Teacher-Student BLSTM Mask Model for Robust Acoustic Beamforming

Zhaoyi Liu* and Yuexian Zou*†
* Peking University Shenzhen Graduate School, Shenzhen, China
† Peng Cheng Laboratory, Shenzhen, China
E-mail: zouyx@pku.edu.cn

Abstract—**Microphone array beamforming has been approved to be an effective approach for suppressing adverse interferences. Recently, acoustic beamformers employing neural networks (NN) for time-frequency (T-F) mask prediction, termed as Mask-BF, have received tremendous interest and shown great benefits as a front-end for distant automatic speech recognition (ASR). However, our preliminary experiments using Mask-BF for ASR task show that the mask model trained with only simulated training data underperforms when the real-recording data appears in the testing stage, where a data mismatch problem occurs. In this study, we aim at reducing the impact of the data mismatch on the mask model. Our research is quite intuitive that the real-recording data can be used together with the simulated data to make the mask model more robust against data mismatch problem. Specifically, two bi-directional long short-term memory (BLSTM) models, are designed as a teacher mask model and a student mask model, respectively. The teacher mask model is trained with simulated data, and it is then employed to generate the soft mask labels for both simulated and real-recording data separately. Then, the simulated data and the real-recording data with generated soft mask labels form the new training data to train the student mask model. As a result, a novel T-S mask BF is developed accordingly. Our T-S mask BF is evaluated as a front-end for ASR on the CHiME-3 dataset. Experimental results show that the generalization ability of our T-S mask BF is enhanced where we obtain relative 4% word error rate (WER) reduction compared to the baseline Mask-BF in the real-recording test set.**

I. Introduction

Distant automatic speech recognition (ASR) has attracted a tremendous amount of attention in recent years with the growing demands for many applications, such as interactions among people and smart home devices by speech [1, 2]. However, for far-field practical applications, background noise and reverberation degrade speech quality as well as the performance of the ASR system, especially under low signal-to-noise ratio (SNR) conditions.

The array beamforming is an efficient multi-channel speech enhancement approach, where the minimum variance distortionless response (MVDR) beamformer and generalized Eigenvalue (GEV) beamformer are the two most popular methods [3, 4]. The performance of the MVDR and GEV heavily depends on the estimation of the correlation matrix of the target speech as well as the correlation matrix of the interferences [5]. Many researches [1, 3] show that there are many constraints on estimating the good correlation matrices, such as array geometry, the distance between the speaker, and the direction of arrival (DOA) of the target speaker. It is clear that for real-life applications where the acoustic environments are unknown, complicated, and time-varying, the traditional beamforming techniques show poorer generalization capacity. Motivated by the outstanding performance of deep neural networks for ASR task, recently, acoustic beamforming technique based on deep learning has been proposed [6, 7] which is data driven and does not subject to these constraints. Moreover, learning-based acoustic beamforming methods have shown their outstanding capacity in dealing with real-world far-field acoustic environment and can be viewed as a general beamformer which has much higher application value. For example, in CHiME-3 and CHiME-4 challenges [8], the learning-based T-F masking models have been developed for beamforming [9, 10], which achieve the state-of-art. In [5], a BLSTM mask model has been designed and trained. In this study, researchers treat the multi-channel signals separately where one speech mask and one noise mask are learned for each channel's signal. Finally, these masks are combined to generate the final mask by median pooling. The beamforming weights are computed as the principal generalized eigenvector of the speech and noise covariance matrices.

In principle, the key to the learning-based acoustic beamforming methods is to learn a monaural time-frequency (T-F) mask model using deep neural networks [1, 11]. Then the spatial covariance matrices of target speech and noise can be estimated for beamforming more accurately. Therefore, training a good T-F mask model is essential to perform beamforming efficiently. Literature study shows that the state-of-art T-F mask models are trained with parallel simulated data and tested with real-recording data, which leads a data mismatch problem between the training and testing conditions [12, 13], such as testing in the real acoustic environment. Moreover, through experiments, it is noted that there are many factors affect the performance of the T-F mask models, such as the level of similarity between the simulated and real-recording data, different acoustic scene conditions in training, development and testing set.

In this paper, we propose to use the teacher-student (T-S) learning scheme [14, 15, 16] to utilize the real-recording data in the training stage of a BLSTM mask model to reduce the impact of the data mismatch for our beamformer, which is termed as T-S mask BF approach. Firstly, we propose to train a *teacher* mask model (TMM). This TMM takes a magnitude spectrum of the simulated noisy signal as input and predicts

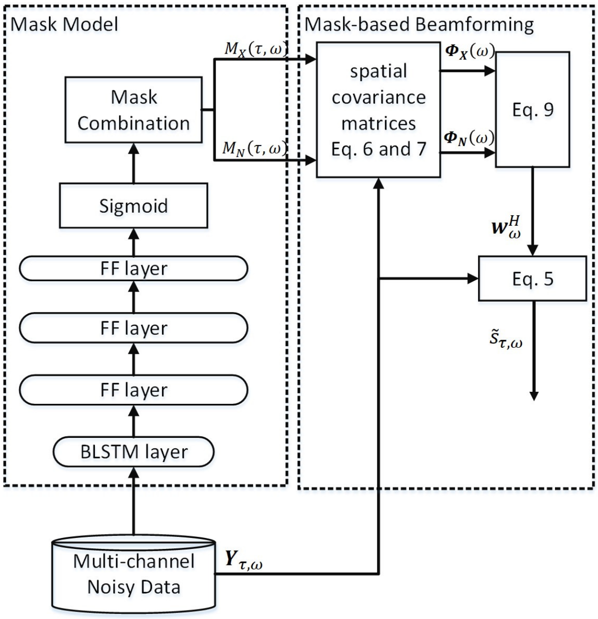

Fig. 1. Diagram of mask-based acoustic beamforming

masks of speech and noise, respectively. Since there are no labels of the real-recording data, we make use of the well-trained TMM as the label generator to predict the soft mask labels of real-recording data as well as the simulated data which form new training data. Secondly, a *student* mask model (SMM) is trained to predict the mask where both simulated data and real-recording data with their soft mask labels are used. In the end, the SMM and TMM is used to generate the enhanced speech and masked out interference, which helps to derive a robust beamformer.

The remainder of this work is organized as follows. In section II, we present the related work of Mask-BF. Section III describes the proposed approach in detail. Detailed experimental corpus, setups, and results are presented in Section IV. Finally, the conclusions are summarized in Section V.

II. RELATED WORK

For presentation completeness, the diagram of the baseline [5] is shown in Fig. 1, which is termed as Mask-BF in this study since it is a data-driven mask model for improving the performance of beamforming. Mask-BF consists of two main blocks: the mask model and mask-based beamforming. The details will be given in part A and B separately.

A. Mask model

The mask model is trained in a fully supervised manner to estimate two masks: the speech mask and the noise mask. In the training stage, the mask model utilizes the simulated data from the training corpus. The magnitude spectrum of the noisy speech is taken as the input of the neural network. The training targets are ideal binary masks (IBM) for speech $IBM_X(\tau,\omega) \in \{0,1\}$ and noise $IBM_N(\tau,\omega) \in \{0,1\}$. The IBMs at frame τ in frequency bin ω are defined based on the SNR ratio with thresholding as:

TABLE I
CONFIGURATIONS OF BLSTM MASK MODEL

Layer	Units	Type	Activation	Dropout
L1	256	BLSTM	Tanh	0.5
L2	513	Feedforward 1	ReLU	0.5
L3	513	Feedforward 2	ReLU	0.5
L4	1026	Feedforward 3	Sigmoid	0.0

$$\mathbf{IBM_N} = \begin{cases} 1, & \frac{||X(\tau,\omega)||}{||N(\tau,\omega)||} < 10^{th_{\mathbf{N}}}, \\ 0, & else. \end{cases} \tag{1}$$

$$\mathbf{IBM_X} = \begin{cases} 1, & \frac{||X(\tau,\omega)||}{||N(\tau,\omega)||} > 10^{th_{\mathbf{X}}}, \\ 0, & else. \end{cases} \tag{2}$$

where $||X(\tau,\omega)|| \in \mathbb{R}_{\geqslant 0}$ and $||N(\tau,\omega)|| \in \mathbb{R}_{\geqslant 0}$ are power spectra of the speech signal and the noise signal at each T-F unit (τ,ω), respectively. In order to achieve the best results, the two threshold $th_{\mathbf{N}}$ and $th_{\mathbf{X}}$ are manually chosen to be different from each other.

The BLSTM mask model is trained to estimate the speech mask $M_X(\tau,\omega)$ and the noise mask $M_N(\tau,\omega)$ at each T-F bin (τ,ω), respectively. Table I shows the configurations of the BLSTM mask model [5]. The BLSTM mask model is trained by using the binary cross-entropy (BCE) cost function. Let's define the total time steps as T and the total number of frequency bins as W. Then the BCE cost function is given by [13]:

$$\begin{aligned} Loss =& BCE(IBM_v, M_v) \\ \overset{def}{=}& \frac{1}{T}\frac{1}{W}\sum_{v\in\{X,N\}}\sum_{\tau=1}^{T}\sum_{\omega=1}^{W} IBM_v\left(\tau,\omega\right) log(M_v\left(\tau,\omega\right)) \\ &+ (1-IBM_v\left(\tau,\omega\right))log(1-M_v\left(\tau,\omega\right)) \end{aligned} \tag{3}$$

where IBM_X and IBM_N are given in (2) and (1), respectively. $M_X(\tau,\omega)$ and $M_N(\tau,\omega)$ are the estimated masks of speech and noise, respectively.

In the testing stage, the masks for each channel are predicted by the BSLTM mask model separately for testing data and then combined to a single mask by using a median operation.

B. Mask-based beamforming

In the short-time Fourier transform (STFT) domain, the received noisy signal from multiple microphones can be expressed as:

$$\mathbf{Y}_{\tau,\omega} = \mathbf{X}_{\tau,\omega} + \mathbf{N}_{\tau,\omega} \tag{4}$$

where $\mathbf{Y}_{\tau,\omega}$, $X_{\tau,\omega}$ and $\mathbf{N}_{\tau,\omega}$ represent STFT of the observed noisy signal, speech and noise respectively.

A beamformer is adapted to estimate the speech from observed noisy signal $\mathbf{Y}_{\tau,\omega}$. The output of the beamformer, $\tilde{s}_{\tau,\omega}$, is calculated by:

$$\tilde{s}_{\tau,\omega} = \mathbf{w}_\omega^{\mathbf{H}}\mathbf{Y}_{\tau,\omega} \tag{5}$$

where superscript H denotes conjugate transpose. The $\mathbf{w}_\omega$ presents the beamforming coefficient vector.

In this study, with the estimated speech mask $M_X(\tau,\omega)$ and noise mask $M_N(\tau,\omega)$ by BLSTM mask model, spatial covariance matrices of speech and noise are computed as:

$$\mathbf{\Phi_X}(\omega) = \sum_{\tau=1}^{T} M_X(\tau,\omega)\,\mathbf{Y}_{\tau,\omega}\mathbf{Y}^{\mathbf{H}}_{\tau,\omega} \tag{6}$$

$$\mathbf{\Phi_N}(\omega) = \sum_{\tau=1}^{T} M_N(\tau,\omega)\,\mathbf{Y}_{\tau,\omega}\mathbf{Y}^{\mathbf{H}}_{\tau,\omega} \tag{7}$$

These spatial covariance matrices are used to compute the beamforming coefficient vector $\mathbf{w}_\omega$. Motivated by the Generalized Eigenvalue (GEV) beamformer are more suitable than MVDR for reverberant environments [5]. The GEV beamformer is employed and its beamforming coefficient vector is calculated by:

$$\mathbf{w}_{GEV}(\omega) = \underset{\mathbf{w}}{argmax}\frac{\mathbf{w^H\Phi_X}(\omega)\,\mathbf{w}}{\mathbf{w^H\Phi_N}(\omega)\,\mathbf{w}} \tag{8}$$

This optimization in Eq. (8) is equivalent to solving the following eigenvalue problem:

$$\left\{\mathbf{\Phi_N^{-1}}(\omega)\mathbf{\Phi_X}(\omega)\right\}\mathbf{w_{GEV}}(\omega) = \lambda\mathbf{w_{GEV}}(\omega) \tag{9}$$

where $\mathbf{w_{GEV}}(\omega)$ is the eigenvector of $\left\{\mathbf{\Phi_N^{-1}}(\omega)\mathbf{\Phi_X}(\omega)\right\}$ and λ is the corresponding eigenvalue.

Finally, following the method proposed in [5], the GEV beamformer is then further adapted by the blind analysis normalization (BAN), which is able to reduce arbitrary distortion of GEV beamformer.

III. PROPOSED METHOD

According to the principle of GEV beamformer introduced in Section II, we note that the estimation of $\mathbf{\Phi_X}(\omega)$ and $\mathbf{\Phi_N}(\omega)$ is the key to obtain a good beamformer for target speech enhancement, which is directly related to the estimation of speech and noise masks. Our preliminary experiments using Mask-BF for ASR task show that the mask model trained using all simulated training data underperforms when the real-recording data is used in the prediction stage, where a data mismatch problem occurs. Motivated by the good results brought by BLSTM mask model, in this study, we focus on improving the capability of mask learning by introducing the real-recording data which is able to alleviate the mismatch between the training data and the testing data. Our research motivation is quite intuitive that the real-recording data can be used together with the simulated data to train a better mask model. In this work, a Teacher-Student (T-S) learning scheme is proposed where a teacher mask model (TMM) and a student mask model (SMM) are designed. The training strategy is proposed to use both simulated data and real-recording data. As a result, our developed front-end is termed as T-S mask BF. Fig. 2 illustrates the framework of the proposed approach. The details are given in the following context.

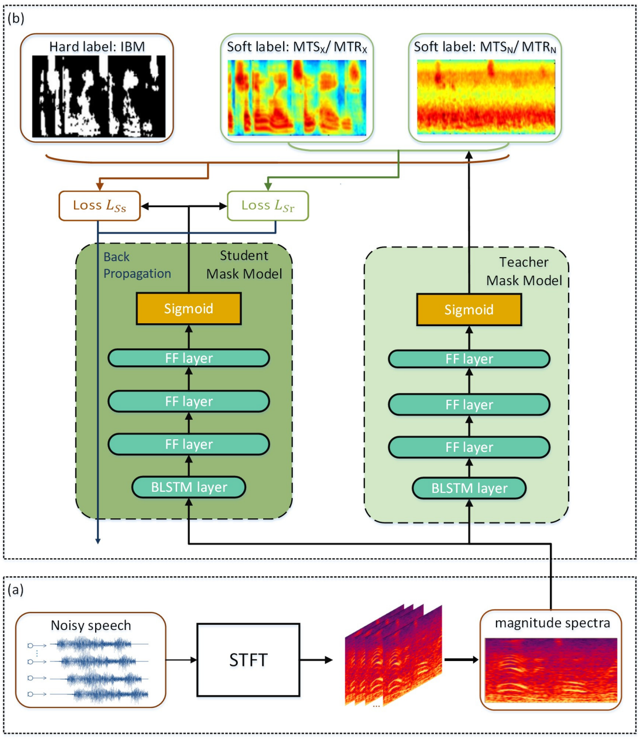

Fig. 2. The framework of our proposed Teacher-student mask model (T-S mask). (a) Feature extraction: Obtain the short-time Fourier transforms (STFT) of the noisy signals and calculate their magnitude spectra $|\mathbf{Y}|_{\tau,\omega}$. Use the magnitude spectrum of ith channel $Y_i^m(\tau,\omega)$ as the input of the mask model. (b) Proposed T-S mask model: The well-trained teacher mask model generates the estimated speech soft mask labels MTS_X and MTR_X as well as noise soft mask labels MTS_N and MTR_N as additional labels to student mask model.

A. Teacher mask model (TMM)

The training processes and architecture of the teacher mask model (TMM) are the same as the baseline (BLSTM mask model), which are described in Section II. Note that the TMM is only trained by using simulated training data.

As shown in Fig. 2, in our design, the well-trained TMM is used as the soft label generator for the real-recording data as well as the simulated data. Therefore, for the simulated data, we can get two soft masks: $MTS_X(\tau,\omega) \in [0,1]$, and $MTS_N(\tau,\omega) \in [0,1]$ for speech and noise, respectively. For the real-recording data, another two masks can be generated as well, which are denoted as $MTR_X(\tau,\omega) \in [0,1]$, and $MTR_N(\tau,\omega) \in [0,1]$. In the end, we can form three different training data: {*simulated noisy data, corresponding $IBMs$*}, {*simulated noisy data, corresponding $MTSs$*}, and {*real-recording noisy data, corresponding $MTRs$*}. The example of these mask labels is given in the (b) of Fig. 2. These training data will be used to train our student mask model.

B. Student mask model (SMM)

The structure of student mask model (SMM) is the same as the baseline (BLSTM mask model) described in Section II-A. In the SMM training stage, we use different cost functions to train our SMM with the simulated data and the real-recording data, respectively.

For the simulated data training, we consider the following cost functions for speech L_{Ss_X} and noise L_{Ss_N}:

$$\begin{aligned} L_{Ss\_X} =& (1-\pi)BCE(IBM_X(\tau,\omega), MS_X(\tau,\omega)) \\ &+ \pi MSE(MTS_X(\tau,\omega), MS_X(\tau,\omega)) \end{aligned} \tag{10}$$

$$\begin{aligned} L_{Ss\_N} =& (1-\pi)BCE(IBM_N(\tau,\omega), MS_N(\tau,\omega)) \\ &+ \pi MSE(MTS_N(\tau,\omega), MS_N(\tau,\omega)) \end{aligned} \tag{11}$$

where $MS_X(\tau,\omega)$, and $MS_N(\tau,\omega)$ denote the estimated speech mask and noise mask by SMM, respectively. The $IBM_X(\tau,\omega)$ and $IBM_N(\tau,\omega)$ are the hard mask labels of speech and noise, respectively. The hyper-parameter $\pi \in [0,1]$ is the linear interpolation weight. For soft mask labels, the mean squared error (MSE) between the inferred mask prediction and the target soft mask label is used as the cost function. The MSE is defined as follow:

$$\begin{aligned} &MSE(MTS_v(\tau,\omega), MS_v(\tau,\omega)) \overset{def}{=} \\ &\frac{1}{T}\frac{1}{W}\sum_{v\in\{X,N\}}\sum_{\tau=1}^{T}\sum_{\omega=1}^{W}\|MS_v(\tau,\omega) - MT_v(\tau,\omega)\|_2^2 \end{aligned} \tag{12}$$

The final cost function of SMM for simualated data, termed as L_{Ss} is expressed as:

$$L_{Ss} = (L_{Ss\_X} + L_{Ss\_N})/2 \tag{13}$$

For real-recording data training, in our proposed T-S mask model, the soft mask labels of the real-recording data enable be obtained from TMM. The cost function for SMM on real-recording data, termed as L_{Sr}, is defined as:

$$\begin{aligned} L_{Sr} =& [MSE(MTR_X(\tau,\omega), MS_X(\tau,\omega)) \\ &+ MSE(MTR_N(\tau,\omega), MS_N(\tau,\omega))]/2 \end{aligned} \tag{14}$$

where the MSE cost function is defined in Eq. (12).

The SMM has been trained on the simulated data and real-recording data with loss L_{Ss} and loss L_{Sr}, respectively. When the SMM predicts the speech and noise mask for each microphone channel, we calculate the beamforming coefficient vector by using the method shown section II-B.

IV. Experiments

We evaluate the proposed Teacher-Student mask model for beamforming (T-S mask BF) approach on ASR tasks using the CHiME-3 corpus [8]. The T-S mask BF is used as a front-end for the ASR system.

A. Corpus

The CHiME-3 corpus includes real-recording data and simulated data generated by artificially mixing the incorporations of Wall Street Journal (WSJ) corpus sentences spoken with 4 different noisy environments including cafe (CAFE), the street junction (STR), public transport (BUS) and pedestrian area (PED). This corpus is recorded by using a 6-channel microphone array attached to a tablet device. The corpus is divided into 3 respective subsets. The first one is the training set, composing 8738 (1600 real + 7138 simulated) noisy utterances. The second one is the development set (dt_05), containing 3280 (1640 real + 1640 simulated) noisy utterances. The third one is the evaluation set (et_05), including 2640 (1320 real + 1320 simulated) noisy utterances.

B. Speech Recognition

To facilitate the comparisons, the original baseline back-end of CHiME-3 is used. It features two different acoustic models (AM). One is based on a Gaussian Mixture Model (GMM) AM [17, 18] and the other is based on a deep neural network (DNN) AM [8], which contains 7 layers with 2048 Sigmoid units. Both of them are trained by using Kaldi speech recognition toolkit [19]. For the language model (LM), a standard WSJ 5K word tri-gram LM [8] is uesd in all experiments.

C. Experimental setups and results

As frontend processing, we compare the proposed approach T-S mask BF with another two beamforming algorithms. The first one is the weighted delay-and-sum beamformer. It is implemented using the BeamformIt! toolkit [20] where the DOA estimate is obtained from GCC-PHAT [3] and a two-step Viterbi post-processing technique is used to avoid instabilities. The second one is proposed by Heymann *et al.* [5] described in Section II-A.

The results of these ASR experiments are shown in Table II. From Table II, the results reveal that all of the Mask-BF methods outperform that of the BeamformIt by a large margin for the real-recording data, despite sharing the same back-end. Meanwhile, we can see that the ASR performance of the student mask models with different hyper-parameters π are better than the ASR performance of the teacher mask model (original BLSTM mask model) as expected. Specifically, for the GMM AM, adding the Teacher-Student (T-S) learning scheme results in relative the best improvement rate of the real-recording evaluation data up to 4.3%. For the DNN AM, the proposed T-S mask BF approach can obtain relative 3.9% WER the most reduction compared to the teacher mask model for the real-recording evaluation data. This means that utilizing T-S learning scheme to estimate mask can reduce the impact of the data mismatch of Mask-BF by pooling real-recording data with simulated data in the SMM training stage. Thus, the proposed T-S mask BF approach is able to generalize better in real-life practical applications.

We also compared our proposed SMMs with different values of hyper-parameter π. The purpose of this experiment is to find out that using which linear interpolation weight can make SMM achieve best ASR performance. From Table II (rows 5-14) we find that the parameter choice of $\pi = 0.95$ gives the best performance amongst the different values we tried.

V. Conclusion

Motivated by the data mismatch problem for Mask-BF results from training simulated data and testing real data,

TABLE II
OVERVIEW OF THE AVERAGE WERS (%) FOR DIFFERENT BEAMFORMING METHODS ON THE CHIME-3

	Parameters	Front-end	Back-end	DEV		EVAL	
				simu	*real*	*simu*	*Real*
1	–	BeamformIt	GMM	11.64	20.53	12.79	37.31
2	–	BeamformIt	DNN	9.43	20.16	11.92	33.26
3	–	Teacher	GMM	10.8	11.77	11.59	17.97
4	–	Teacher	DNN	9.04	10.83	11.34	16.87
5	$\pi = 0.00$	Student	GMM	10.77	11.61	11.48	17.54
6	$\pi = 0.35$	Student	GMM	9.87	11.43	11.19	16.38
7	$\pi = 0.65$	Student	GMM	10.26	11.05	11.28	15.64
8	$\pi = 0.95$	Student	GMM	**8.38**	**9.31**	**9.86**	**13.7**
9	$\pi = 1.00$	Student	GMM	8.4	9.37	9.89	13.79
10	$\pi = 0.00$	Student	DNN	8.96	10.78	11.07	16.36
11	$\pi = 0.35$	Student	DNN	8.61	10.39	10.54	15.8
12	$\pi = 0.65$	Student	DNN	8.83	10.1	10.75	14.83
13	$\pi = 0.95$	Student	DNN	**7.71**	**8.9**	**9.08**	**12.97**
14	$\pi = 1.00$	Student	DNN	7.72	8.96	9.13	13.09

we propose a teacher-student learning scheme for mask-based acoustic beamforming (T-S mask BF). In our T-S mask BF, the well-trained teacher mask model (TMM) is used as the soft label generator for both the real-recording data and the simulated data separately, so that the real-recording data can be pooled with the simulated data for mask model. Then, a student mask model (SMM) is trained to predict the mask where both simulated data and real-recording data with their soft mask labels generated by TMM are used. Experimental results show that our proposed T-S mask BF approach can make the SMM more robust against data mismatch problem and improve the performance of ASR.

ACKNOWLEDGMENT

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274). Special acknowledgements are given to AOTO-PKUSZ Joint Research Center for Artificial Intelligence on Scene Cognition & Technology Innovation for its support

REFERENCES

[1] Mohammad Hasanzadeh Mofrad and Daniel Mosse. Speech recognition and voice separation for the internet of things. *Proceedings of the 8th International Conference on the Internet of Things*, page 8, 2018.

[2] Dong Yu and Jinyu Li. Recent progresses in deep learning based acoustic models. *IEEE/CAA Journal of Automatica Sinica*, 4(3):396–409, 2017.

[3] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-golan, and Alexey Ozerov. A Consolidated Perspective on Multimicrophone Speech Enhancement and Source Separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 25(4):692–730, 2017.

[4] Alexander Krueger, Ernst Warsitz, and Reinhold Haebumbach. Speech Enhancement With a GSC-Like Structure Employing Eigenvector-Based Transfer Function Ratios Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(1):206–219, 2011.

[5] Jahn Heymann, Lukas Drude, and Reinhold Haebumbach. Neural network based spectral mask estimation for acoustic beamforming. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 196–200, 2016.

[6] Xueliang Zhang, Zhongqiu Wang, and Deliang Wang. A Speech Enhancement Algorithm by Iterating Single- and Multi-Microphone Processing and Its Application to Robust ASR. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 276–280, 2017.

[7] Yuxuan Wang, Arun Narayanan, and Deliang Wang. On training targets for supervised speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 22(12):1849–1858, 2014.

[8] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe. The third 'CHiME' Speech Separation and Recognition Challenge: Analysis and Outcomes. *Computer Speech Language*, 46:605–626, 2017.

[9] 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. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 436–443, 2015.

[10] Yanhui Tu, Jun Du, Lei Sun, Feng Ma, and Chinhui Lee. On Design of Robust Deep Models for CHiME-4 Multi-Channel Speech Recognition with Multiple Configurations of Array Microphones. *INTERSPEECH*, pages 394–398, 2017.

[11] Deliang Wang and Jitong Chen. Supervised Speech Separation Based on Deep Learning: An Overview. *IEEE Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018.

[12] 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*, 46:535–557, 2017.

[13] Ying Zhou and Yanmin Qian. Robust Mask Estimation By Integrating Neural Network-Based and Clustering-Based Approaches for Adaptive Acoustic Beamforming. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 536–540, 2018.

[14] Geoffrey E Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the Knowledge in a Neural Network. *arXiv: Machine Learning*, 2015.

[15] Aswin Shanmugam Subramanian, Szu Jui Chen, and Shinji Watanabe. Student-teacher learning for BLSTM mask-based speech enhancement. *Interspeech*, pages 3249–3253, 2018.

[16] Ladislav Mosner, Minhua Wu, Anirudh Raju, Sree Hari Krishnan Parthasarathi, Kenichi Kumatani, Shiva Sundaram, Roland Maas, and Bjorn Hoffmeister. Improving Noise Robustness of Automatic Speech Recognition via Parallel Data and Teacher-student Learning. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[17] Li Deng, Patrick Kenny, Matthew Lennig, Vishwa Gupta, Franz Seitz, and Paul Mermelstein. Phonemic hidden markov models with continuous mixture output densities for large vocabulary word recognition. *IEEE Transactions on Signal Processing*, 39(7):1677–1681, 1991.

[18] Daniel Povey, Lukas Burget, Mohit Agarwal, Pinar Akyazi, Feng Kai, Arnab Ghoshal, Ondej Glembek, Nagendra Goel, Martin Karafiat, Ariya Rastrow, et al. The subspace Gaussian mixture model-A structured model for speech recognition. *Computer Speech Language*, 25(2):404–439, 2011.

[19] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The Kaldi Speech Recognition Toolkit. *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[20] Xavier Anguera, Chuck Wooters, and Javier Hernando. Acoustic Beamforming for Speaker Diarization of Meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.