LSTM-Based Iterative Mask Estimation and Post-Processing for Multi-Channel Speech Enhancement

Yan-Hui Tu* and Jun Du* and Lei Sun* and 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, sunlei17@mail.ustc.edu.cn
‡ Georgia Institute of Technology, Atlanta, Georgia, USA
E-mail: chinhui.lee@ece.gatech.edu

***Abstract*—Recently, we propose an iterative mask estimation (IME) approach to improve the conventional complex Gaussian mixture model (CGMM) based beamforming and yield the best multi-channel speech recognition accuracy in CHiME-4 challenge. In this study, we focus on multi-channel speech enhancement and present a novel approach via long short-term memory (LSTM) based IME and post-processing. First, an LSTM is adopted to estimate the ideal ratio mask (IRM) to improve the mask estimated by a CGMM. Then, the improved mask is used to derive a beamformer. Finally, the IME-based beamformed speech is processed by the LSTM-based regression model. Experiments on the CHiME-4 simulation data show that LSTM-based IME approach can improve the PESQ performance comparing to unprocessed signals, with relative PESQ improvements of 17.33% and 13.89%, and the LSTM-based post-processing can further yield performance gains based on the IME approach, with relative PESQ improvements of 11.42% and 10.00% for 6-channel and 2-channel cases, respectively.**

I. Instructions

Multi-channel enhancement in distant-talking scenarios based on the use of microphone arrays has become an important part of everyday life with the emergence of speech-enabled applications on multi-microphone portable devices due to its 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). Speech enhancement algorithms that reduce noise without much damaging the target speech are therefore desired for improving the speech quality. For multi-channel speech enhancement, representative algorithms in this category include multi-channel Wiener filtering [1], blind source separation [2], and beamforming [3], [4]. And beamforming is a popular approach,for example, the minimum variance distortionless response (MVDR) beamformer. A key to achieving a high-quality beamformer is how to construct a steering vector that represents the acoustic propagation [5]. 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 estimate accurately. In [4], a method was developed to steer a beamformer using the time-frequency (T-F) masks estimated by a complex Gaussian mixture model (CGMM), which was demonstrated to be beneficial to ASR in real-life scenarios.

Deep learning techniques are becoming increasingly popular in many speech research areas, notably ASR [6]. In [7], [8], deep neural networks (DNNs) were utilized for single-channel enhancement and shown to be superior to some early speech enhancement algorithms in improving some objective measures, such as short-time objective intelligibility (STOI) [9] and segmental SNR (SSNR, in dB) [10]. Different neural network architectures have been adopted in single-channel speech enhancement for ASR, and they have demonstrated a significant increase in ASR performance [11], [12], [13]. The input features of these approaches are magnitude or log-magnitude spectra in the short-time-Fourier-transform (STFT) domain [14]. The ideal ratio mask (IRM) [15] has also been shown to obtain a good speech enhancement performance.

In this paper, we focus on the IME approach [16] to beamforming by leveraging upon information obtained from long short-term memory (LSTM) based ideal ratio mask (IRM) estimation, and single channel post-processing based on the LSTM-based regression model for multi-channel speech enhancement. First, a deep learning approach is used to estimate signal statistics, e.g., IRM, in order to improve the mask estimated by a complex Gaussian mixture model (CGMM). Then, the improved mask is used to derive a beamformer. Finally, the IME-based beamformed speech is processed by the LSTM-based single channel regression model.

The remainder of this paper is organized as follows. In Section II, we present an overview of the system. In Section III, we present a detailed description of our proposed IME approach and LSTM-based post-processing. Section IV presents the enhancement performance of our proposed approach on the CHiME-4 Challenge. Finally, we summarize our findings in Section VI.

II. The Proposed Framework

A block diagram of the proposed deep learning framework is shown in Fig. 1. "LSTM-IRM" and "LSTM-DM" denote the IRM estimation and feature direct mapping using the LSTM regression model, respectively. For the IRM estimation, the LSTM-IRM model is trained using the log-power spectral (LPS) features of training data as input features and the IRM as reference. And the LSTM-DM is trained using the LPS features of training data as input features and the clean features as reference. The LPS features offering perceptually relevant parameters are adopted [10]. IRM is used to represent the speech presence probability at each time-frequency point in speech separation [17].

The beamformer process of test data is divided into four successive steps, namely, beamforming initialization, LSTM-based signal statistics (IRM) estimation, beamforming, LSTM-based post-processing. First, beamformed speech from random channel data is initialized and a T-F mask of test speech is obtained by online CGMM-based beamforming. Then, the IRM estimated by a well-trained LSTM-IRM model is used to improve the initial mask where the LSTM-IRM model uses the LPS features of the initial beamformed speech. Next, the improved mask is adopted to steer the beamformer, thereby obtaining the beamformed speech. Finally, the beamformered speech is processed by the LSTM-DM model. The detail is presented in the following subsections.

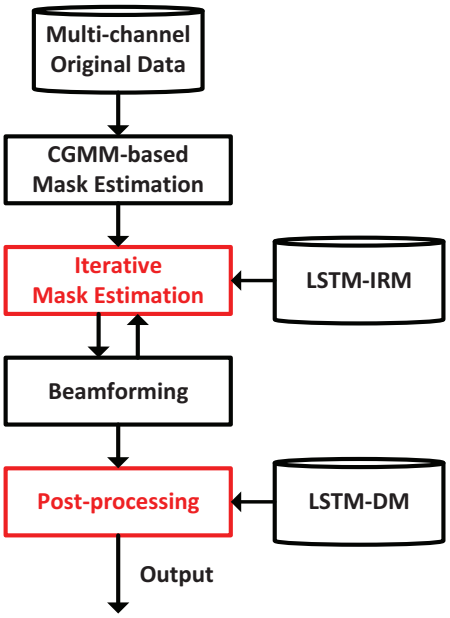

Fig. 1. A block diagram of the proposed approach.

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

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

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

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 [5]; $\boldsymbol{R}_{\boldsymbol{xx}}(k)$ and $\boldsymbol{R}_{\boldsymbol{nn}}(k)$ are the spatial correlation matrix of target and noise, respectively. In [4], 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.

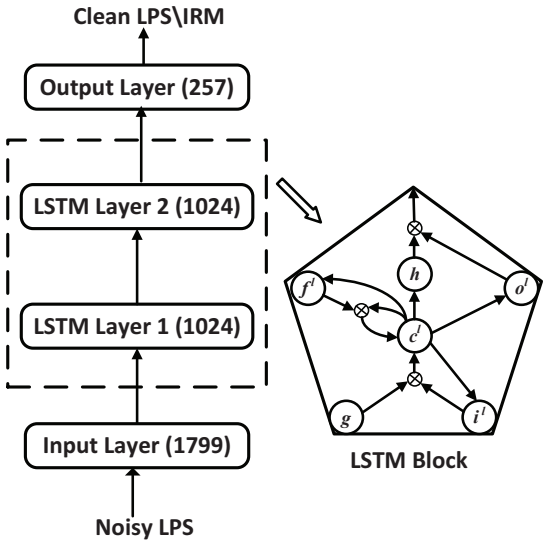

Fig. 2. The architecture of LSTM-IRM and LSTM-DM models.

A. Architecture of LSTM-IRM and LSTM-DM models

The LSTM architecture is shown in Fig. 2. The input layer of the LSTM is a 1799-dimensional vector of noisy LPS features with 7 frame expansion and 257 frequency bins. Each node of output layer adopts a sigmoid activation function. The two hidden layers are 1024 cells for each intend to leverage upon the memory structure that is capable of capturing the temporal constraints. The output layer of the LSTM is a 257-dimensional vector of the clean LPS features for LSTM-DM model or IRM features for LSTM-IRM model. The key components, namely, memory cell state $\boldsymbol{c}^l$, input gate $\boldsymbol{i}^l$, forget gate $\boldsymbol{f}^l$, and output gate $\boldsymbol{o}^l$, are shown in Fig. 2. With this architecture, the network can determine what information to store, update, discard, and output.

To train the LSTM-IRM model, supervised fine-tuning is used to minimize the mean squared error (MSE) between the LSTM-IRM output IRM, denoted as $M_{\text{NN}}(k,l)$ and the reference IRM, denoted as $M_{\text{ref}}(k,l)$, which is defined as

$$E_{\text{NN}} = \sum_{k}\sum_{l}\left[M_{\text{NN}}(k,l) - M_{\text{ref}}(k,l)\right]^2. \quad (2)$$

And to train the LSTM-DM model, supervised fine-tuning is used to minimize the mean squared error (MSE) between the LSTM-DM output LPS features, denoted as $L_{\text{NN}}(k,l)$ and the reference LPS features denoted as $L_{\text{ref}}(k,l)$, which is defined as

$$E_{\text{NN}} = \sum_{k}\sum_{l}\left[L_{\text{NN}}(k,l) - L_{\text{ref}}(k,l)\right]^2. \quad (3)$$

B. LSTM-IRM based IME

Acoustic context information along both the time axis (with multiple neighboring frames) and frequency axis (with full frequency bins) can be fully exploited by the LSTM to obtain a good mask estimate in adverse environments, which is strongly complementary with the conventional CGMM-based approach to retain robustness. 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. And the estimated $M_{\text{NN}}(k,l)$ is combined with $M_{\text{CGMM}}(k,l)$ to yield an improved mask $M_I(k,l)$, i.e.,

$$M_I(k,l) = \sqrt{M_{\text{CGMM}}(k,l)M_{\text{NN}}(k,l)}. \quad (4)$$

C. LSTM-DM based post-processing

In this section, we discuss the LSTM-based post-processing for single channel enhancement. For the beamformer, the aim of the approach is to improve the SNR without destroying the target speech, and it is hard to completely eliminate the noise. While for the LSTM-based regression model, it can eliminate the noise even at low SNR, but the target speech maybe destroyed. We proposed a post-processing based on the LSTM-based regression model. In order to further improve the enhancement performance, the IME-based beamformered speech is processing by the LSTM-based single channel regression model.

Fig. 3 plots the spectrograms of a processed utterance using different enhancement approaches from the test set of simulated data. Figures 3 (a) and (b) present the spectrograms of the original speech from clean speech and channel 5, respectively. Figures 3 (c) and (d) present the spectrograms of the speech processed by the LSTM-DM model and IME-based approach. Although there still exits some noise in Fig. 3 (d), the approach obviously improve the SNRs and the target speech is almost not destroyed. Fig. 3 (e) presents the spectrogram of the IME-based beamformered speech processed by LSTM-based regression model. Comparing to Fig. 3 (c), we can find that the post-processing based on the beamformered speech can almost eliminate the noise, which can further improve the enhancement performance.

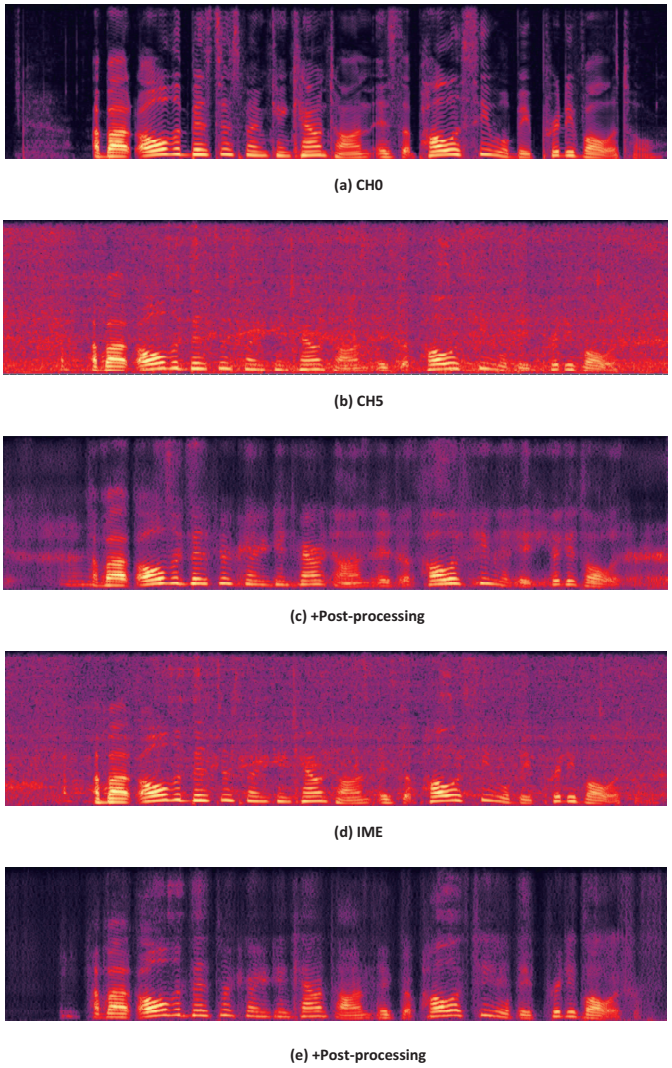

Fig. 3. The spectrograms using different enhancement approaches.

IV. Experimental Evaluation

Now, we present the experimental evaluation of our framework using the CHiME-4 data [18], which was designed to study real-world scenarios where a person is 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). CHiME-4 offers three tasks (1-channel, 2-channel, and 6-channel) with different testing scenarios. In this paper, we focus on the 2-channel and 6-channel cases to make the paper concise. And we construct the simulated test data by mixing clean utterances with environmental noise recordings using the techniques described in [19], and each environment has 40 utterances.

We use the LSTM-based regression model for IRM estimation (LSTM-IRM) and LPS features direct mapping (LSTM-DM). For the LSTM fine-tuning, the learning rate is set to 0.001 for 50 epochs, and the mini-batch size is set to 128. The architecture of LSTM is 1799-1024-1024-257, with 11 frames of LPS features for input layer. And output layer is one frame of clean LPS features or IRM for LSTM-DM or LSTM-IRM, respectively. The training data of the LSTM-IRM model is the channel 5 training data provided by the official (about 18 hours), and the training data of the LSTM-DM model is constructed by mixing clean utterances with environmental noise recordings using the techniques described in [19] (about

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT ENHANCEMENT METHODS FOR THREE TRACKS.

Track	Methods	Measure	BUS	CAF	PED	STR	AVG
1ch	CH5	PESQ	1.82	1.72	2.17	1.74	1.86
		STOI	0.78	0.83	0.79	0.82	0.81
	+PP	PESQ	2.16	2.12	2.46	2.28	2.26
		STOI	0.83	0.82	0.81	0.84	0.83
2ch	IME	PESQ	2.07	1.98	2.43	2.17	2.16
		STOI	0.84	0.85	0.85	0.89	0.85
	+PP	PESQ	2.39	2.25	2.56	2.41	2.40
		STOI	0.86	0.84	0.82	0.83	0.83
6ch	IME	PESQ	2.23	2.12	2.52	2.16	2.25
		STOI	0.86	0.91	0.85	0.88	0.88
	+PP	PESQ	2.48	2.48	2.67	2.52	2.54
		STOI	0.87	0.87	0.81	0.84	0.86

36 hours).

Table I presents the performance comparison of different enhancement methods for 1ch, 2ch, and 6ch tracks. First, for 1ch track, "CH5" and "+PP" denote the original speech from channel 5 and the enhancement speech by the post-processing with the LSTM-DM model. We can find that the original speech from channel 5 processed by LSTM-based regression model can improve the performance, with a relative PESQ improvement of 17.70% across all test sets. Second, for 6ch track, Table I, "IME" denotes the IME-based beamformer. We can observe that although IME-based approach achieves a similar PESQ performance to the single-channel LSTM-DM approach ("+PP" in 1ch), the significant improvement of the STOI performance is yielded (from 0.83 to 0.86 in average). Furthermore, the LSTM-DM based post-processing ("+PP" in 6ch) can achieve an additional PESQ improvement of 11.42% over the IME approach across all test sets, which demonstrates the effectiveness of LSTM-based post-processing. For 2ch track, the similar observation to 6ch track could be made, namely with a relative PESQ improvement of 13.89% from "CH5" to "IME", and with a relative PESQ improvement of 10.00% from "IME" to "+PP". Finally, for both 2ch and 6ch cases, IME with post-processing generates the best PESQ (speech quality) while IME achieves the best STOI (speech intelligibility).

V. ACKNOWLEDGE

This work was supported in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant XDB02070006, and in part by the National Key Research and Development Program of China under Grant 2017YFB1002200.

VI. CONCLUSION

In this paper, we have proposed IME approach and post-processing based on LSTM regression model. Although there still exits some noise in IME-based beamformered speech, the approach obviously improve the SNRs and the target speech is almost not destroyed. And the LSTM-based postprocessing can almost eliminate the noise, which can further improve the enhancement performance.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Y. Tu, J. Du, Y. Xu, L. Dai, and C. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *International Symposium on Chinese Spoken Language Processing.(ISCSLP)*, 2014.

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of timecfrequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[10] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions." in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2008.

[11] 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.

[12] 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.

[13] 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.

[14] R. J. Mcaulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[15] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[16] 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.

[17] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," *Blind Source Separation*, pp. 349–368, 2014.

[18] 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.

[19] ——, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933—1950, 2007.