IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2018.DOI

Robust Beamforming for Speech Recognition Using DNN-based Time-Frequency Masks Estimation

WENBIN JIANG, FEI WEN, (MEMBER, IEEE), AND PEILIN LIU, (Senior Member, IEEE)
Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China (e-mail: jwb361@sjtu.edu.cn; wenfei@sjtu.edu.cn; liupeilin@sjtu.edu.cn)
Corresponding author: Fei Wen (e-mail: wenfei@sjtu.edu.cn).

ABSTRACT This paper addresses the robust beamforming problem for speech recognition using a novel time-frequency mask estimator. The beamformer first estimates the time-frequency mask using a deep neural network (DNN), based on which the covariance matrices of the target speech and noise are computed. Then, the beamformer coefficients are directly obtained via generalized eigenvector decomposition. To achieve accurate covariance matrix estimation for robust beamforming, we propose a DNN based mask estimator which can exploit the spatial features of the multi-channel microphone signals. The proposed mask estimator leverages the spatial information of the microphone array by using multi-channel signals to estimate a speech-aware mask and a noise-aware mask simultaneously. Using the target-specified masks, accurate covariance matrices of the target speech and noise can be obtained from the observation independently. Experiments on CHiME4 data sets demonstrate that, compared with the baseline toolkit (BeamformIt) and the winner in the CHiME3 challenge, the proposed method achieves better results both in terms of perceptual speech quality and speech recognition error rate.

INDEX TERMS Acoustic beamforming, multi-channel speech enhancement, deep neural network, robust speech recognition

I. INTRODUCTION

BENEFITED from the successful application of deep neural network (DNN) into automatic speech recognition (ASR), the accuracy of ASR is closely approaching human's level in close-talking and noise-free scenarios. However, in many practical applications with far-field and/or noisy conditions, accurate and reliable speech recognition is still a challenging problem. In such applications, a microphone array is usually employed to enhance the received noisy speech signals before ASR via beamforming techniques. In conventional beamforming methods, accurate steering vector estimation is paramount for effective noise reduction. However, in practical applications, steering vector estimation is vulnerable to inaccurate knowledge, such as the geometry of the array, or a plane wave assumption.

To overcome this limitation, a robust minimum variance distortionless response (MVDR) beamformer using time-frequency mask has been proposed in [1], which does not need a priori knowledge on the array geometry or far-field assumption on the source. This method first estimates the time-frequency masks, based on which the covariance matrix of the desired speech signals is computed. Then, the steering vector is estimated via eigenvector decomposition and used for MVDR beamforming. Moreover, a robust beamformer which does not need the estimation of the steering vector has been proposed in [2]. In this method, the covariance matrices of the target speech and the noise are firstly computed based on the time-frequency masks, and then, the beamformer coefficients are directly obtained via generalized eigenvector decomposition. Both the methods [1], [2] can achieve state-of-the-art ASR performance.

For the beamforming methods [1], [2], the key is an effective time-frequency mask estimation for the speech or noise, based on which the covariance matrices of the target speech signal and the background noise can be obtained. For time-frequency mask estimation, many model-based methods have been proposed, such as employing Gaussian mixture model [1], [3] or employing Watson mixture model [4]. However, these model-based estimators operate on each individual frequency separately, and do not have the capability of capturing the speech characteristics. Meanwhile, model-based estimators may lead to erroneous results due to distribution

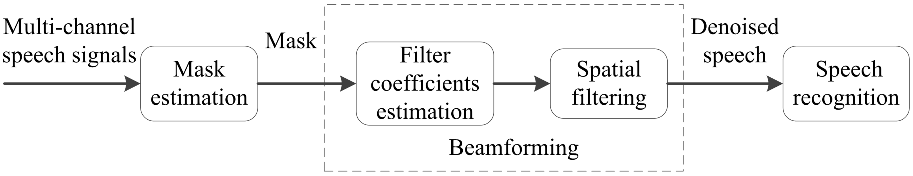


FIGURE 1: Framework of multi-channel speech enhancement for robust speech recognition.

model mismatch. To address these problems, a more popular method is applying data-driven approaches (e.g. DNN) to estimate the time-frequency mask, either estimating single channel mask [5] or estimating multi-channel masks independently followed by a median operation [2]. Nevertheless, these DNN based methods have not made full use of the spatial features of the multi-channel signals.

In this work, we propose a DNN-based time-frequency mask estimator for robust beamforming. The proposed estimator uses multi-channel signals as training data and is expected to learn multi-target specified ideal ratio mask (IRM) (i.e., an IRM for noise-aware channel and an IRM for speech-aware channel). Unlike the DNN architecture [5] which only predicts single channel mask, or the DNN architecture [2] which predicts multi-channel masks independently using a median operation to obtain the mask, the proposed DNN architecture can make full use of the spatial features of the multi-channel signals and would be more robust in practical applications. Specifically, we first estimate the time-frequency mask for the speech and noise via a multi-target specified DNN. Then, covariance matrices of the target speech and the noise are computed, based on which the beamforming filtering coefficients are directly obtained via generalized eigenvector decomposition.

The rest of this paper is organized as follows. Section II introduces related works. We present the framework of the proposed robust acoustic beamforming in Section III. Then, we present the details of the multi-target specified DNN architecture for supervised time-frequency estimation in Section IV. Experimental results are provided in Section V. Finally, Section VI and VII provide discussion and conclusion.

II. RELATED WORKS

Acoustic beamforming or multi-channel speech enhancement [29], [30] aims to enhance the desired signal coming from a particular direction while suppressing undesired interference in other directions. Traditional methods (e.g. the popular MVDR beamformer) rely heavily on accurate steering vector estimation. In practical applications, such beamformers are vulnerable to steering vector estimation errors caused by improper assumptions (e.g., far-field assumption), inaccurate knowledge (e.g., array geometry), and erroneous direction-of-arrival (DOA) estimation of the desired source, especially for mobile devices. For example, in the scenarios with handheld devices, the far-field assumption of the source would introduce an error to steering vector estimation.

To address these problems, most recent studies on multi-channel speech enhancement use robust beamformer to fight against the mismatches and steering vector estimation errors [1, 2, 5]. Fig. 1 illustrates the framework of multi-channel speech enhancement for robust speech recognition. It mainly consists of mask estimation, filter coefficients estimation, and spatial filtering, in which the latter two steps are beamforming.

For mask estimation, a clustering-based method has been proposed in [1], which uses the expectation-maximization (EM) algorithm to estimate the spatial covariance matrix, for noise and target speech. The EM algorithm is performed on each frequency bin independently, and the mask is obtained by the posterior probability of the cluster. On the other hand, many DNN-based mask estimation methods have been studied, both for single-channel speech enhancement [8], [25]–[28] and for multi-channel speech enhancement [2], [5], [31]–[33]. Various masks methods have been evaluated in [8], such as the ideal binary mask (IBM), target binary mask (TBM), and ideal ratio mask (IRM), etc. The evaluation results showed that IRM obtained the best performance for the single channel speech enhancement. The very recent work in [2] uses two IBMs as the training target in the multi-channel speech enhancement task. The results demonstrated that the DNN-based mask estimation method yields better accuracy than the clustering-based method.

To achieve robust beamforming, it is popular to use robust MVDR [1], [22]–[24] or blind acoustic beamforming [2], [4], [6]. The robust MVDR beamforming method does not require DOA estimation, and the steering vector is obtained by eigenvector decomposition of the target speech covariance matrix. This makes the estimated steering vector more accurate and robust than that obtained by traditional methods. Moreover, blind acoustic beamforming method [6] even does not need the estimation of the steering vector, the spatial filter coefficients are obtained by the generalized eigenvector decomposition of the masked covariance matrices.

III. FRAMEWORK OF ROBUST ACOUSTIC BEAMFORMING

In this paper, we use the robust acoustic beamforming framework discussed above, and adopt DNN to estimated multi-target specified ideal ratio masks. Fig. 2 shows the framework of the proposed robust acoustic beamformer. It consists of DNN-based time-frequency mask estimation, beamforming filter coefficients calculation, and blind acoustic beamforming.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2870758, IEEE Access

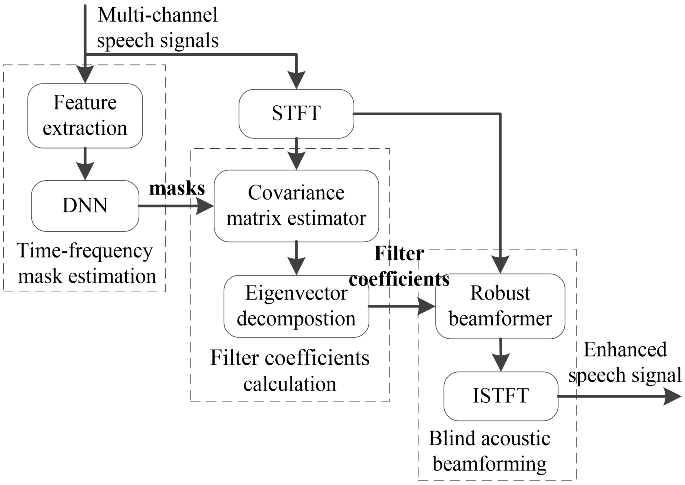

Multi-channel speech signals
Feature extraction
DNN
masks
Time-frequency mask estimation
STFT
Covariance matrix estimator
Eigenvector decomposition
Filter coefficients calculation
Filter coefficients
Robust beamformer
ISTFT
Enhanced speech signal
Blind acoustic beamforming

FIGURE 2: Block diagram of the proposed robust acoustic beamforming method.

The beamformer is operated in the short-time Fourier transform (STFT) domain. Let $\mathbf{y}_{f,t} \in \mathbb{C}^{M\times 1}$ denote the STFT of the M channel microphone array output signals at frequency bin f and time frame t. The enhanced signal is obtained by applying a linear filter $\mathbf{w}_f$ to the observed multi-channel signals

$$\mathbf{x}_{f,t} = \mathbf{w}_f^H \mathbf{y}_{f,t}. \tag{1}$$

Using the criterion of maximizing the signal-to-noise-ratio (SNR) [6], the optimal beamformer coefficient vector $\mathbf{w}_f^o$ is determined via solving the following problem

$$\mathbf{w}_f^o = \arg\max_{\mathbf{w}_f} \frac{\mathbf{w}_f^H \mathbf{R}_f^{(x)} \mathbf{w}_f}{\mathbf{w}_f^H \mathbf{R}_f^{(n)} \mathbf{w}_f} \tag{2}$$

where $\mathbf{R}_f^{(x)}$ and $\mathbf{R}_f^{(n)}$ are respectively the covariance matrices of the desired speech signal and the undesired noise. The solution to the maximization problem (2) is given by generalized eigenvector decomposition as

$$\mathbf{R}_f^{(x)} \mathbf{w}_f = \lambda \mathbf{R}_f^{(n)} \mathbf{w}_f. \tag{3}$$

For this generalized eigenvector decomposition based beamforming method, the key is accurate covariance matrices estimation of the desired signal and the undesired noise.

Inspired by computational auditory scene analysis (CASA), we can define the ideal mask of the desired speech, $m_{f,t}^{(x)}$, and the undesired noise, $m_{f,t}^{(n)}$, in each time-frequency unit. Thus, the covariance matrices of the speech and noise can be obtained by

$$\mathbf{R}_f^{(x)} = \frac{1}{\sum_t m_{f,t}^{(x)}} \sum_t m_{f,t}^{(x)} \mathbf{Y}_{f,t} \tag{4}$$

$$\mathbf{R}_f^{(n)} = \frac{1}{\sum_t m_{f,t}^{(n)}} \sum_t m_{f,t}^{(n)} \mathbf{Y}_{f,t} \tag{5}$$

respectively, where $\mathbf{Y}_{f,t} = \mathbf{y}_{f,t}\mathbf{y}_{f,t}^H$ is the power spectrum of the observed multi-channel noisy signals.

The proposed beamforming method does not require the s-teering vector to compute the beamforming filter coefficients. It only requires covariance matrices estimation of the desired speech signal and the undesired noise, which are obtained by masking the observed noisy speech using the time-frequency mask. Accordingly, accurate time-frequency mask estimation is the key to this beamformer. In the following, we propose a novel supervised time-frequency mask estimation method.

IV. TIME-FREQUENCY MASK ESTIMATION

Supervised time-frequency mask estimation method, typically an elaborately trained DNN, has shown great advantage over conventional model-based methods [2]. For the training target, the ideal ratio mask, which is closely related to the frequency-domain Wiener filter [7], usually yields better performance than the ideal binary mask in the monaural speech enhancement [8]. In this work, we use multi-target specified ideal ratio masks as the training target.

A. MULTI-TARGET SPECIFIED IDEAL RATIO MASKS

Unlike IBM using a binary value for a time-frequency unit, IRM uses a floating-point value ranged in [0, 1]. Let $m_{f,t}^{(x)} \in [0,1]$ denote an IRM for clean speech at frequency bin f and time t, it is defined as

$$m_{f,t}^{(x)} = \left(\frac{X_{f,t}^2}{X_{f,t}^2 + N_{f,t}^2} \right)^{\alpha} \tag{6}$$

where $X_{f,t}^2$ and $N_{f,t}^2$ denote spectral power of speech and noise of a time-frequency unit, respectively. α is a parameter to scale the mask and set to 0.5 empirically. Similarly, the IRM for noise $m_{f,t}^{(n)} \in [0,1]$ is defined as

$$m_{f,t}^{(n)} = \left(\frac{N_{f,t}^2}{X_{f,t}^2 + N_{f,t}^2} \right)^{\alpha}. \tag{7}$$

Intuitively, the summation of $m_{f,t}^{(x)}$ and $m_{f,t}^{(n)}$ should be 1, and it seems to be superfluous to define two masks. But, in our implementation of the multi-target specified masks, the summation of the two masks is not equal to 1 (i.e. $m_{f,t}^{(x)} + m_{f,t}^{(n)} \neq 1$). We use a speech-aware channel to obtain $m_{f,t}^{(x)}$, and use a noise-aware channel to obtain $m_{f,t}^{(n)}$. For example, for the CHiME4 6-channels dataset [9], the channel 5 which faces forward and is closest to the target speaker would be selected as the speech-aware channel. Correspondingly, the channel 2 which faces backward and is farthest to the target speaker would be selected as the noise-aware channel.

The main advantage of the proposed multi-target specified masks is that it leverages the power of microphone array to obtain more reliable masks for speech (speech-aware channel) and noise (noise-aware channel). Moreover, it does not require the exact microphone array geometry or DOA of the target speaker. It just requires a priori knowledge that which channel is more likely to receive the desired speech and which channel is more likely to receive the undesired noise.

B. DNN BASED MASKS ESTIMATION

We use a multi-task DNN architecture [10]–[12] to jointly learn the two masks. By sharing representations between related tasks, the multi-task learning can be expected to achieve better performance than training an ensemble of models. The diagram of the proposed DNN for multi-target specified masks is shown in Fig. 3. Using multi-channel noisy features as input, the multi-task DNN is expected to jointly estimate the IRM for speech (denoted by IRM_x) and the IRM for noise (denoted by IRM_n) simultaneously.

As shown in Fig. 3, the DNN follows a multilayer architecture. A stack of full-connected layers using rectified linear units (ReLUs) activation function compose the shared hidden layer. Two individual full-connected layers using sigmoid activation function compose the task-specific output layer. It should be noted that other types of sophisticated network architectures, e.g., convolutional neural network (C-NN) and long short-term memory recurrent neural network (LSTM-RNN) network, may be good alternatives to the full-connected layer in the shared layer. However, the discussion of different DNN architectures is beyond the scope of this paper.

For the DNN input, we use the integrated multi-channel signals rather than the single channel signal [6] or the separated multi-channel signals [2]. Besides, in order to capture the temporal context of the speech, the input frames of each channel are concatenated in to super-vectors consisting of $2c+1$ frames (c left, 1 center, and c right frames). As a result, the dimension of the input feature is $F \times (2c+1) \times M$, where F is the number of frequency bins of the input feature and M is the number of microphone array channels. We jointly learn the IRM for speech and the IRM for noise, the cost function combines the cost of the two masks

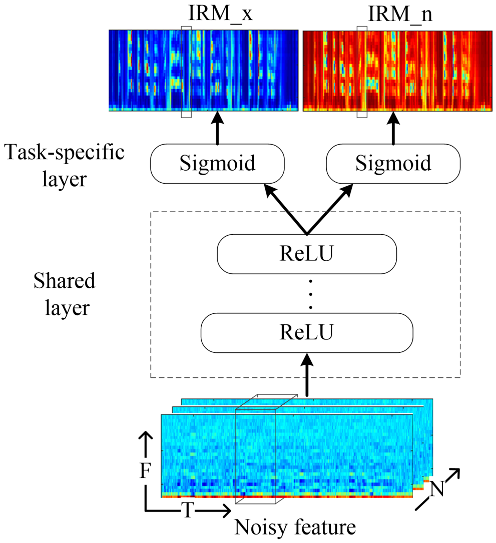


FIGURE 3: Diagram the proposed DNN for time-frequency mask estimation.

$$E = \sum_{f,t} \left[\beta (m_{f,t}^{(x)} - \tilde{m}_{f,t}^{(x)})^2 + (1-\beta)(m_{f,t}^{(n)} - \tilde{m}_{f,t}^{(n)})^2 \right] \tag{8}$$

where β is a weight coefficient of the two mean squared error (MSE) items, which is empirically set to 0.5.

Making full use of the multi-channel and multi-frame observations, the proposed DNN architecture has the capability of exploiting the spatial information of the microphone array and the temporal context information of the speech. To achieve this, we feed the DNN with multi-channel and multi-frame noisy features, and train it to estimate both speech-aware mask and noise-aware mask simultaneously. It should be noticed that the DNN requires the target-aware masks only in the supervised training stage, and the trained DNN can predict the two masks in the estimation stage.

V. EXPERIMENTAL RESULTS

We perform the experiments using the CHiME4 data corpus. The corpus consists of real and simulated 6-channels audio data taken from the 5k WSJ0-Corpus with four different types of noise. There are totally 8738 utterances for training, 3280 utterances for validation, and 2640 utterances test. More details on CHiME4 data are available in [9]. The speech data are framed to 400 samples using a hamming window with 100 samples shift (75% overlapped). The parameters for the experiment are summarized in Table 1. We use the CHiME4 baseline speech enhancement method (BeamformIt [13]), the front-end of the winner in the CHiME3 challenge [1] (denoted by CGMM-MVDR), and a beamformer using DNN-based IBM [2] (denoted by DNN-IBM) for comparison. In addition, the single channel noisy speech is also considered for comparison (denoted by None).

TABLE 1: Parameters of the experiments

Parameters	Value
Training set	8738
Validation set	3280
Test set	2640
Duration of each utterance	3∼6s
Number of microphones	6
Sampling frequency	16kHz
Frame length	25ms
Frame shift	6.25ms
Window function	Hamming

A. PREPROCESS: MICROPHONE FAILURE DETECTION

The separated clean speech and noise are available for the simulated data, but they are not available for the real data. Moreover, there may be microphone failures due to the unstable connection of the device or masking by the user's hands or clothes. Fig. 4 demonstrates the spectrogram of an utterance taken from the development set, from which we can see that the fourth channel of the microphone failed to record the speech. In this case, the performance of the beamformer will deteriorate drastically. In order to obtain reliable multi-channel signals, we firstly perform microphone failure detection by checking both normalized cross-correlation coefficients and signal power of each channel. Then, we fill the failed channel with the next-to-last channel (channel 5, in most instances). Finally, an estimation of clean speech and noise for each channel is obtained using the official toolkit.

B. FEATURE EXTRACTION AND DNN TRAINING

Mel-frequency cepstrum coefficients (MFCCs) of the multi-channel signals are extracted as the input of the network. 26 MFCCs are calculated for each frame, and 3 frames of temporal context are concatenated, in 6 microphone channels. That produces a $26 \times (2 \times 3 + 1) \times 6 = 1092$ dimension vector for one input feature. These features are normalized to have zero-mean and unit variance before feeding to the network.

The settings for the DNN training are as follows. From the input layer to the output layer, the proposed network has 1092, 1024, 512, 512, 512 and two 512 units, respectively. The weights of all layers are initialized by a Xavier initializer [14], and all the biases are initialized with zeros. An Adam optimizer [15] with a fixed learning rate set to 0.001 is adopted to optimize the DNN. In order to prevent the DNN from overfitting, L2 regularization of the weight (parameter, λ, set to 10^{-5}) and dropout strategy [16] (dropout rate, p, set to 0.5) are adopted. Additionally, batch normalization (BN) [17] for all layers except the output layer is applied. The training process is stopped when the validation error does not decrease anymore in 3 epochs. The parameters of the each layers and the parameters for the training are summarized in Table 2 and Table 3, respectively. We implement the DNN training algorithm in Python with the help of 'Tensorflow' [18] and carry out the training and evaluation procedures on a workstation with two Intel Xeon E5-2630 CPUs and four GTX 1080 GPUs.

TABLE 2: Parameters of the each layer

Layer	Units	Activation	Dropout	BN
L1	1092	ReLU	Yes	Yes
L2	1024	ReLU	Yes	Yes
L3	512	ReLU	Yes	Yes
L4	512	ReLU	Yes	Yes
L5	512	ReLU	Yes	Yes
L6	521/512	Sigmoid	No	No

TABLE 3: Parameters for training

Parameteres	Value
Learning rate	0.001
L2 regularization	10^{-5}
Dropout rate	0.5
Max epochs	30
Waiting epochs	3

C. TIME-FREQUENCY MASK ESTIMATION

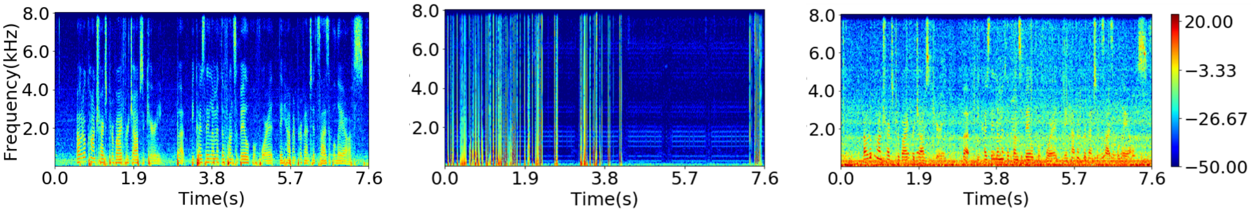
First, we evaluated the accuracy of time-frequency mask estimation. The masks estimated by the compared methods for an utterance from the evaluation set (F05_444C020G_BUS) are shown in Fig. 5. Panel (a) shows the training target calculated with (6) from channel 5, (b) shows the mask estimated by the CGMM method, (c) shows the output of DNN-IBM, and d) shows the estimated speech-weight mask of the proposed network. It can be seen that the estimated mask of the model-based method, which operates on each frequency separately, contains salt-and-pepper like noise, especially in high-frequency bins. In contrast, both the DNN-based methods do not suffer from such noise. However, the DNN-IBM, which learns a binary mask in the training stage and predicts a soft mask in the testing stage, prone to make a wrong prediction. For example, as can be seen in Fig. 5 (c), the DNN-IBM method introduces unnecessary noise at 0.6s and misses necessary speech peaks at 4.7s. In comparison, the proposed method, which directly learns a soft mask, gives a more accurate estimation, as shown in Fig. 5 (d).

Generally, we can evaluate the accuracy of mask estimation quantitatively. Yet, due to the proposed network using the mask in Fig. 5 (a) as the learning target, such comparison result is unfair and has not been provided here. We will quantitatively evaluate the performance of speech enhancement and speech recognition in the following.

D. SPEECH ENHANCEMENT

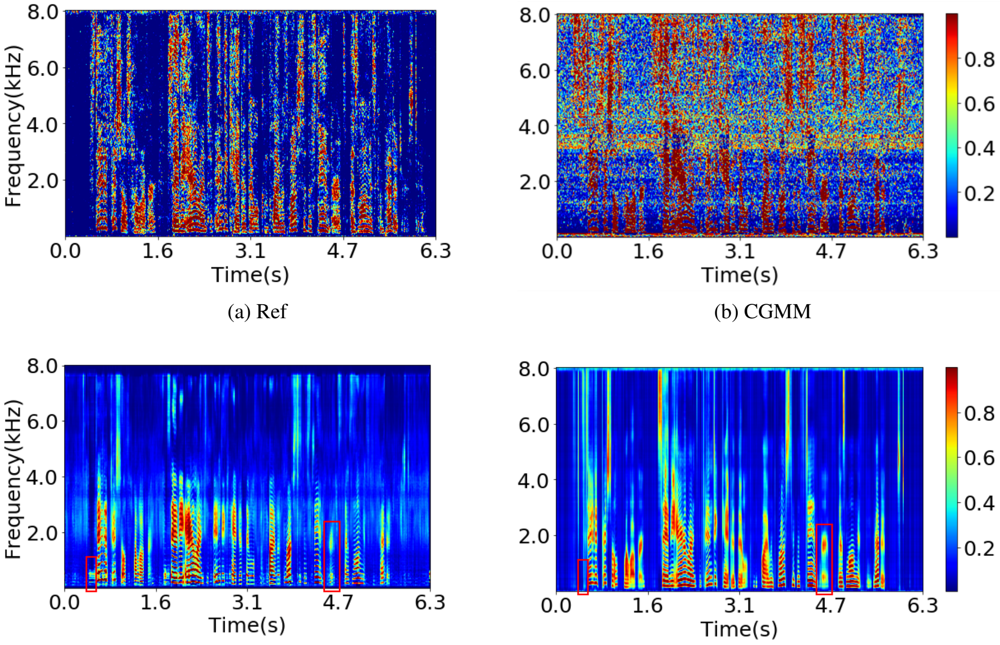
We use signal to distortion ratios (SDR) in dB [20], short-time objective intelligibility (STOI) [21], and speech quality (PESQ) [19] to evaluate the quality of the enhanced speech signal. The close-talking microphone recordings (i.e. channel 0) are considered to be the underlying clean speech.

The SDR and STOI scores for speech quality test on the development set and simulation set are shown in Table 4. The results demonstrate that, our approach outperforms the state-of-art CGMM-MVDR and DNN-IBM algorithms. For example, for the simulated data in development set, the STOI



(a) Reference channel (b) Failure channel (Channel 4) (c) Success channel (Channel 5)

FIGURE 4: Spectrogram of the utterance *F04_053C0111_BUS* taken from the development set



(a) Ref (b) CGMM

(c) DNN-IBM (d) Proposed

FIGURE 5: Comparison of the estimated time-frequency masks.

TABLE 4: SDR (dB) and STOI for speech quality test of the model-based beamforming methods

Method	Critera	Development		Evaluation	
		simu	real	simu	real
None	SDR STOI	4.03 0.86	-5.10 0.52	4.94 0.81	-5.64 0.43
BeamformIt	SDR STOI	5.48 0.88	-4.46 0.56	6.21 0.86	-4.93 0.47
CGMM-MVDR	SDR STOI	11.47 0.94	-2.24 0.59	11.83 0.93	-3.73 0.50
DNN-IBM	SDR STOI	5.42 0.89	-2.16 0.59	4.40 0.88	-5.75 0.51
Proposed	SDR STOI	11.35 0.95	-0.88 0.60	12.17 0.95	-4.29 0.51

score of CGMM-MVDR and DNN-IBM are 0.88 and 0.89 respectively, while that of the proposed method is 0.95.

The results in Table 4 are average scores over various environment conditions, i.e., bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR). It is practical to evaluate the performance of the bemforming methods in various environments. The PESQ results of the compared methods in the four environments are shown in Table 5. From the results we can see that, for the real data, the PESQ score is the highest in the PED condition while be the lowest in the BUS condition. This is because the pedestrian area is much quiet than in the bus. In addition, we can conclude that the beamformer is robust to various types of noise.

E. SPEECH RECOGNITION

Finally, speech recognition performance is evaluated using Kaldi toolkit [34]. The CHiME4 official ASR models (i.e., GMM, DNN, and DNN+RNNLM) are used in this evaluation. Word error rates (WERs in percentage) for speech recognition accuracy of the compared methods are shown in Table 6. The results show that our approach outperforms the CHiME4 front-end baseline and CGMM-MVDR on both

TABLE 6: WERs for speech recognition on the development set and simulation set

Method	ASR backend	Development			Evaluation		
		simu	real	avg	simu	real	avg
None	GMM	24.46	22.18	23.32	33.32	37.57	35.45
	DNN	15.64	14.68	15.16	24.13	27.68	25.91
	DNN+RNNLM	12.96	11.57	12.27	20.85	23.71	22.28
BeamformIt	GMM	14.34	12.98	13.66	21.33	21.80	21.57
	DNN	9.06	8.16	8.61	14.27	14.98	14.63
	DNN+RNNLM	6.73	5.79	6.26	10.93	11.50	11.22
CGMM-MVDR	GMM	11.60	11.08	11.34	15.22	17.55	16.39
	DNN	7.14	7.58	7.36	10.31	12.45	11.38
	DNN+RNNLM	5.12	5.33	5.23	7.62	9.32	8.47
DNN-IBM	GMM	11.92	10.87	11.40	14.19	14.84	14.52
	DNN	7.69	6.94	7.32	9.85	10.12	9.99
	DNN+RNNLM	5.52	4.81	5.16	7.04	7.26	7.15
Proposed	GMM	10.20	10.84	10.52	12.01	15.73	13.87
	DNN	6.63	7.19	6.91	8.19	10.73	9.46
	DNN+RNNLM	4.62	4.89	4.76	5.69	7.62	6.67

TABLE 5: PESQ scores for speech quality test of the DNN-based beamforming.

Method	Environment	Development		Evaluation	
		simu	real	simu	real
BeamformIt	BUS	2.41	2.19	2.36	2.26
	CAF	2.19	2.49	2.15	2.56
	PED	2.39	2.64	2.14	2.59
	STR	2.25	2.30	2.15	2.57
CGMM-MVDR	BUS	2.73	2.33	2.76	2.38
	CAF	2.38	2.73	2.48	2.74
	PED	2.65	2.81	2.55	2.64
	STR	2.55	2.46	2.52	2.62
DNN-IBM	BUS	2.55	2.65	2.57	2.72
	CAF	2.22	2.76	2.31	2.75
	PED	2.41	2.83	2.35	2.69
	STR	2.37	2.65	2.39	2.71
Proposed	BUS	2.78	2.71	2.79	2.57
	CAF	2.43	2.99	2.56	2.88
	PED	2.68	3.02	2.60	2.76
	STR	2.59	2.76	2.61	2.86

simulated and real data. Compared with DNN-IRM, our method has lower WERs almost in all cases, expect for the real data on the evaluation set. This is because many tricks (e.g. voiced/unvoiced splitting, bins edge nulling, and frame edge padding) have been used in the DNN-IBM approach, which might match the speech recognizer for the real data. Our method have not used these tricks. Even so, our method has lower WERs on the average.

VI. DISCUSSION

As introduced in Section IV, the proposed method uses multi-target specified IRMs as learning targets, and uses multi-channel and multi-frame observations feed to the DNN. This architecture leverages the power of microphone array to predict more reliable masks, i.e. speech-aware channel to obtain the speech mask and noise-aware channel to obtain the noise mask.

In comparison with CGMM-MVDR [1], the DNN-based methods jointly estimate all frequency bin of the mask. The difference between theses two types of methods can be seen in Fig. 5. The CGMM-MVDR, using the EM-based clustering method to perform estimation on each frequency bin independently, gives rise to salt-and-pepper like noise in the spectral mask.

The proposed method is different from the DNN-IBM method [2] in the following aspects. Firstly, the DNN architecture in [2] operates on each channel separately (i.e., the DNN learns to predict 6 channel masks), and the output mask is obtained by median operation over all channels. In contrast, our approach feeds the DNN with integral observations of the microphone array, and uses the speech-aware channel to predict the speech mask. Meanwhile, the noise-aware channel is used to predict the noise mask. Secondly, in the learning phase, the DNN in [2] uses binary mask as the learning target, while in the prediction phase, the output of the DNN is soft mask. The mismatch may introduce unnecessary noise, which can be seen in Fig. 5 (c). In contrast, our proposed method uses the soft mask as the learning target and it does not suffer from such mismatch (see Fig. 5 (d)). Thirdly, the method in [2] only uses the current frame to predict the mask, while our method uses adjacent c frames to capture the temporal context of the speech. Besides, in order to reduce the dimension of the concatenated super-vectors, we use MFCC as features rather than the spectral magnitude.

VII. CONCLUSION

In this paper, we proposed a novel robust acoustic beamformer using supervised time-frequency mask estimation. It does not need the estimation of the steering vector of the source and does not make any assumption on the distribution of the signal in the mask estimation. Thus, it does not suffer from steering vector estimation error, model mismatch caused by inaccurate knowledge on the array, and/or improper distribution assumptions on the signal and noise. We use multi-target specified IRM estimated by a multi-task DNN to obtain the covariance matrices of the target speech signal and noise. Experiments on the CHiME4 datasets demonstrated that the proposed method can achieve state-of-the-art performance both in terms of speech enhancement quality and speech recognition.

REFERENCES

[1] T. Higuchi, N. Ito, T. Yoshioka and T. Nakatani, "Robust MVDR beamforming using time-frequency masks for online/offline ASR in noise," in *Proc. ICASSP*, 2016, pp. 5210–5214.

[2] J. Heymann, L. Drude and R. Haeb-Umbach, "Neural network based spectral mask estimation for acoustic beamforming," in *Proc. ICASSP*, 2016, pp. 5210–5214.

[3] N. Q. K. Duong, E. Vincent and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[4] D. H. Tran Vu and R. Haeb-Umbach, "Blind speech separation employing directional statistics in an expectation maximization framework," in *Proc. ICASSP*, 2010, pp. 241–244.

[5] H. Erdogan, J. R. Hershey, and S. Watanabe, *et al.*, "Improved MVDR beamforming using single-channel mask prediction networks," in *Proc. Interspeech*, 2016, pp. 1981–1985.

[6] E. Warsitz and R. Haeb-Umbach, "Blind acoustic beamforming based on generalized eigenvalue decomposition," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 5, pp. 1529–1539, 2007.

[7] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. ed. Boca Raton, FL, USA: CRC, 2013.

[8] Y. Wang, A. Narayanan and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

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

[10] M. L. Seltzer, and J. Droppo, "Multi-task learning in deep neural networks for improved phoneme recognition," in *Proc. ICASSP*, 2013, pp. 6965–6969.

[11] L. Deng, G. Hinton and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *Proc. ICASSP*, 2013, pp. 8599–8603.

[12] D. S. Williamson, Y. Wang and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, March 2016.

[13] X. Anguera, C. Wooters and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2011–2022, 2007.

[14] G. Xavier and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Thirteenth Int. Conf. Artificial Intelligence Statistics*, pp. 249–256. 2010.

[15] D. Kinga, and J. Ba Adam, "A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.

[16] N. Srivastava, G. Hinton, and A. Krizhevsky, *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp.1929–1958, 2014.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. ICML*, pp. 448–456. 2015.

[18] M. Abadi and A. Agarwal, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint*, arXiv:1603.04467, 2016.

[19] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 229–238, 2008.

[20] E. Vincent, R. Gribonval, and C. FÃľvotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[21] C. H. Taal, R. C. Hendriks, and R. Heusdens, "An algorithm for intelligibility prediction of timeâĂŞfrequency weighted noisy speech," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[22] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Trans. Audio Speech Lang. Process.*, vol. 47, no. 10, pp. 2677–2684, 1999.

[23] K. Kumatani, J. McDonough, B. Rauch, *et al.*, "Beamforming with a maximum negentropy criterion," *IEEE Trans. Audio Speech Lang. Process.*, vol. 17, no. 5, pp. 994–1008, 2009.

[24] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 2, pp. 260–276, 2010.

[25] Y. Wang and D. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. ICASSP*, 2015, pp. 4390–4394.

[26] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 3, pp. 483–492, 2016.

[27] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. ICASSP*, 2013, pp. 7092–7096.

[28] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 20, no. 10, pp. 1702–1726, 2018.

[29] J. Tu, Y. Xia, and S. Zhang, "A complex-valued multichannel speech enhancement learning algorithm for optimal tradeoff between noise reduction and speech distortion," *Neurocomputing*, vol. 267, pp. 333–343, 2017.

[30] J. Tu and Y. Xia, "Effective Kalman filtering algorithm for distributed multichannel speech enhancement," *Neurocomputing*, vol. 275, pp. 144–154, 2018.

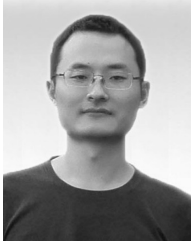
[31] X. Xiao, S. Watanabe, H. Erdogan, *et al.*, "Deep beamforming networks for multi-channel speech recognition," in *Proc. ICASSP*, 2016, pp. 5745–5749.

[32] S. Araki, T. Hayashi, M. Delcroix, *et al.*, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. ICASSP*, 2015, pp. 116–120.

[33] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, 2016.

[34] D. Povey, A. Ghoshal, G. Boulianne, *et al.* "The Kaldi speech recognition toolkit," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, 2011.

WENBIN JIANG received the M.S. degree in electronic engineering from Hangzhou Dianzi University, China, in 2012, and the Ph.D. degree in communications and information engineering from Shanghai Jiao Tong University in 2018. His research interests include speech signal processing and machine learning.



FEI WEN (M'15) received the B.S. degree from the University of Electronic Science and Technology of China (UESTC) in 2006, and the Ph.D. degree in communications and information engineering from UESTC in 2013. Since December 2012, he has been a lecturer at the Air Force Engineering University. Now he is a research fellow of Department of Electronic Engineering in Shanghai Jiao Tong University. His main research interests are nonconvex optimization, large-scale numerical optimization, statistical signal processing, and machine learning.

PEILIN LIU (M'99) received the Ph.D. degree from the University of Tokyo majoring in Electronic Engineering in 1998 and worked there as a Research Fellow in 1999. From 1999 to 2003 she worked as a Senior Researcher for Central Research Institute of Fujitsu, Tokyo. Her research mainly focuses on Signal processing, low power computing architecture, Application-oriented SoC design and Verification. She is now a professor of Department of Electronic Engineering in Shanghai Jiao Tong University, executive director of Shanghai Key Laboratory of Navigation and Location Based Service and responsible for a series of important projects, such as BDSSoC Platform Development, Low power and High-Performance communication DSP. Prof. Liu is the chair of Shanghai chapter of IEEE Circuit and System.