ON TIME-FREQUENCY MASK ESTIMATION FOR MVDR BEAMFORMING WITH APPLICATION IN ROBUST SPEECH RECOGNITION

Xiong Xiao[1], *Shengkui Zhao*[2], *Douglas L. Jones*[2], *Eng Siong Chng*[1,3], *Haizhou Li*[1,3,4,5]

[1]Temasek Laboratories, Nanyang Technological University (NTU), Singapore.
[2]Advanced Digital Sciences Center, Singapore.
[3]School of Computer Science and Engineering, NTU, Singapore.
[4]Department of ECE, National University of Singapore, Singapore.
[5]Institute for Infocomm Research, A*STAR, Singapore.

xiaoxiong@ntu.edu.sg

ABSTRACT

Acoustic beamforming has played a key role in the robust automatic speech recognition (ASR) applications. Accurate estimates of the speech and noise spatial covariance matrices (SCM) are crucial for successfully applying the minimum variance distortionless response (MVDR) beamforming. Reliable estimation of time-frequency (TF) masks can improve the estimation of the SCMs and significantly improve the performance of the MVDR beamforming in ASR tasks. In this paper, we focus on the TF mask estimation using recurrent neural networks (RNN). Specifically, our methods include training the RNN to estimate the speech and noise masks independently, training the RNN to minimize the ASR cost function directly, and performing multiple passes to iteratively improve the mask estimation. The proposed methods are evaluated individually and overally on the CHiME-4 challenge. The results show that the proposed methods improve the ASR performance individually and also work complementarily. The overall performance achieves a word error rate of 8.9% with 6-microphone configuration, which is much better than 12.0% achieved with the state-of-the-art MVDR implementation.

Index Terms— beamforming, robust speech recognition, time-frequency mask, neural networks, long short-term memory.

1. INTRODUCTION

Despite the rapid progress in automatic speech recognition (ASR) due to the use of deep neural networks (DNN) and large training corpus [1], recognizing speech reliably from far-field recordings is still a challenging task. Robust recognition of far-field recorded speech has attracted significant amount of attention in the speech community. Several benchmarking tasks have been devoted to evaluate the progress of this field, e.g. the REVERB challenge [2], CHiME-3/4 [3, 4] challenges, and the AMI meeting transcription task [5].

Microphone array beamforming is one of the most effective approaches to improve the robustness of far-field ASR. Due to the simultaneous recording of speech signals at different locations, a microphone array provides an extra spatial dimension in which the target signal and interference could be separable if they come from different directions. Several beamforming techniques have been studied for the robust ASR applications. The delay-and-sum (DS) beamforming [6] delays the microphone signals according to the time difference of arrival (TDOA) to synchronize and add all the channels together to enhance the target signal. As the target signal is added constructively while the interference signals are added destructively, the signal-to-noise ratio (SNR) is improved. In DS, the TDOA or steering vector needs to be estimated prior to the beamforming. A more advanced beamforming is the minimum variance distortionless response (MVDR) beamforming [7, 8], which also makes use of the spatial information of the interference. The MVDR beamforming minimizes the power of the beamformed signal, while keeping the gain of the target direction at unity to preserve the target signal.

A critical component of the MVDR beamforming is the estimation of the spatial statistics of the target speech and noise. The spatial information of the target signal is included in the speech spatial covariance matrix (SCM) while the spatial and temporal information of the noise is included in noise SCM. In practice, the noise SCM can be estimated from the speech absent time-frequency (TF) bins, while the speech SCM can be estimated from speech dominant TF bins. Hence, the problem becomes the accurate estimation of a TF mask which specifies whether a TF bin is speech dominant or noise dominant. Several methods have been used for the TF mask estimation. In [10], a complex Gaussian mixture model [11] is used, where each source is represented by one Gaussian at each TF bin. Recently, the studies in [12, 13, 14, 15] propose to use the long short-term memory (LSTM) recurrent neural networks (RNN) to predict the TF masks and significantly improved the ASR performance in the CHiME-3 challenge. In these studies, the LSTM is trained to predict the ideal binary masks (IBM) of the speech (and noise). The IBMs are estimated from the ground truth speech and noise recordings with manually optimized thresholds during training stage.

In this paper, we investigate several ways to train the LSTM mask predictor for further improving ASR performance. Motivated by our recent success of joint optimization of acoustic model and a beamforming network [16, 17], we propose to refine the LSTM mask predictor by minimizing the ASR's cost function directly. This approach has two advantages. First, the mask predictor refined with ASR cost will be more optimal for the ASR task in theory than the mask predictor trained with IBMs. Second, the manual tuning in creating the IBMs (e.g. the setting of thresholds) can be avoided, which leads to more reliable and reproducible results. In addition, we also investigate the multiple passes of mask estimation and find it further boosts the ASR performance significantly.

The rest of the paper is organized as follows. In section 2, we present the techniques for robust and optimal estimation of TF masks for the MVDR beamforming. In section 3, experimental settings and results are presented and discussed. In section 4, conclusions are drawn and future works are discussed.

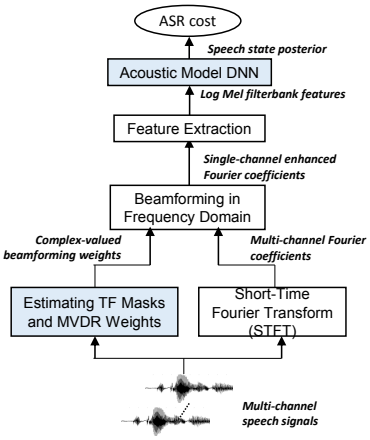

Fig. 1. A computational graph that covers beamforming, feature extraction, and acoustic modeling. The shaded boxes represent trainable modules, while white boxes are deterministic modules.

2. TIME-FREQUENCY MASK ESTIMATION

2.1. System Overview

The proposed beamforming system for ASR is shown in Fig. 1. There are two trainable modules in the diagram (shown in blue color). One is the mask estimation and weight prediction module and the other is the acoustic model. The rest modules are deterministic. The mask estimation module is trained to directly minimize the ASR cost function. This can be achieved by training the mask predictor using the gradients back-propagated from the ASR cost layer, such as cross entropy of frame-level phone classification. Although we can optimize the mask predictor and acoustic model simultaneously, in this work, we freeze the acoustic model that is already trained and only update mask predictor.

2.2. Mask Estimation

The details of the mask estimation subnet is shown in Fig. 2. From the multi-channel input signals, we first extract the short time Fourier transform (STFT) coefficients and log power spectrum for all channels. The complex-valued STFT coefficients will be used for SCM estimation and beamforming, while the magnitude features are used for mask prediction using an LSTM. Although phase information between channels could be useful for estimating TF masks, we choose to work on single channel mask estimation in this work, following the practice in [13]. One advantage of single channel mask prediction is that the LSTM mask predictor can be applied to all kinds of array configurations.

The network structure of LSTM based mask predictor is shown in Fig. 3. The input features are the log power spectrum of current frame, concatenated with its delta and acceleration versions. Hence, the dimension of the input feature vector will be $3K$ where K is the number of unique frequency bins. A single LSTM layer with H memory cells are used. The hidden activations of the LSTM layer $\mathbf{h}_t$ are mapped to the speech mask vector $\mathbf{m}_t^s$ and noise mask vector $\mathbf{m}_t^n$ using two different affine transforms. Sigmoid functions are applied after affine transforms to ensure that the predicted masks contain values between 0 and 1. The speech and noise mask estimations share the same LSTM layer to reduce model parameters.

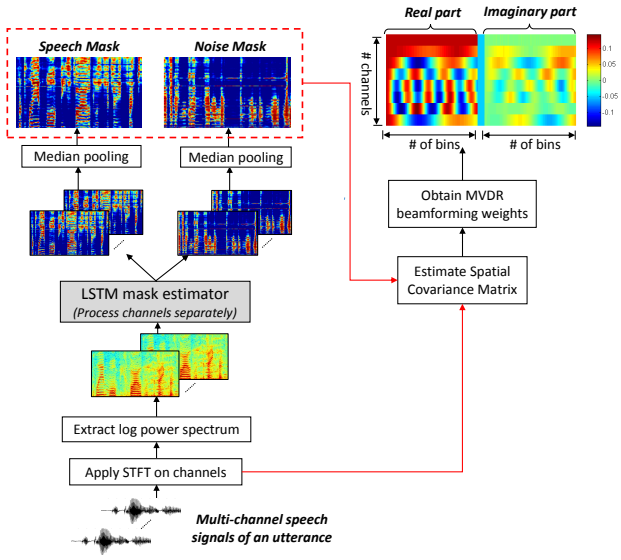

Fig. 2. Details of mask prediction and MVDR weight estimation.

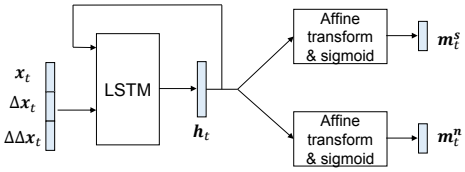

Fig. 3. Structure of the LSTM based mask predictor.

As a pair of speech and noise masks are generated from each input channel, we pool the masks over the channels to obtain one pair of final masks as shown in the upper left part of Fig. 2. Four types of pooling functions are compared, including mean, median, min, and max. The pooling functions are applied to each TF bin independently.

2.3. MVDR Weights Computation

With the estimated speech and noise masks, we estimate the SCMs of speech and noise for each frequency bin as follows.

$$\phi_{ss}(f) = \frac{\sum_{t=1}^{T} \hat{m}_t^s(f)\mathbf{y}_t(f)\mathbf{y}_t^H(f)}{\sum_{t=1}^{T} \hat{m}_t^s(f)} \quad (1)$$

$$\phi_{nn}(f) = \frac{\sum_{t=1}^{T} \hat{m}_t^n(f)\mathbf{y}_t(f)\mathbf{y}_t^H(f)}{\sum_{t=1}^{T} \hat{m}_t^n(f)} \quad (2)$$

where t and f are the time frame and frequency indices, respectively. T is the number of frames in an utterance. $\mathbf{y}_t(f) = [y_{t,1}(f), ..., y_{t,J}(f)]^T$ is the vector of observed STFT coefficients of all the J microphones. $\hat{m}_t^s(f)$ and $\hat{m}_t^n(f)$ are the predicted speech and noise masks, respectively. In Equations (1-2). we are estimating the average SCMs of each sentence. It is also possible to estimate online SCMs by averaging over a moving window instead of the whole sentence. The online estimation is left for future study.

After the SCMs are estimated, we can compute the MVDR beamforming weights as $\mathbf{w}(f) = \frac{\phi_{nn}^{-1}(f)\phi_{ss}(f)\mathbf{u}}{\mathrm{Tr}(\phi_{nn}^{-1}(f)\phi_{ss})}$, where $\mathbf{u}$ is a column vector whose elements are all 0's except that the element

corresponding to the reference channel (fixed to the first channel in this study) being 1 [18]. Tr$(\cdot)$ denotes the trace of a matrix.

2.4. Training Cost Functions

Two types of cost functions are used to train the LSTM mask predictor. The first cost function is the mean square error (MSE) between the predicted masks and the IBMs. The IBMs of speech and noise are obtained as follows

$$m_t^s(f) = 1[\beta_t(f) > \theta_s] \quad (3)$$
$$m_t^n(f) = 1[\beta_t(f) < \theta_n] \quad (4)$$

where $1[c]$ is an indicator function that equals to 1 if c is true and 0 otherwise. $\beta_t(f)$ is the oracle local SNR at frame t and frequency bin f and can be obtained for simulated data for which we have separated clean and noise signals. In previous works such as [13], the thresholds θ_s and θ_n are set differently such that only confident speech (noise) TF bins are marked as speech (noise) in the IBM. This is to reduce the false alarm rate of mask estimation and produce reliable SCM estimation. However, the optimal setting of the thresholds may require manual tuning. In this work, we use the training on IBMs only to initialize the mask predictor. Hence, we first set $\theta_s = \theta_n = 0dB$ although it is not optimal for ASR. We then apply the fine tuning of mask predictor on ASR to optimize the mask prediction for the ASR task.

After the MSE training of mask predictor, we plug the pretrained model into the graph in Fig. 1 and fine tune the LSTM with ASR cost function (cross entropy of frame level senone classification). As only speech masks are predicted in the MSE training stage, we need to initialize the affine transform for noise mask (see Fig. 3) as $\mathbf{A}^n = -\mathbf{A}^s$ and $\mathbf{b}^n = -\mathbf{b}^s$, where $\mathbf{A}^s$ and $\mathbf{b}^n$ are the linear transform and bias vector for speech mask prediction. Such initialization makes the initial noise mask and speech mask sum to 1 for all TF bins.

3. EXPERIMENTS

We evaluate the performance of the proposed methods on the CHiME-4 speech recognition challenge [4]. The official ASR baseline system is used except for two changes. First, to facilitate learning in Fig. 1, Mel-frequency cepstral coefficients (MFCC) features with feature space maximum likelihood linear regression (fMLLR) speaker adaptation is replaced with log Mel filterbank features. This is because it is difficult to pass the gradient through the speaker adaptation stage as the fMLLR transforms are estimated dynamically on the features, which depends on the beamforming module. The filterbank feature vectors (40 dimensions) are processed by utterancewise mean normalization. No pre-emphasis or DC removal is applied. Delta and acceleration features are appended and then 11 frames of feature vectors are cascaded to form the input for the DNN acoustic model. Second, the acoustic model is trained from data pooled from all the 6 channels of the training data, as we found that training with all channels increase the robustness of the acoustic model on evaluation data.

The LSTM mask predictor contains only one hidden layer with 1024 memory cells. 512-point fast Fourier transform is used for beamforming, resulting in 257 dimensional mask vectors for each frame. For the fineing with ASR cost function, we plug in the MSE trained mask predictor and cross entropy (CE) trained acoustic model into the graph in Fig. 1, and only update the mask predictor but freeze the acoustic model. In this way, we can optimize the mask predictor specifically for an existing acoustic model.

Table 1. Recognition word error rate (WER %) on the CHiME-4 6-channel track. "Split Mask" specifies whether we estimate speech and noise masks separately. LM: "3" means trigram, "5" means 5-gram, while "R" is RNN LM rescoring.

Sys No	Settings						Dev		Eval	
	#ch for mask	ASR cost	Split Mask	Pooling	#Pass	LM	Real	Simu	Real	Simu
1	1-channel track						12.4	14.8	21.6	22.0
2	Delay-and-sum (BeamformIt)						8.2	9.4	13.6	14.2
3	Traditional MVDR						7.6	6.6	12.0	8.2
4	Use only the first channel	No	No	No	1	3	8.3	7.1	12.8	19.5
5		Yes			1		7.3	6.4	10.9	15.2
6					3		6.4	6.1	9.4	11.1
7			Yes		1		6.5	6.1	10.1	11.9
8					3		6.1	6.0	9.0	9.9
9	Estimate masks for all 6 channels, then pool the masks			max	1		6.6	6.0	10.2	10.0
10				min	1		6.6	6.0	10.3	8.9
11				mean	1		6.4	5.9	9.8	9.2
12				median	1		6.2	6.0	9.5	8.9
13					3		6.1	5.9	8.9	9.6
14					3	5	4.8	4.9	7.4	7.9
15					3	R	4.1	4.3	6.3	6.9

3.1. Baseline Results

The ASR performance of the proposed methods and baselines are shown in Table 1 in terms of word error rates (WER). Two baseline systems are used, one is official CHiME-4 baseline which is a delay-and-sum (DS) beamformer implemented in the BeamformIt toolkit [19]. The BeamformIt uses Viterbi decoding to track time difference of arrival of target signal and also estimates the gains of the target signal in the channels. It is observed from the table that the DS beamforming (system 2 in the table) produces significant improvement over the 1-channel track, e.g. WER on real eval set reduced from 21.6% to 13.6%. The MVDR beamforming (system 3) is found to outperform the DS beamforming of BeamformIt. We used the MVDR beamforming proposed in [20] which uses eigen-decomposition based signal gain estimation. Note that the MVDR makes use of the noise information (both estimated within the test sentence and the 0.5s noise immediately before the test sentence) while the DS beamforming does not.

From system 4 onwards in Table 1, we examine the results obtained by MVDR beamforming using masks prediction. System 4 predicts speech mask from only the first channel of the array and MSE trained LSTM. The results of system 4 are slightly worse than the MVDR baseline, except a big increase in WER for the simulated eval set, which is due to the fact that the first channel of this test set is much noisier than other channels, hence resulting in poorly predicted mask and beamforming.

3.2. Fine Tuning Mask Predictor with ASR cost

We first examine whether refining the LSTM mask predictor by using the ASR cost function improves the performance of ASR. By comparing system 5 and 4, it is observed that the refinement improves the performance in every test set significantly. Note that we still only predict speech mask in system 5, so the number of free parameters are the same for both systems and the improvement is only from the use of ASR cost function.

We also examine the masks generated by system 4 and 5 in Fig.4 (a-b). For both systems, the noise and speech masks sum to 1 for every TF bin. It is observed that both systems produce reasonable speech masks, except that the baby crying (high-pitch harmonics from frame 300) also appears in the speech mask. System 5 (refined by ASR cost) predicts slightly better mask than system 4 in that the baby's crying is largely removed in the speech mask. We also observe that the system 5 generally assigns more TF bins to speech mask and less bins to noise mask, as compared to system 4 (the MSE training with IBMs). This effect is simiar to using a smaller threshold θ_s in (3).

3.3. Separate estimation of noise and speech masks

We then examine the effect of estimating the speech and noise masks separately. By comparing system 5 (only estimate speech mask) and 7 (estimate two masks separately), it is observed that the separate estimation of the two masks is beneficial for ASR. Comparing the masks generated from system 5 and 7 in Fig. 4, it is observed that the separately estimated speech and noise masks are more conservative than their counterparts produced by system 5. Those TF bins containing both significant energy of speech and noise do not appear in both the speech and noise masks. Such masks could result in purer estimation of speech and noise SCMs and hence better beamforming performance. In [12, 13], it is also encouraged to predict more conservative masks by setting θ_s larger than θ_n in (3-4). An advantage of using the ASR cost function for training is that we don't need to set any thresholds manually. Fig. 4 shows the clean, noisy, and beamformed (by system 7) log spectrum. Significantly improved spectrum is observed. The power of the baby crying is also reduced.

3.4. Multiple passes of mask estimation

Next, we investigate the effect of multiple-pass mask estimation on ASR performance. As the input speech is significantly improved by beamforming, it is reasonable to use the beamformed signal for more accurate mask estimation. Such process could repeat several times and the masks could be improved iteratively. Note that the same LSTM mask predictor is used in all the passes. By comparing systems 5 to 6, and systems 7 to 8, we observe that using 3 passes of mask estimation significantly boosts the performance of ASR, producing up to about 20% relative WER reduction on various test sets.

3.5. Pooling of masks of multiple channels

Finally, we compare different pooling strategies for combining the masks predicted from all channels. By comparing system 7 to system 9-12, we observe that most pooling functions improve the performance of ASR, with median pooling be the best and max pooling the worst. In [13], median pooling is also used.

We also combine all the techniques investigated above in system 13. Comparing system 13 to system 12, performance is improved except for eval simu set. Comparing system 13 to system 8, performance is marginally improved for all test sets. The results show that the multi-pass mask estimation and mask pooling, both improving the ASR independently, are not working well with each other. One reason could be that the mask predictor are trained without the knowledge of multi-pass mask estimation or pooling. More optimal way is to include these steps in the graph of Fig. 1 during training. We will investigate this direction in the future. The best WER obtained by mask based MVDR is 8.9% for real eval set. This represent a 3.1% absolute WER reduction compared to conventional MVDR

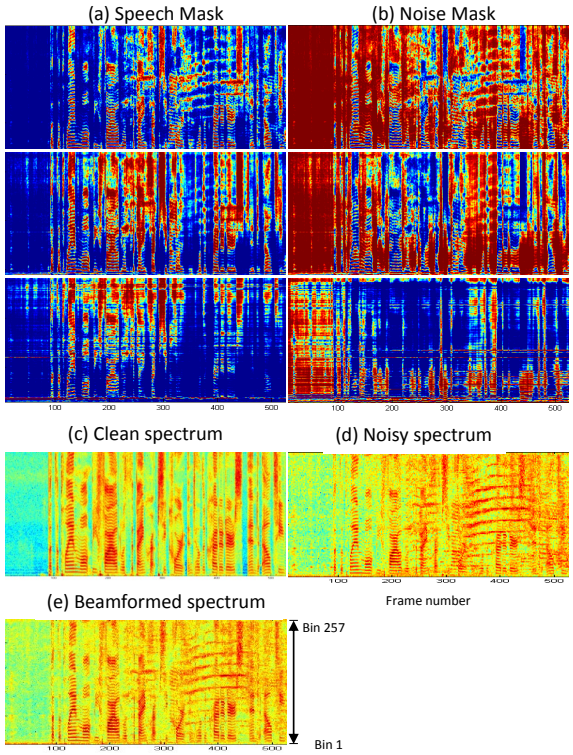

Fig. 4. Predicted masks for utterance "F01_423C020L_BUS" (up to 5s) in dt05, bus condition, simulated set. (a-b) speech and noise masks generated by system 4 (top), system 5 (middle), and system 7 (bottom). (c) clean spectrogram. (d) noisy spectrogram (channel 1). (e) enhanced spectrogram by system 7. Log spectrogram in (c-e) are mean normalized. There are baby crying from frame 300.

beamforming, despite the fact that the proposed method does not use TDOA tracking and noise samples before the test utterances, which are both used in the MVDR baseline of system 3.

Matlab based recipes for both the MSE and CE training of mask predictor for the CHiME-4 task are available in [21].

4. CONCLUSIONS AND FUTURE WORKS

In this paper, we investigated several ways to improve the mask-based MVDR beamforming for the ASR application. The main idea is to fine tune the mask predictor to directly minimizing the ASR cost function. This not only improves performance of the MVDR beamforming, but also reduces the heuristics in designing the IBMs in the training stage. We also showed that several other methods also improve the ASR performance, including separate estimation of noise and speech masks, multiple passes of mask estimation, and pooling of masks of different channels. Significant improvement is obtained by using the proposed methods over baseline MVDR beamforming. In the future, we will extend the current work in several ways, such as online tracking of moving speakers and noise statistics.

5. ACKNOWLEDGMENTS

Thanks to Mr. Chenglin Xu for creating simulated noisy and reverberant array data for MSE training of mask predictor. This work is supported by DSO funded project MAISON DSOCL14045 and A-STAR funded HCCS programme.

6. REFERENCES

[1] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.

[3] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The thirdchime'speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, 2015.

[4] 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 and Language, to appear*.

[5] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[6] 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.

[7] Jack Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[8] Lloyd J Griffiths and Charles W Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, 1982.

[9] Simon Doclo and Marc Moonen, "GSVD-based optimal filtering for single and multimicrophone speech enhancement," *IEEE Transactions on Signal Processing*, vol. 50, no. 9, pp. 2230–2244, sep 2002.

[10] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, Atsunori Ogawa, Keisuke Kinoshita, Masakiyo Fujimoto, Chengzhu Yu, Wojciech J Fabian, Miquel Espi, Takuya Higuchi, et al., "The ntt chime-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 436–443.

[11] Nobutaka Ito, Shako Araki, Takuya Yoshioka, and Tomohiro Nakatani, "Relaxed disjointness based clustering for joint blind source separation and dereverberation," in *Acoustic Signal Enhancement (IWAENC), 2014 14th International Workshop on*. IEEE, 2014, pp. 268–272.

[12] Jahn Heymann, Lukas Drude, Aleksej Chinaev, and Reinhold Haeb-Umbach, "Blstm supported gev beamformer front-end for the 3rd chime challenge," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 444–451.

[13] Lukas Drude Jahn Heymann and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*. IEEE, 2016.

[14] Takaaki Hori, Zhuo Chen, Hakan Erdogan, John R Hershey, Jonathan Le Roux, Vikramjit Mitra, and Shinji Watanabe, "The merl/sri system for the 3rd chime challenge using beamforming, robust feature extraction, and advanced speech recognition," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 475–481.

[15] Hakan Erdogan, John Hershey, Shinji Watanabe, Michael Mandel, and Jonathan Le Roux, "Improved mvdr beamforming using single-channel mask prediction networks," in *INTERSPEECH*, 2016.

[16] 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 *ICASSP*. IEEE, 2016.

[17] Xiong Xiao, Shinji Watanabe, Eng Siong Chng, and Haizhou Li, "Beamforming networks using spatial covariance features for far-field speech recognition," in *APSIPA ASC*, 2016.

[18] Mehrez Souden, Jacob Benesty, and Sofiène Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 2, pp. 260–276, 2010.

[19] 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.

[20] Shengkui Zhao, Xiong Xiao, Zhaofeng Zhang, Thi Ngoc Tho Nguyen, Xionghu Zhong, Bo Ren, Longbiao Wang, Douglas L Jones, Eng Siong Chng, and Haizhou Li, "Robust speech recognition using beamforming with adaptive microphone gains and multichannel noise reduction," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 460–467.

[21] Xiong Xiao, "SignalGraph: a Matlab based deep learning toolkit for signal processing," in *https://github.com/singaxiong/SignalGraph*, 2016.