JOINT TRAINING OF COMPLEX RATIO MASK BASED BEAMFORMER AND ACOUSTIC MODEL FOR NOISE ROBUST ASR

Yong Xu[†], *Chao Weng*[†], *Like Hui*[‡*], *Jianming Liu*[†], *Meng Yu*[†], *Dan Su*[†], *Dong Yu*[†]

Tencent AI lab[†] The Ohio State University, USA[‡]

ABSTRACT

In this paper, we present a joint training framework between the multi-channel beamformer and the acoustic model for noise robust automatic speech recognition (ASR). The complex ratio mask (CRM), demonstrated to be more effective than the ideal ratio mask (IRM), is proposed to estimate the covariance matrix for the beamformer. Minimum Variance Distortionless Response (MVDR) beamformer and Generalized Eigenvalue (GEV) beamformer are both investigated under the CRM-based joint training architecture. We also propose a robust mask pooling strategy among multiple channels. A long short-term memory (LSTM) based language model is utilized to re-score hypotheses which further improves the overall performance. We evaluate the proposed methods on CHiME-4 challenge dataset. The CRM based system achieves a relative 10% reduction on word error rate (WER) compared with the IRM based system. Without sequence discriminative training, our best single system already achieves an average WER 2.72% on the test set which is comparable to the state-of-the-art.

Index Terms— Joint training, CHiME-4, complex ratio mask, speech recognition, beamforming

1. INTRODUCTION

Multichannel ASR in challenging noisy environments, e.g, low signal-to-noise ratio (SNR), far-field, overlapped speech, etc., has attracted lots of research efforts recently. The series of CHiME speech separation and recognition challenges [1, 2] were designed to encourage researchers to develop advanced techniques to solve these problems. Beamforming is one of the most useful techniques .

A complex Gaussian mixture model (CGMM) based time-frequency mask estimation was proposed in [3, 4] to help to calculate the steering vector for the MVDR beamforming. They got the best performance on CHiME-3 challenge [3]. Inspired by this work, some subsequent deep learning based mask estimation and beamforming methods were developed. In [5], an iterative mask estimation based on deep neural network was proposed to improve the conventional CGMM based method and they ranked 1st in the CHiME-4 challenge [5]. Meanwhile, a bidirectional LSTM based time-frequency mask estimation was proposed in [6] to estimate the covariance matrix for the MVDR or GEV beamformer. Similar methods can be also found in [7, 8, 9, 10]. These mask-based beamforming methods only need to know the time-frequency masks or the speech presence probability without needing the knowledge of microphone array geometry. Even for the more challenging cocktail party scenario with overlapped speech problem in CHiME-5 [2], careful estimated masks through speaker dependent separation models followed by a GEV beamforming also achieved the best performance [11]. Among all of these methods, the key question is how to reliably estimate masks. In this paper, we propose to use complex ratio mask (CRM) [12] to replace the commonly used real-valued mask. To the best of our knowledge, this is the first time to use CRM to estimate the covariance matrix for the multichannel beamforming. CRM has been verified to be more effective than the ideal ratio mask in the monaural speech separation task [12].

Recently, joint training between the beamformer and acoustic model also attracts lots of interest. Joint training means that the gradients derived from the ASR loss will back-propagate through all the way from the acoustic model to the complex valued beamforming and the mask estimation networks. A unified architecture was proposed in [13] to jointly optimize the multichannel enhancement and the ASR components. Xiao *et al.* [14] also proposed to adopt ASR cost to directly train the RNN based mask estimator. Zmolikova *et al.* [15] proposed to use ASR criterion to optimize the multichannel beamformer for recognizing the speech corrupted by the overlapping speakers. Meanwhile, Beamnet was proposed in [16] to jointly train the acoustic model and the beamforming module to bridge the mismatch between the front-end and the back-end. However, only the real-valued mask was investigated in these papers. In this work, we propose to jointly train the complex ratio mask based beamformer and the acoustic model to expect better ASR performance.

The contribution of this paper is three-fold. First, we propose to use complex ratio mask to replace the real-valued mask to estimate the covariance matrix for the beamformer,

[*]Like Hui performed the work while she was a research intern at Tencent AI lab, Bellevue, USA

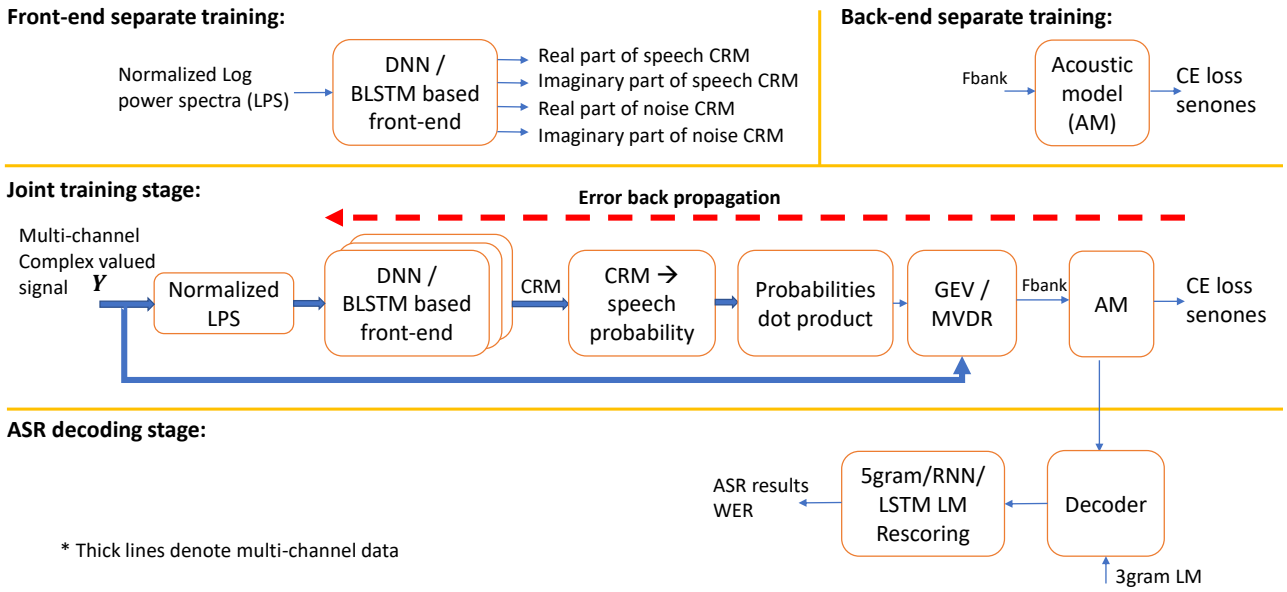

Fig. 1. The proposed joint training system between the complex ratio mask (CRM) based beamformer and the acoustic model.

and demonstrate that CRM can lead to a more reliable beamformer. Second, we jointly train the CRM-based beamformer and the AM with using a more robust mask/probability pooling method. Third, with simple cross-entropy criterion, we already achieved comparable ASR performance with the state-of-the-art single system [17]. Note that in this paper, we did not use I-vector technique and lattice-free maximum mutual information (MMI) training method as [17] did.

2. PROPOSED JOINT TRAINING SYSTEM

Fig. 1 shows the whole framework of the proposed joint training system. The front-end mask estimator and the back-end acoustic model are first separately trained. Then the joint training is performed by concatenating them together.

2.1. Complex ratio mask front-end estimator

To the best of our knowledge, complex ratio mask has not been investigated for multiple channel speech recognition. CRM was first proposed in [12] for monaural speech separation. The mathematical derivation of the speech and the noise complex mask are defined as,

$$S(t,f) = M^s(t,f) * Y(t,f) \tag{1}$$

$$N(t,f) = M^n(t,f) * Y(t,f) \tag{2}$$

where $Y(t,f)$, $S(t,f)$ and $N(t,f)$ denote the noisy speech, clean speech and the pure noise at time frame t and frequency bin f in the complex domain. The complex mask of speech and noise are represented as $M^s(t,f)$ and $M^n(t,f)$ in the complex domain. '$*$' indicates complex multiplication. Note that (t,f) will be omitted below for simplicity. Then the Eq. (1) and Eq. (2) can be extended in the complex-domain forms:

$$(S_r + jS_i) = (M_r^s + jM_i^s) * (Y_r + jY_i) \tag{3}$$

$$(N_r + jN_i) = (M_r^n + jM_i^n) * (Y_r + jY_i) \tag{4}$$

Where the subscript r and i denote the real part and the imaginary part, respectively. Then the complex masks of speech and noise are derived as:

$$M_r^s + jM_i^s = \frac{Y_r S_r + Y_i S_i}{Y_r^2 + Y_i^2} + j\frac{Y_r S_i - Y_i S_r}{Y_r^2 + Y_i^2} \tag{5}$$

$$M_r^n + jM_i^n = \frac{Y_r N_r + Y_i N_i}{Y_r^2 + Y_i^2} + j\frac{Y_r N_i - Y_i N_r}{Y_r^2 + Y_i^2} \tag{6}$$

As [12] did, the complex masks should be compressed into the hyperbolic tangent to avoid the number infinity problem:

$$CRM = K\frac{1 - e^{-C \cdot M}}{1 + e^{-C \cdot M}} \tag{7}$$

This operation compresses complex mask values into $[-K, K]$. C controls its steepness. M denotes the mask. As [12] did, $K = 10$ and $C = 0.1$ are selected. Then as the top-left of Fig. 1 shows, DNN or BLSTM was used to train a front-end to predict the real part and the imaginary part of speech or noise CRM. Mean square error (MSE) was used as the training loss.

2.2. Mask-based beamforming

Mask-based beamforming was successful in CHiME-3 and CHiME-4 challenges [3, 5]. Even in CHiME-5 challenge with overlapped speech problem, the mask-based beamforming is still promising if the speaker dependent mask could be estimated properly [11].

2.2.1. GEV beamformer

With the complex masks derived in subsection 2.1, the speech or noise probability (or ideal ratio mask) at each time frame can be estimated as:

$$\hat{p}^{\nu}(t,f) = \frac{|\hat{M}^{\nu} * Y(t,f)|^2}{|\hat{M}^{s} * Y(t,f)|^2 + |\hat{M}^{n} * Y(t,f)|^2} \tag{8}$$

where ν denotes either the speech or the noise. Comparing to the commonly used ideal ratio/binary mask in [5, 6], there is no independence assumption [18] between the noise and speech in the CRM calculation.

The speech or noise probability $\hat{p}^{\nu}(t,f)$ is estimated independently for each microphone channel. Then the probabilities of all channels are pooling into one probability as,

$$\bar{p}^{\nu}(t,f) = \prod_{d=1}^{D} \hat{p}_d^{\nu}(t,f) \tag{9}$$

where d denotes the microphone channel index. Different from [6, 16], we proposed to use multiplication pooling as in Eq. (9) among different microphone channels here to replace the mean pooling (median pooling in the testing). Our multiplication pooling here is a consistent operation between the training and the testing. We found it was more robust than the mean-median pooling. Because the multiplication pooling can pick out the most confident T-F unit where speech exists. A similar idea can be found in [10]. But our system is a joint training system, we did not use the discontinuous indicator function as [10] did. Then the speech or noise covariance matrix can be estimated as,

$$\mathbf{\Phi}_{\nu\nu}(f) = \frac{\sum_{t=1}^{T} \bar{p}^{v}(t,f)\mathbf{Y}(t,f)\mathbf{Y}(t,f)^{\mathrm{H}}}{\sum_{t=1}^{T} \bar{p}^{v}(t,f)} \tag{10}$$

Where T denotes the total frame number of an utterance. The superscript H represents the conjugate transpose. We also tried $\mathbf{\Phi}_{\mathrm{SS}} = \sum_{t=1}^{T} \hat{\mathbf{S}}\hat{\mathbf{S}}^{\mathrm{H}}$ to directly calculate the covariance matrix considering that the speech or noise of each channel can be easily estimated through the predicted complex ratio masks. But it did not perform well in our experiments compared to the slightly conservative way in Eq. (10).

GEV is aimed at maximizing a posteriori signal-to-noise ratio (SNR) [6], and the beamforming vector $\mathbf{w}_f^{\mathrm{GEV}}$ is as,

$$\mathbf{w}^{\mathrm{GEV}}(f) = \underset{\mathbf{w}(f)}{\operatorname{argmax}} \frac{\mathbf{w}^{\mathrm{H}}(f)\mathbf{\Phi}_{\mathrm{SS}}(f)\mathbf{w}(f)}{\mathbf{w}^{\mathrm{H}}(f)\mathbf{\Phi}_{\mathrm{NN}}(f)\mathbf{w}(f)} \tag{11}$$

2.2.2. MVDR beamformer

MVDR beamformer is to find a weight vector to target into the speech direction without speech distortion while depressing the noise from other directions [10]. The MVDR beamforming vector can be calculated as [10] did,

$$\mathbf{w}^{\mathrm{MVDR}}(f) = \frac{\mathbf{\Phi_{NN}}(\mathbf{f})^{-1}\mathbf{c}(\mathbf{f})}{\mathbf{c}(f)^{\mathrm{H}}\mathbf{\Phi_{NN}}(\mathbf{f})^{-1}\mathbf{c}(\mathbf{f})} \tag{12}$$

where $\mathbf{c}(f)$ is the steering vector which is calculated through the principal component analysis (PCA) of the speech covariance matrix $\mathbf{\Phi_{SS}}(\mathbf{f})$. An alternative MVDR beamformer is derived from the multichannel Wiener filter [10] as $\mathbf{w} = \frac{\mathbf{\Phi_{NN}}^{-1}\mathbf{\Phi_{SS}}}{\mathrm{trace}(\mathbf{\Phi_{NN}}^{-1}\mathbf{\Phi_{SS}})}\mathbf{u}$, $\mathbf{u}$ is the one-hot vector representing a reference microphone. Its advantage is to avoid the PCA, but its performance was worse than the one with Eq. (12) [10].

2.3. Joint training with acoustic model

Wide residual BLSTM (WRBN) based acoustic model was used in our paper, and it is similar to the structure in [19, 20][1]. WRBN is an improved version over the convolutional, long short-term memory, fully connected deep neural network (CLDNN). WRBN achieved good performance in CHiME-4 challenge [20]. The model details can be found in [20]. We firstly trained a WRBN acoustic model based on the log-mel filter bank features as the top-right part of Fig. 1 shows. Then it was used as our initial model for the joint training.

For the joint training, as Fig. 1 shows, we concatenate all of the front-end modules with the back-end modules together. We jointly optimize the models on a whole utterance with full back-propagation through time using a cross-entropy (CE) criterion. The gradients derived from the CE loss will go through the real-valued domain AM, log-mel filter bank feature extraction, complex-valued domain beamformer, real-valued domain mask operation and mask prediction. It means that the ASR loss will finetune the complex mask estimation networks to directly serve for the ASR performance.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Dataset and experiments setup

The database we used in this paper is CHiME-4 six channel data [21]. For the complex mask estimation network, we use DNN and also BLSTM to predict the complex masks. The DNN has three fully connected layers with 1024 hidden units for each layer. The BSLTM has one bi-directional LSTM layer with 512 units in the front, followed by two fully connected layer with 1024 units for each layer. Dropout rate is 0.2 for both of them. Note that the DNN used for mask estimation here is larger than the DNN in [6]. The details of WRBN based acoustic model with 2042 senones output can be found in [20, 19]. All of the front-end mask estimators, complex-valued beamforming operations and back-end AM are conducted on Tensorflow-1.9.0 with Python 3.6.4. To stabilize the training, the beamforming part was conducted on CPU while other parts were done on GPU. The learning rate for joint training was set to 1.0e-6.

[1]Thanks Peidong Wang (Ohio Univ.) for sharing the WRBN code.

3.2. Results and discussions

Table 1 presents WER results of disjoint training systems by decoding with 3-gram LM. The proposed CRM-based beamforming is consistently demonstrated to be more effective than the IRM-based beamforming on the simulation and the real sets. Because there is nearly no assumption in the definition of CRM while IRM needs the independence assumption between the speech and the noise [18]. The GEV beamforming is superior to the MVDR beamforming where [16] has similar observations on CHiME-4 challenge. On the real case of the test set, CRM-GEV can reduce the WER from 6.10 to 5.82 by comparing with the IRM-GEV.

Table 1. WERs of disjoint training system with 3-gram LM decoding on the development set and the test set.

Mask	Beamforming	Dev (%)		Test (%)	
		real	simu	real	simu
IRM	MVDR	4.40	4.21	6.15	5.30
IRM	GEV	4.36	4.29	6.10	5.16
CRM	MVDR	4.36	4.15	5.96	5.07
CRM	GEV	**4.28**	**4.11**	**5.82**	**4.96**

With Table 2, we show the joint training performance on the development set and the evaluation set by comparing to the disjoint training systems. Compared to the disjoint training systems, the joint training systems can obtain slightly larger improvements on the simulation set than on the real set. For example, CRM-JointTr can reduce the WER from 4.96 to 4.79 on the simulation case of the test set while reduce the WER from 5.82 to 5.74 on the real case. This is because that there are much more simulation data than the real data (1600 utterances for real and 7138 utterances for simu) in the training set. Furthermore, with the 5-gram LM plus the RNNLM or LSTM re-scoring, the performance can get large improvements. LSTMLM is more effective than the RNNLM. Finally, the proposed CRM-based joint training system can get a relative 10% WER reduction comparing with the IRM-based joint training system, e.g., reducing the WER from 3.35 to 3.03 on the real case of the test set.

Table 3 shows the WER comparisons with state-of-the-art systems on the CHiME-4 test set. Compared with the CHiME-4 baseline system [21], our best system can achieve 75.7% relative improvement on average. The most comparable systems are UPB Beamnet [16, 20] which is also a joint training system, our proposed CRM-based joint training system are better than their real-valued mask based systems. Our system is also superior to the MERL system [7] and Ohio MVDR system [10]. As the USTC system [5] is a fusion model, we here compare to the state-of-the-art single system, namely JHU interspeech2018 system [17]. Without using any adaptation, I-vector and LF-MMI training criterion which are adopted in [17], we already match the performance of the JHU system [17] on average. In our method, we found

Table 2. WERs between disjoint training systems and joint training systems. The beamforming is GEV all belows. The basic disjoint/joint training system uses 3-gram LM. Then the joint training systems adopts 5-gram LM plus RNNLM or LSTMLM re-scoring.

System	**Dev (%)**			**Test (%)**		
	real	simu	ave	real	simu	ave
IRM-Disjoint	4.36	4.29	4.33	6.10	5.16	5.63
CRM-Disjoint	4.28	4.11	4.20	5.82	4.96	5.39
IRM-JointTr	4.34	4.10	4.22	6.10	4.79	5.45
+ RNNLM	3.18	3.13	3.16	4.58	3.36	3.97
+ LSTMLM	2.35	2.24	2.30	3.35	2.59	2.97
CRM-JointTr	4.25	4.01	4.13	5.74	4.79	5.27
+ RNNLM	3.03	3.01	3.02	4.32	3.33	3.83
+ LSTMLM	**2.18**	**2.10**	**2.14**	**3.03**	**2.40**	**2.72**

BLSTM-based CRM front-end did not improve the performance by comparing to the DNN-based CRM front-end. But our DNN model is much larger than the DNN model where only 2 layers were adopted in [6]. As BLSTM is trained on the whole utterance, it might decrease its generalization capability due to the context dependency.

Table 3. WER comparisons with state-of-the-art systems on the CHiME-4 test set.

System	real	simu	ave
CHiME-4 Baseline [21]	11.5	10.9	11.2
UPB GEV-Beamnet-JointTr [16]	5.42	3.95	4.69
UPB GEV + Adaptation [20]	3.48	2.76	3.12
MERL GEV + Adaptation [7]	3.81	2.94	3.38
Ohio MVDR + Adaptation [10]	3.65	3.09	3.37
JHU GEV + I-vector + LF-MMI [17]	**2.74**	2.66	**2.70**
Proposed BLSTM-CRM-JointTr *	3.25	2.41	2.83
Proposed DNN-CRM-JointTr *	3.03	**2.40**	2.72

* we haven't used adaptation, I-vector and LF-MMI, but in the plan

4. CONCLUSIONS

In this paper, we proposed a complex ratio mask (CRM) based beamforming to replace the commonly used real-valued mask based beamforming. The joint training between the CRM-based beamforming and the acoustic model are also conducted with using a more robust mask pooling method. The proposed CRM-based joint training system can get a relative 10% WER reduction compared with the real-valued mask based system. Without using any speaker adpatation, I-vector and Lattice-free maximum mutual information (LF-MMI) training criterion, our method has already been highly competitive to the state-of-the-art system [17]. In the near future work, we plan to utilize speaker adaptation and LF-MMI to further optimize our CRM-based joint training system.

5. REFERENCES

[1] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *ASRU*. IEEE, 2015, pp. 504–511.

[2] Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal, "The fifth CHiME speech separation and recognition challenge: Dataset, task and baselines," *arXiv preprint arXiv:1803.10609*, 2018.

[3] Takuya Yoshioka, Nobutaka Ito, Marc Delcroix, Atsunori Ogawa, Keisuke Kinoshita, et al., "The NTT CHiME-3 system: Advances in speech enhancement and recognition for mobile multi-microphone devices," in *ASRU*. IEEE, 2015, pp. 436–443.

[4] Takuya Higuchi, Nobutaka Ito, Takuya Yoshioka, and Tomohiro Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *ICASSP*. IEEE, 2016, pp. 5210–5214.

[5] Yan-Hui Tu, Jun Du, Lei Sun, Feng Ma, and Chin-Hui Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," *Proc. Interspeech*, pp. 394–398, 2017.

[6] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *ICASSP*. IEEE, 2016, pp. 196–200.

[7] Hakan Erdogan, Tomoki Hayashi, John R Hershey, Takaaki Hori, Chiori Hori, et al., "Multi-channel speech recognition: Lstms all the way through," in *CHiME-4 workshop*, 2016.

[8] 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.

[9] Xiong Xiao, Chenglin Xu, Zhaofeng Zhang, Shengkui Zhao, Sining Sun, Shinji Watanabe, Longbiao Wang, Lei Xie, Douglas L Jones, Eng Siong Chng, et al., "A study of learning based beamforming methods for speech recognition," in *CHiME 2016 workshop*, 2016, pp. 26–31.

[10] Zhong-Qiu Wang and DeLiang Wang, "Mask weighted stft ratios for relative transfer function estimation and its application to robust asr," in *ICASSP*. IEEE, 2018, pp. 5619–5623.

[11] Jun Du, Tian Gao, Lei Sun, and et al., "The ustc-iflytek system for chime-5 challenge," *Proc. CHiME-5*, 2018.

[12] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[13] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, and Xiong Xiao, "Unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1274–1288, 2017.

[14] 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 *ICASSP*, 2017, pp. 3246–3250.

[15] Katerina Zmolikova, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Tomohiro Nakatani, and Jan Černockỳ, "Optimization of speaker-aware multichannel speech extraction with asr criterion," in *ICASSP*, 2018, pp. 6702–6706.

[16] Jahn Heymann, Lukas Drude, Christoph Boeddeker, Patrick Hanebrink, and Reinhold Haeb-Umbach, "Beamnet: End-to-end training of a beamformer-supported multi-channel asr system," in *ICASSP*, 2017, pp. 5325–5329.

[17] Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, and Shinji Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," *arXiv preprint arXiv:1803.10109*, 2018.

[18] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[19] Peidong Wang and DeLiang Wang, "Filter-and-convolve: A cnn based multichannel complex concatenation acoustic model," in *ICASSP*, 2018, pp. 5564–5568.

[20] Jahn Heymann, Lukas Drude, and Reinhold Haeb-Umbach, "Wide residual blstm network with discriminative speaker adaptation for robust speech recognition," in *Proceedings of CHiME-4 workshop*, 2016, pp. 12–17.

[21] 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*, vol. 46, pp. 535–557, 2017.