming outperforms the CGMM-based method on the real data. Further, utilizing IRM mask labels leads to a slightly better result compared to the IBM usage.

Table 1. Baseline: Average WER (%) for using different mask estimators on the CHiME-4 development set (dt_05) and evaluation set (et_05).

Mask estimator	dev		eval	
	real	simu	real	simu
CGMM	5.47	4.76	8.66	5.72
BLSTM-IBM	5.06	5.03	7.56	6.36
BLSTM-IRM	4.63	4.97	6.79	6.59

5.2. Evaluation on real training data augmentation

Real training data augmentation for NN-based mask modeling is evaluated. The real data in the CHiME-4 training set is firstly processed with CGMM-based mask estimator, so that these real data with the soft masks can be pooled with simulated data together for NN-mask model training. Both IBM and IRM labels were evaluated on the experiments, which results are displayed in Table 2. It is observed that introducing real data with CGMM-estimated soft masks for the BLSTM-mask beamformer modeling outperforms the baselines trained only on the simulated data. The improvement on the real test condition is particularly larger, while the gain on the simulated test condition is small or even slightly decreased. Since the mismatch between the original simulated training and test conditions is small, adding real data in training actually increases the mismatch for the simulated test condition. In contrast, adding real training data for model optimization can reduce the mismatch on the real test condition, thus contributes to a better performance for real applications. Consistent with the results in Table 1, the IRM usage still outperforms the IBM usage. Based on these results, augmented real training data with IRM labels are used in the following experiments.

Table 2. Average WER (%) comparison of different training data usages for BLSTM-mask beamformer.

Target	Training-data	dev		eval	
		real	simu	real	simu
IBM	SIMU	5.06	5.03	7.56	6.36
	SIMU + REAL	4.73	4.96	6.88	**6.24**
IRM	SIMU	4.63	4.97	6.79	6.59
	SIMU + REAL	**4.59**	**4.83**	**6.51**	6.61

5.3. Evaluation on adaptive neural network based acoustic beamformer

For adaptation experiments, we compared the performance of three different adaptation methods, i.e. re-training, LIN and LHUC as described in Section 4. The CGMM-based mask estimator is used in the first pass to generate the soft mask labels for the test data. After obtaining the soft labels, the NN-mask model is adapted for each speaker. The results are illustrated in Table 3. It shows that the proposed adaptive NN-based beamformer, using LIN or LHUC structure, can obtain significant gains on almost all test conditions compared to the normal unadaptive beamformer. In contrast, the re-training approach's ability to enhance the performance is limited. This promising result further demonstrates, that the acoustic mismatch can be significantly reduced by the proposed adaptive neural network based front-end beamformer. The proposed new beamformer using both data augmentation and adaption can obtain relative $\sim$15.0% WER reduction compared to the conventional NN-mask beamforming [10] for the real data.

Table 3. Average WER (%) comparison of different adaptive BLSTM-mask beamformers.

Adaptation	dev		eval	
	real	simu	real	simu
—	4.59	4.83	6.51	6.61
retraining	4.59	4.93	6.42	6.60
LIN	**4.28**	4.83	6.38	6.27
LHUC	4.35	**4.70**	**6.17**	**6.22**

Finally, an investigation whether the adaptation benefits on the front-end beamformer is exclusive to the usual adaptation gain on the back-end acoustic modeling is carried out. The proposed adaption technique is performed on both front-end beamformer and back-end acoustic model. Results are shown in Table 4. It shows that significant improvements can be obtained no matter adaption is performed on front-end or back-end. The adaptation in these two stages are not exclusive. Moreover, the system applied the adaptation on both beafmormer and acoustic model can get an additional gain on all test conditions.

Table 4. Average WER (%) comparison of adaptation technologies applied on front-end beamformer and back-end acoustic modeling.

BF-adapt	AM-adapt	dev		eval	
		real	simu	real	simu
$\times$	$\times$	4.59	4.83	6.51	6.61
$\times$	$\surd$	4.11	4.12	5.48	4.98
$\surd$	$\times$	4.35	4.70	6.17	6.22
$\surd$	$\surd$	**3.83**	**3.98**	**5.16**	**4.83**

6. CONCLUSION

Inspired by the training-testing mismatch problem for the NN-mask based beamforming algorithm arisen from the simulated training and real testing data, several strategies are proposed, including real training data augmentation and adaptive NN-mask beamformer development. In order to enable the system to utilize the real training data in NN-mask estimator training, the unsupervised clustering-based model is applied on the real data first to generate the soft masks, so that the real data can be pooled with the simulated data for NN-mask modeling. Adding real training data can reduce the training-testing mismatch obviously. Moreover, for the first time, adaptation techniques are performed on this front-end NN-mask based beamformer, leading to a further improvement of the system performance. Another finding is that the adaptation benefits on the front-end beamformer and back-end acoustic modeling are not exclusive, thus by adapting the system on both two stages, an additional gain on all testing condition on CHiME-4 has been demonstrated.

7. REFERENCES

[1] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[2] Frank Seide, Gang Li, and Dong Yu, "Conversational speech transcription using context-dependent deep neural networks.," in *Interspeech*, 2011, pp. 437–440.

[3] Barry D Van Veen and Kevin M Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[4] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[5] Mehrez Souden, Jacob Benesty, and Sofiène 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.

[6] Xiong Xiao, Shinji Watanabe, Hakan Erdogan, Liang Lu, John Hershey, Michael L Seltzer, Guoguo Chen, Yu Zhang, Michael Mandel, and Dong Yu, "Deep beamforming networks for multi-channel speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5745–5749.

[7] Dang Hai Tran Vu and Reinhold 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*. IEEE, 2010, pp. 241–244.

[8] Mehrez Souden, Shoko Araki, Keisuke Kinoshita, Tomohiro Nakatani, and Hiroshi Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[9] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5210–5214.

[10] Jahn Heymann, Lukas Drude, and Reinhold 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.

[11] Hakan Erdogan, John R Hershey, Shinji Watanabe, Michael I Mandel, and Jonathan Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks.," in *INTERSPEECH*, 2016, pp. 1981–1985.

[12] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Complex angular central gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Signal Processing Conference (EUSIPCO), 2016 24th European*. IEEE, 2016, pp. 1153–1157.

[13] Christoph Boeddeker, Patrick Hanebrink, Lukas Drude, Jahn Heymann, and Reinhold 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.

[14] Lukas Pfeifenberger, Matthias Zöhrer, and Franz Pernkopf, "Dnn-based speech mask estimation for eigenvector beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 66–70.

[15] Xiong Xiao, Shengkui Zhao, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "On time-frequency mask estimation for mvdr beamforming with application in robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3246–3250.

[16] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[17] Hendrik Meutzner, Shoko Araki, Masakiyo Fujimoto, and Tomohiro Nakatani, "A generative-discriminative hybrid approach to multi-channel noise reduction for robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5740–5744.

[18] Aditya Arie Nugraha, Antoine Liutkus, and Emmanuel Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.

[19] Tomohiro Nakatani, Nobutaka Ito, Takuya Higuchi, Shoko Araki, and Keisuke Kinoshita, "Integrating dnn-based and spatial clustering-based mask estimation for robust mvdr beamforming," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 286–290.

[20] Hank Liao, "Speaker adaptation of context dependent deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7947–7951.

[21] Joao Neto, Luís Almeida, Mike Hochberg, Ciro Martins, Luis Nunes, Steve Renals, and Tony Robinson, "Speaker-adaptation for hybrid hmm-ann continuous speech recognition system," 1995.

[22] Pawel Swietojanski and Steve Renals, "Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models," in *Spoken Language Technology Workshop (SLT), 2014 IEEE*. IEEE, 2014, pp. 171–176.

[23] Ernst Warsitz and Reinhold 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] John Garofalo, David Graff, Doug Paul, and David Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.