Online LSTM-based Iterative Mask Estimation for Multi-Channel Speech Enhancement and ASR

Yan-Hui Tu* and Jun Du* and Nan Zhou* and Chin-Hui Lee†
* University of Science and Technology of China, Hefei, Anhui, China
E-mail: tuyanhui@mail.ustc.edu.cn, jundu@ustc.edu.cn, zhounan1@mail.ustc.edu.cn
† Georgia Institute of Technology, Atlanta, Georgia, USA
E-mail: chinhui.lee@ece.gatech.edu

***Abstract*—Accurate steering vector estimation is the key point for a beamformer which suppresses the background noise to improve the noisy speech quality and intelligibility. Recently, time-frequency masking approach, which estimates the steering vectors that are utilized for a beamformer, is popular in this field. In particular, we have proposed an iterative mask estimation (IME) approach to improve the complex Gaussian mixture model (CGMM) based beamforming and yield the best system for multi-channel ASR in CHiME-4 challenge [1]. And in [2], we also demonstrated that our algorithm could improve the speech quality (PESQ) and intelligibility (STOI) for multi-channel speech enhancement. In this study, we focus on the online processing of our IME algorithm for multi-channel speech enhancement and ASR, which achieves comparable performance to the offline version. In addition, a regression long short-term memory recurrent neural network (LSTM-RNN) for a multiple-target joint learning is utilized, denoted as LSTM-MT, to replace two separate models in [2]. Experiments on the CHiME-4 simulation data show that the online IME algorithm can improve the enhancement performance, e.g., PESQ from 2.18 to 2.58 and STOI from 86.85 to 94.51, which is comparable to those obtained by offline IME. Furthermore, the LSTM-MT based post-processing can achieve an additional PESQ improvement from 2.58 to 2.71. And experiments on the CHiME-4 real data show that the online IME approach outperforms the online CGMM-based approach, with a relative word error reduction (WER) of 14.49%.**

I. Instructions

Recently, hands-free speech communication is more and more popular for many applications, such as multi-microphone portable devices and automatic speech recognition (ASR) systems, due to the provided convenience and flexibility. However, the speech signals recorded by distant microphones are often corrupted by reverberation and background noise, leading to considerable degradation in speech quality, particularly at low signal-to-noise ratios (SNRs). The key point of speech enhancement algorithms is to reduce noise without large distortions to the target speech. For multi-channel speech enhancement, representative algorithms include multi-channel Wiener filtering [3], blind source separation [4], and beamforming [5], [6]. And beamforming is a popular approach, e.g., the minimum variance distortionless response (MVDR) beamformer. How to construct a steering vector that represents the acoustic propagation [7] is the key to achieving a high-quality beamformer. Conventionally, some *a priori* knowledge is used to construct the steering vector, e.g., the geometry of the microphone array and the direction of arrival (DOA) information. But its robustness often becomes a problem in real-life environments where the acoustic propagation information is not known and difficult to be estimated accurately. In [6], the authors provide a new form of using the time-frequency (T-F) masks estimated by a complex Gaussian mixture model (CGMM) to steer a beamformer, which is demonstrated to be quite effective for ASR in real-life situation.

On the other hand, deep learning techniques are becoming increasingly popular in speech recognition areas [8], [9]. Different deep neural network (DNN) architectures have been adopted in single-channel speech enhancement for ASR, and they have demonstrated a significant increase in recognition performance [10], [11], [12], [13]. Some preliminary studies on using deep learning approaches for multi-channel speech enhancement have also been conducted. In [14], [15], the signals obtained using multi-channel speech enhancement algorithms were directly used as the input signals for neural-network-based enhancement models. In [16], bidirectional long short-term memory (BLSTM) [17] was adopted to estimate signal statistics to steer the beamformer for multi-channel speech enhancement. It was also demonstrated in [18] that DNN-based source spectra estimation is helpful for steering a multi-channel filter. In [19], they proposed multi-channel enhancement joint with acoustic modeling in a DNN framework. The raw time-domain waveform was directly modeled by beamforming, which leveraged upon differences in the fine time structure of the signal at different microphones to filter energy arriving from different directions. In [20], an end-to-end framework was proposed by encompassing microphone array signal processing for noise suppression within the acoustic encoding network, allowing the beamforming components to be optimized jointly within the recognition architecture.

In this study, we focus on the online processing of our iterative mask estimation (IME) algorithm [1] for multi-channel speech enhancement and ASR. This work is comprehensively extended from our recent paper [2] with new contributions listed as follows. First, a regression long short-term memory recurrent neural network (LSTM-RNN) for a multiple-target joint learning is utilized, denoted as LSTM-MT, to replace two separate models in [2]. Second, for our online algorithm, the estimated spatial correlation matrix of target and noise of previous batch is adopted to steer the beamformer of current batch frame by frame. Accordingly the beamformed speech is obtained only with current frame plus three-frame delay, while in [21], the online CGMM-based beamformed speech is obtained in a manner of the batch delay with quite a few frames. Finally, the spatial correlation matrix of target and noise at each batch is estimated by the combination of CGMM-based and LSTM-based approach. The experiments on the CHiME-4 real data sets show that the online IME approach outperforms the online CGMM-based approach, with a relative word error reduction (WER) of 14.92%.

The remainder of this paper is organized as follows. In Section II, we present an overview of the related work. In Section III, a detailed description of the proposed online IME approach and LSTM-based post-processing is given. Section IV shows the enhancement and ASR performance of our proposed approach on the CHiME-4 challenge. Finally, we summarize our findings in Section V.

II. Related Work

As shown in Fig. 1, our proposed offline IME approach in [2] was used for the multi-channel speech enhancement. At the training stage, two LSTM-based regression models, denoted as LSTM-IRM and LSTM-DM, were trained for the ideal ratio mask (IRM) estimation and direct mapping of the clean speech, respectively. At the test stage, the beamformed speech and T-F mask of the whole utterance were estimated by CGMM-based beamforming. Then, the IRM estimated by the trained LSTM-IRM model was utilized to improve the mask estimation. Next, the improved mask was adopted to steer the beamforming. Finally, the beamformed speech was processed by the trained LSTM-DM model.

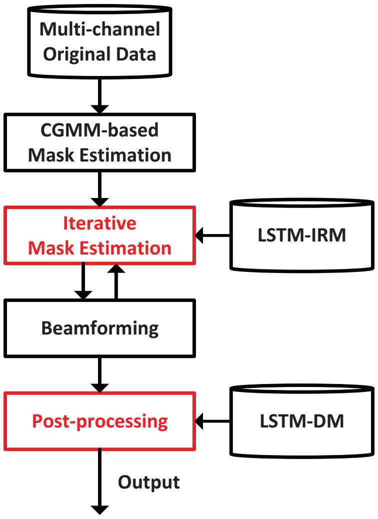
Multi-channel Original Data
CGMM-based Mask Estimation
Iterative Mask Estimation
LSTM-IRM
Beamforming
Post-processing
LSTM-DM
Output

Fig. 1. A block diagram of the offline IME approach.

III. Online IME-based Beamformer and LSTM-based Post-processing

A. Mask-based beamforming

We use minimum variance distortionless response (MVDR) beamformer which maximizes the SNR of the beamformer output in each frequency bin k, leading to the beamformer coefficients:

$$\boldsymbol{w}(k) = \frac{\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(k)\boldsymbol{g}(k)}{\boldsymbol{g}^H(k)\boldsymbol{R}_{\boldsymbol{nn}}^{-1}(k)\boldsymbol{g}(k)}, \tag{1}$$

where $\boldsymbol{g}(k)$ is the signal propagation vector, which is in the same form as the so-called steering vector in the literature of array beamforming [7]; $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ and $\boldsymbol{R}_{\boldsymbol{nn}}(k)$ are the spatial correlation matrix of target and noise, respectively. In [6], an approach using a speech spectral model based on CGMM was proposed to estimate the time-frequency masks, denoted as $M_{\text{CGMM}}(k,l)$. The parameters of the CGMM are full-rank spatial correlation matrices, which provide some flexibility to address the spatial fluctuation of the steering vector.

B. Architecture of LSTM-MT model

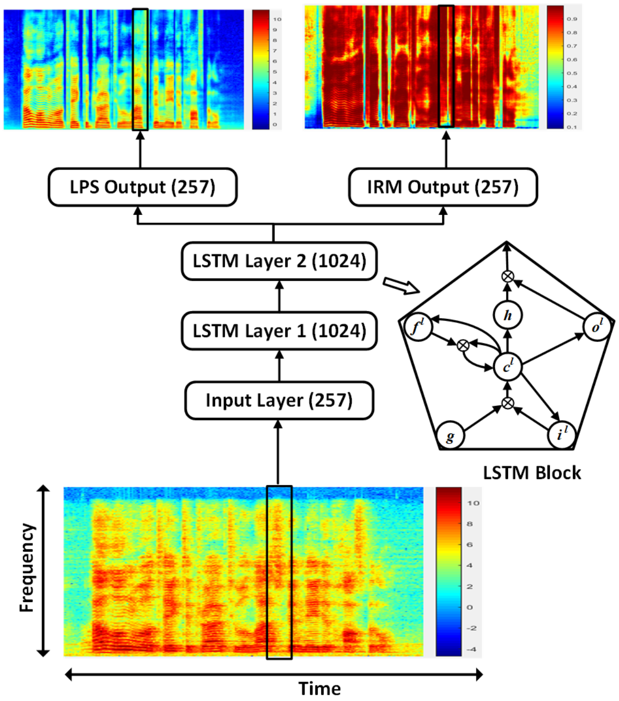

Fig. 2. The architecture of LSTM-MT.

Fig. 2 shows the architecture of the LSTM-based multi-target learning, which can be trained to learn the complex transformation from the noisy log-power spectra (LPS) features to clean LPS features and IRM, denoted as LSTM-MT. Acoustic context information along both the time axis and frequency axis (with full frequency bins) can be fully exploited by the LSTM to obtain a good mask estimation in adverse environments. The estimated IRMs are restricted to be in the range between zero and one, which can be directly used to represent the speech presence probability. The IRM as the learning target is defined as follows.

$$M_{\text{ref}}(k,l) = S_{\text{PS}}(k,l) / \left[S_{\text{PS}}(k,l) + N_{\text{PS}}(k,l)\right], \tag{2}$$

where $S_{\text{PS}}(k,l)$ and $N_{\text{PS}}(k,l)$ are clean and noise versions of power spectral features at the T-F unit (k,l).

Because the training of this LSTM-MT model requires a large amount of time-synchronized stereo-data with the IRM and LPS of training data pairs, the training data are synthesized by adding different types of noise to the clean speech utterances with different SNR levels. Note that the specified SNR levels in the training stage are expected to address the problem of SNR variation in the test stage with real speech data. To train the LSTM-MT model, supervised fine-tuning is used to minimize the mean squared error (MSE) between both of the LSTM-LPS output $\hat{X}_{\text{LPS}}(k,l)$ and the reference LPS $X_{\text{ref}}(k,l)$, and the LSTM-IRM output $\hat{M}_{\text{IRM}}(k,l)$ and the reference IRM $M_{\text{ref}}(k,l)$, which is defined as

$$\begin{aligned} E_{\text{MT}} = \sum_{k,l} \Big[& (\hat{X}_{\text{LPS}}(k,l) - X_{\text{ref}}(k,l))^2 \\ & + (\hat{M}_{\text{IRM}}(k,l) - M_{\text{ref}}(k,l))^2 \Big]. \end{aligned} \tag{3}$$

This MSE is optimized using the stochastic gradient descent based back-propagation method in a mini-batch mode.

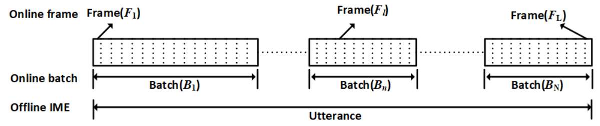

Fig. 3. The illustration of online and offline IME algorithm.

C. Online LSTM-MT based IME

In [21], the CGMM-based beamforming was extend to online version for ASR, but the beamformed speech was obtained after a batch time-delay. To solve this problem, we give a new form of our proposed online LSTM-based IME algorithm for multi-channel speech enhancement frame by frame. For an observed signal, we consider its previous frames as a sequence of block-batches as show in Fig. 3. $n = 1, 2, , N$ is the batch index, and $l = 1, 2, , L$ is the frame index. In the beginning of each utterance, the first batch (B_1) usually contains more frames than the successive batches to make sure this batch contains target speech, and the offline IME algorithm [1] is utilized to estimate the initial $\boldsymbol{R_{xx}}(k, 1)$ and $\boldsymbol{R_{nn}}(k, 1)$ with LSTM-MT. To avoid the delay, the noisy speech at B_1 is processed by LSTM-MT model with direct mapping output. For the n-th ($n > 1$) batch B_n, the $\boldsymbol{R_{xx}}(k, n)$ and $\boldsymbol{R_{nn}}(k, n)$ are recursively obtained by $\boldsymbol{R_{xx}}(k, n-1)$ and $\boldsymbol{R_{nn}}(k, n-1)$ according to Eq. (5) and Eq. (6). And the $\mathbf{M}(k, l)$ is obtained by the LSTM-MT model with IRM output. α is set to 0.75. Accordingly, the estimated $\boldsymbol{R_{xx}}(k, n)$ and $\boldsymbol{R_{nn}}(k, n)$ in the current batch is mainly dependent on $\boldsymbol{R_{xx}}(k, n-1)$ and $\boldsymbol{R_{nn}}(k, n-1)$ in the previous batch due to the continuity of speech. Finally, the beamformed speech in the $(n+1)$-th batch is obtained based on the estimated $\boldsymbol{R_{xx}}(k, n)$ and $\boldsymbol{R_{nn}}(k, n)$ frame by frame.

$$\beta = \alpha + (1 - \alpha)\mathbf{M}(k, l) \tag{4}$$

$$\begin{aligned}\boldsymbol{R_{xx}}(k, n) =& \beta \boldsymbol{R_{xx}}(k, n-1) \\ &+ (1-\beta) \sum_{l \in B_n} \mathbf{M}(k, l)\mathbf{y}(k, l)\mathbf{y}^{\mathrm{H}}(k, l)\end{aligned} \tag{5}$$

$$\begin{aligned}\boldsymbol{R_{nn}}(k, n) =& \beta \boldsymbol{R_{nn}}(k, n-1) \\ &+ (1-\beta) \sum_{l \in B_n} (1 - \mathbf{M}(k, l))\mathbf{y}(k, l)\mathbf{y}^{\mathrm{H}}(k, l)\end{aligned} \tag{6}$$

D. LSTM-MT based post-processing

In this section, we discuss the LSTM-based post-processing for beamformed speech. For the beamformer, the original purpose is to improve the SNR without destroying the target speech, and it is difficult to completely eliminate the background noise. While for the LSTM-MT based regression model with direct mapping output, it can eliminate the background noise well, but the target speech maybe destroyed at low SNR situations. So, in this study, the direct mapping output of LSTM-MT is used for post-processing. And in the following Section IV, the experiments can well confirm the above analysis. The whole procedure of our proposed approach is presented as Algorithm 1.

IV. Experimental Evaluation

We conduct the experimental evaluation of our proposed approach using the CHiME-4 data [22], which was designed to investigate real-world scenarios where a person was talking to a mobile tablet device equipped with 6 microphones in a variety of adverse environments. Four conditions were selected: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). For each case, two types of noisy speech data were provided: RealData and SimData. RealData were collected from speakers reading the same sentences from the WSJ0 corpus [23] in the four conditions. SimData were constructed by mixing clean utterances with environmental noise recordings using the techniques described in [24]. CHiME-4 offers three tasks (1-channel, 2-channel, and 6-channel) with different testing scenarios. In this paper, we focus only on the 6-channel case to make the paper concise. The readers can refer to [22] for more detailed information regarding CHiME-4.

Algorithm 1 Online iterative mask estimation and post-processing

```
Input: Multi-channel noisy speech, denoted as $X(k,l)$.
Output: Online IME based beamformed speech $\hat{X}_{\mathrm{IME}}(k,l)$ for ASR
    or post-processing speech $\hat{X}_{\mathrm{PP}}(k,l)$ for enhancement.
for all short-time frames $l = 1, 2, ..., N$ do
    for all frequency bins $k = 1, 2, ..., F$ do
        if $l \in B_1$ then
            Feed $X(k,l)$ into LSTM-MT and obtain estimated speech
            $\hat{X}_{\mathrm{PP}}(k,l)$ and IRM $\hat{M}_{\mathrm{IRM}}(k,l)$.
            if $l$ is the last frame of $B_1$ then
                Compute the initial $\boldsymbol{R_{xx}}(k,1)$ and $\boldsymbol{R_{nn}}(k,1)$ using
                the $\hat{M}_{\mathrm{IRM}}(k,l)$ with offline IME algorithm.
            end if
        end if
        if $l \in B_n$ $(n > 1)$ then
            Use estimated $\boldsymbol{R_{xx}}(k,n-1)$ and $\boldsymbol{R_{nn}}(k,n-1)$ to obtain
            the beamformed speech $\hat{X}_{\mathrm{IME}}(k,l)$ for ASR.
            Feed $\hat{X}_{\mathrm{IME}}(k,l)$ into LSTM-MT and obtain post-
            processing speech $\hat{X}_{\mathrm{PP}}(k,l)$ for enhancement.
            if $l$ is the last frame of $B_n$ then
                Obtain $\boldsymbol{R_{xx}}(k,n)$ and $\boldsymbol{R_{nn}}(k,n)$ according to Eq. (5)
                and Eq. (6).
            end if
        end if
    end for
end for
```

For front-end configurations, speech waveform is sampled at 16 kHz, and the corresponding frame length is set to 512 samples (or 32 msec) with a frame shift of 128 samples. A short-time Fourier transform (STFT) analysis is used to compute the DFT of each overlapping windowed frame. To train the LSTM-MT model, the 257-dimensional feature vector was used for both LPS and IRM targets. The PyTorch tools [25] was used for training. The LSTM-MT architecture is 257-1024*2-514, namely 257 dimension for LPS input features, 2 LSTM layers with 1024 cells for each layer, and 514 nodes for the output T-F LPS and IRM, respectively. The model parameters were randomly initialized. The learning rate for the first ten epochs was initialized as 0.01, then decreased by 0.9 after each epoch, and the number of epochs was 30. To build the training data, clean speech was derived from the WSJ0 corpus [23], and the 4 type noise provided by CHiME-4 in [26] were selected as our noise database. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, were corrupted with the above mentioned 4 noise types at three SNR levels (-5dB, 0dB and 5dB) to build a 36-hour training set, consisting of pairs of clean speech and noisy speech utterances.

A. Enhancement experiments

In this section, we evaluated the online IME algorithm for speech enhancement task. The size was set to 1000 ms for the first batch and 320 ms for the succeeding batches to ensure that the first batch contained the target speech signal. Specifically, the enhanced speech at the first batch was obtained by LSTM-MT with direct mapping output, so we could obtain the enhanced speech signal only with a 32 ms delay, while for the online CGMM-based beamformer in [21], the delay of the algorithm was 1000 ms in this configuration.

TABLE I
THE AVERAGE PESQ AND STOI COMPARISON OF DIFFERENT SYSTEMS ON THE TEST SETS OF SIMDATA.

Measure	Methods	BUS	CAF	PED	STR	AVG
PESQ	CH5	2.32	2.09	2.13	2.19	2.18
	Online CGMM_batch	2.78	2.53	2.63	2.59	2.63
	Online CGMM_frame	2.75	2.49	2.60	2.56	2.60
	Offline IME	2.74	2.50	2.61	2.54	2.59
	Online IME	2.73	2.48	2.60	2.52	2.58
	Online +PP	2.81	2.59	2.76	2.69	2.71
STOI(%)	CH5	88.35	88.49	87.23	86.33	86.85
	Online CGMM_batch	96.63	93.84	94.26	92.91	94.41
	Online CGMM_frame	96.57	93.77	94.16	92.87	94.34
	Offline IME	97.01	94.43	94.65	93.24	94.83
	Online IME	96.77	93.98	94.29	93.02	94.51
	Online +PP	95.41	93.74	94.06	92.71	93.98

Table I presents the performance comparison of online and offline beamformers on the test sets of SimData. First, "CH5" denotes the original speech from channel 5. "Online CGMM_batch" denotes the online CGMM-based beamformer reproduced according to [21], which the beamformed speech is obtained with a batch delay. "Online CGMM_frame" denotes the online CGMM-based beamformer described in Section III, which the beamformed speech is obtained with a frame delay. We could observe that "Online CGMM_frame" approach improved the performance, e.g., PESQ from 2.18 to 2.60 and STOI from 86.85 to 94.34 in average, compared to "CH5". Also, the performance of "Online CGMM_frame" was comparable to that of "Online CGMM_batch", while our proposed algorithm was with a smaller delay. Second, "Offline IME" denotes the offline IME beamformer in [2] and "Online IME" denotes the online IME beamformer proposed in Section III. The proposed online IME beamformer could obtain the comparable performance to offline IME beamforer, e.g., PESQ from 2.59 to 2.58 and STOI from 94.83 to 94.51. Furthermore, the LSTM-based post-processing ("+PP") could achieve an additional PESQ improvement from 2.58 to 2.71 over the online IME approach across all test sets, which demonstrated the effectiveness of LSTM-based post-processing.

B. ASR experiments

In this section, we evaluated the online IME algorithm for ASR task. The baseline ASR system officially provided in [22] was used to evaluate the different beamformers on the test sets of RealData. The acoustic model is a DNN-HMM discriminatively trained with the sMBR criterion [27]. The input of the DNN-HMM is a 440-dimensional feature vector extracted from channel 5, consisting of a 40-dimensional fMLLR [28] with an 11-frame expansion. The language models are 5-gram with Kneser-Ney (KN) smoothing [29] for the first-pass decoding and the simple RNN-based language model [30] for rescoring.

Table II presents the WER comparison of different beamformers averaged on the test sets of RealData. First, we could observe that "Online CGMM_batch" and "Online CGMM_frame" approaches improved the performance, e.g., the relative WER reductions were 63.03% and 62.22%, respectively, compared with "CH5". Second, although the "Online IME" approach generated an average PESQ and STOI similar to that of the "Online CGMM_frame" approach, the online IME approach provided an additional 14.49% relative WER reduction. This result indicates that our proposed online IME approach is quite effective to ASR. Third, although the proposed online IME underperformed offline IME for ASR with a relative WER increase of 5.9%, the enhanced speech could be obtained in an online manner, namely without the whole utterance provided. Finally, although "+PP" could improve the PESQ performance, it degraded the ASR performance, e.g., WER from 7.61% ("Online IME") to 8.25% ("+PP"). The results also confirm our previous analysis in Section III that direct mapping method suppresses the background noise well, but destroys the target speech at low SNR situations.

TABLE II
WER (%) COMPARISON OF DIFFERENT BEAMFORMERS AVERAGED ON THE TEST SETS OF REALDATA.

Measure	Methods	BUS	CAF	PED	STR	AVG
WER	CH5	36.10	24.45	19.39	14.29	23.56
	Online CGMM_batch	12.77	6.79	7.21	8.09	8.71
	Online CGMM_frame	12.94	6.97	7.55	8.15	8.90
	Offline IME	10.12	6.08	6.32	6.15	7.16
	Online IME	10.53	6.34	7.01	6.56	7.61
	+PP	11.84	6.87	7.54	6.76	8.25

V. CONCLUSION

In this paper, we extend our offline IME approach and post-processing based on two LSTM regression models, LSTM-IRM and LSTM-DM, to online IME approach and post-processing based on one LSTM-based multi-target learning regression model with two output, denoted as LSTM-MT. First, the proposed approach utilizes the estimated $\boldsymbol{R_{xx}}$ and $\boldsymbol{R_{nn}}$ at the previous batch to steer the beamformer of current batch frame by frame. So the beamformered speech is obtained only with current frame plus three frame shift delay, while in [21], the beamformed speech is obtained with a batch delay. Second, although the online IME approach generates an average PESQ and STOI similar to those of the online CGMM-based approach, the online IME approach can provide an additional 14.49% relative WER reduction. Finally, because direct mapping method can suppress the background noise well, but it may destroy the target speech at low SNR situations. In the future, the detailed analysis of the influence of initial batch size and successive batch size on the online IME performance will be explored, and also more powerful neural network regression models will be utilized, e.g., BLSTM and convolutional neural network (CNN).

VI. ACKNOWLEDGEMENT

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and MOE-Microsoft Key Laboratory of USTC, and Huawei Noah's Ark Lab.

References

[1] Y.-H. Tu, J. Du, L. Sun, F. Ma, and C.-H. Lee, "On design of robust deep models for chime-4 multi-channel speech recognition with multiple configurations of array microphones," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2017.

[2] Y.-H. Tu, J. Du, L. Sun, and C.-H. Lee, "Lstm-based iterative mask estimation and post-processing for multi-channel speech enhancement," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017.

[3] B. Cornelis, M. Moonen, and J. Wouters, "Performance analysis of multichannel wiener filter-based noise reduction in hearing aids under second order statistics estimation errors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1368–1381, 2011.

[4] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 1, pp. 120–134, 2005.

[5] A. Krueger, E. Warsitz, and R. Haebumbach, "Speech enhancement with a gsc-like structure employing eigenvector-based transfer function ratios estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 206–219, 2011.

[6] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr beamforming using time-frequency masks for online/offline asr in noise," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[7] B. D. Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE Signal Processing Magazine*, vol. 10, no. 3, pp. 4–24, 1988.

[8] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. W. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, p. 82, 2012.

[9] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[10] J. Du, Y. Tu, L. Dai, and C. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 8, pp. 1424–1437, 2016.

[11] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[12] Y. Tu, J. Du, L. Dai, and C. Lee, "Speech separation based on signal-noise-dependent deep neural networks for robust speech recognition," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP)*, 2015.

[13] T. Gao, J. Du, L. Dai, and C. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP)*, 2015.

[14] T. Gao, J. Du, Y. Xu, C. Liu, L. Dai, and C. Lee, "Joint training of dnns by incorporating an explicit dereverberation structure for distant speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, p. 86, 2016.

[15] F. Weninger, J. R. Hershey, J. L. Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. IEEE Global Conf. Signal and Information Process.(GlobalSIP)*, 2014, pp. 577–581.

[16] J. Heymann, L. Drude, A. Chinaev, and R. Haeb-Umbach, "Blstm supported gev beamformer front-end for the 3rd chime challenge," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[17] W. Zaremba, I. Sutskever, and O. Vinyals, "Recurrent neural network regularization," *CoRR*, vol. abs/1409.2329, 2014. [Online]. Available: http://arxiv.org/abs/1409.2329

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

[19] T. N. Sainath, R. J. Weiss, K. W. Wilson, B. Li, A. Narayanan, E. Variani, M. Bacchiani, I. Shafran, A. Senior, K. Chin, A. Misra, and C. Kim, "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, May 2017.

[20] T. Ochiai, S. Watanabe, T. Hori, J. R. Hershey, and X. 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, Dec 2017.

[21] T. Nakatani, T. Nakatani, T. Nakatani, T. Nakatani, T. Nakatani, T. Nakatani, T. Nakatani, T. Nakatani, T. Nakatani, and T. Nakatani, "Online mvdr beamformer based on complex gaussian mixture model with spatial prior for noise robust asr," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 4, pp. 780–793, 2017.

[22] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, 2016.

[23] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csri (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[24] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933—1950, 2007.

[25] PyTorch, "https://pytorch.org."

[26] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Automat. Speech Recognition and Understanding Workshop.(ASRU)*, 2015.

[27] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2013, pp. 2345–2349.

[28] M. J. F. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[29] R. Kenser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP)*, vol. 1, no. 181–184, Detroit, MI, USA, May 1995.

[30] T. Mikolov, M. Karafiat, and L. Burget, "Recurrent neural network based language model," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, no. 1045–1048, Chiba, Japan, Sep. 2010.