Teacher-Student Learning and Post-Processing for Robust BiLSTM Mask-Based Acoustic Beamforming

Zhaoyi Liu[1⋆], Qiuyuan Chen[2⋆], Han Hu[3⋆], Haoyu Tang[4⋆], and YX Zou[1]

[1] School of Shenzhen Graduate, Peking University, Shenzhen, China, 518055
1701213615@sz.pku.edu.cn
[2] College of Computer Science and Technology, Zhejiang University, Hangzhou, China
chenqiuyuan@zju.edu.cn
[3] School of Software, Tsinghua University, Beijing, China
hh17@mails.tsinghua.edu.cn
[4] Department of Electronic Systems, Norwegian University of Science & Technology, Trondheim, Norway, 7050
haoyut@stud.ntnu.no

Abstract. In real-world environments, automatic speech recognition (ASR) is highly affected by reverberation and background noise. A well-known strategy to reduce such adverse interferences in multi-microphone scenarios is microphone array acoustic beamforming. Recently, time-frequency (T-F) mask-based acoustic beamforming receives tremendous interest and has shown great benefits as a front-end for noise-robust ASR. However, the conventional neural network (NN) based T-F mask estimation approaches are only trained in parallel simulated speech corpus, which results in poor performance in the real data testing, where a data mismatch problem occurs. To make the NN-based mask estimation, termed as NN-mask, more robust against data mismatch problem, this paper proposes a bi-directional long short-term memory (BiLSTM) based teacher-student (T-S) learning scheme, termed as BiLSTM-TS, which can utilize the real data during student network training stage. Moreover, in order to further suppress the noise in the beamformed signal, we explore three different mask-based post-processing methods to find a better way to utilize the estimated masks from NN. The proposed approach is evaluated as a front-end for ASR on the CHiME-3 dataset. Experimental results show that the data mismatch problem can be reduced significantly by the proposed strategies, leading to relative 4% Word Error Rates (WER) reduction compared to conventional BiLSTM mask-based beamforming, in the real data test set.

Keywords: Teacher-student learning · mask estimation · robust acoustic beamforming · post-processing · speech recognition.

1 Introduction

Automatic speech recognition (ASR) has attracted amounts of attention in recent years with the growing demands for many applications [9,17], such as mobile devices with

⋆ equal contribution

speech-enabled personal assistants and interaction among smart home devices and people by speech [17]. However, for such real-world far-field practical application scenarios, background noise and reverberation degrades speech quality as well as the performance of the ASR system, especially under low signal-to-noise ratio (SNR) conditions.

Multi-channel speech enhancement [4,8], especially NN-mask for acoustic beamforming [2,7,11], significantly improves the performance of ASR under these circumstances. For example, CHiME-3 and CHiME-4 challenges [1], the NN-mask has been developed for beamforming [2,15], which achieves the state-of-art. In [2], a BiLSTM mask network 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 one channel signal. Finally, the 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 idea of those mask-based acoustic beamforming is to estimate a monaural time-frequency (T-F) mask with a well-trained NN in advance, so that the spatial covariance matrices of speech and noise can be derived for beamforming. Therefore, accurately estimating the T-F mask is essential to perform beamforming efficiently. Note that there are two types of NN training T-F mask targets [16]: one is hard mask target, which is a binary mask constructed from premixed speech and noise signals, such as ideal binary mask (IBM); while the second is soft mask target, which contains the probabilistic information among noise signal class and speech signal class, such as ideal ratio mask (IRM). However, the conventional NN-mask only using parallel simulated speech corpus to train shows the poor performance when it predicts masks in the real data testing, where a data mismatch problem occurs [18].

In this paper, in order to reduce the impact of the data mismatch problem of NN-mask, our proposed approach uses bi-directional long short-term memory based teacher-student (BiLSTM-TS) learning [3,10,14] architecture to utilize the real data information in training phase. Specifically, two BiLSTM mask estimation networks, are designed as a teacher network and a student network, respectively. The teacher network is trained with simulated data, and it is then employed to generate the soft labels for both simulated and real data separately. Then, the student network can be trained by the simulated data and real data with generated soft labels from the well-trained teacher network. In addition, in order to further suppress the noise in the beamformed signal, we explore three different mask-based post-processing methods to find a better way to utilize the estimated masks. The proposed approach and mask-based post-processing methods are evaluated on CHiME-3 dataset [1]. Our proposed approach leads to relative 4% average Word Error Rates (WER) reduction compared to conventional BiLSTM mask-based beamforming, in the real data test set.

In summary, our contributions are as follows:

- We propose a BiLSTM-based teacher-student learning scheme for mask estimation (BiLSTM-TS), which enable the NN-based mask estimator to utilize the real training data in the training stage, in order to reduce the impact of the data mismatch problem.
- We explore various mask-based post-processing ways to utilize the estimated masks to further suppress the noise in the beamformed signal.

The remainder of this work is organized as follows: Section 2 shows the related work of mask-based acoustic beamforming. Our approach is presented in Section 3 in detail. Detailed experimental corpus, metric, setups, and results are discussed in Section 4. Finally, Section 5 summarizes the conclusions.

2 Background

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

$$\mathbf{Y}_{\tau,\omega} = \mathbf{X}_{\tau,\omega} + \mathbf{N}_{\tau,\omega} \tag{1}$$

where $\mathbf{Y}_{\tau,\omega}$, $X_{\tau,\omega}$ and $\mathbf{N}_{\tau,\omega}$ represent STFT vectors of the noisy signal, clean speech and noise respectively. τ and ω denote time index and frequency channel, respectively. The beamformer applies a linear filter $\mathbf{w}_{\omega}^{\mathbf{H}}$ to observed noisy signal $\mathbf{Y}_{\tau,\omega}$ to produce an beamformed speech signal, $\tilde{s}_{\tau,\omega}$, as follow:

$$\tilde{s}_{\tau,\omega} = \mathbf{w}_{\omega}^{\mathbf{H}} \mathbf{Y}_{\tau,\omega} \tag{2}$$

where superscript H denotes conjugate transpose.

Fig. 1 shows the diagram of the recently proposed mask-based acoustic beamforming. In the stage of time-frequency mask estimation, multiple microphones receive a set of noisy speech signals, generate a speech mask and a noise mask for each microphone by treating the microphone array as several independent microphones. Then the estimated masks are condensed to a single speech mask and a single noise mask by using a median filter.

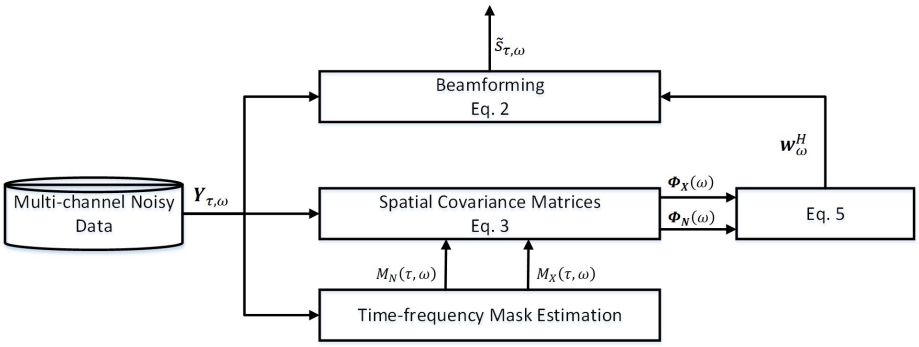

Fig. 1: Processing flow of mask-based beamforming.

With the estimated clean speech mask $M_X(\tau,\omega)$ and noise mask $M_N(\tau,\omega)$ by BiLSTM network, spatial covariance matrices of speech $\Phi_X(\omega)$ and noise $\Phi_N(\omega)$ are computed as:

$$\Phi_v(\omega) = \sum_{\tau=1}^{T} M_v(\tau,\omega)\, \mathbf{Y}_{\tau,\omega} \mathbf{Y}_{\tau,\omega}^{\mathbf{H}} \qquad \mathbf{v} \in \{\mathbf{X}, \mathbf{N}\} \tag{3}$$

Then these spatial covariance matrices compute beamformer coefficients $\mathbf{w}_\omega$. In this study, we propose to maximize the SNR of the beamformer output in each frequency bin separately leading to the Generalized Eigenvalue (GEV) beamformer [2] with coefficients:

$$\mathbf{w_{GEV}}\left(\omega\right) = \mathbf{argmax}_{\mathbf{w}} \frac{\mathbf{w^H \Phi_X}\left(\omega\right)\mathbf{w}}{\mathbf{w^H \Phi_N}\left(\omega\right)\mathbf{w}} \tag{4}$$

This optimization problem is equivalent to solving the following eigenvalue problem:

$$\left\{\Phi_N^{-1}\Phi_X\right\}\mathbf{w_{GEV}}\left(\omega\right) = \lambda \mathbf{w_{GEV}}\left(\omega\right) \tag{5}$$

where $\mathbf{w_{GEV}}\left(\omega\right)$ is the eigenvector of $\left\{\Phi_N^{-1}\Phi_X\right\}$ and λ is the corresponding eigenvalue. Finlly, the blind analysis normalization (BAN) [2] is used as a post-filter of beamformer to reduce arbitrary distortion of the GEV beamformer.

3 Approach

3.1 BiLSTM teacher-student learning for mask estimation

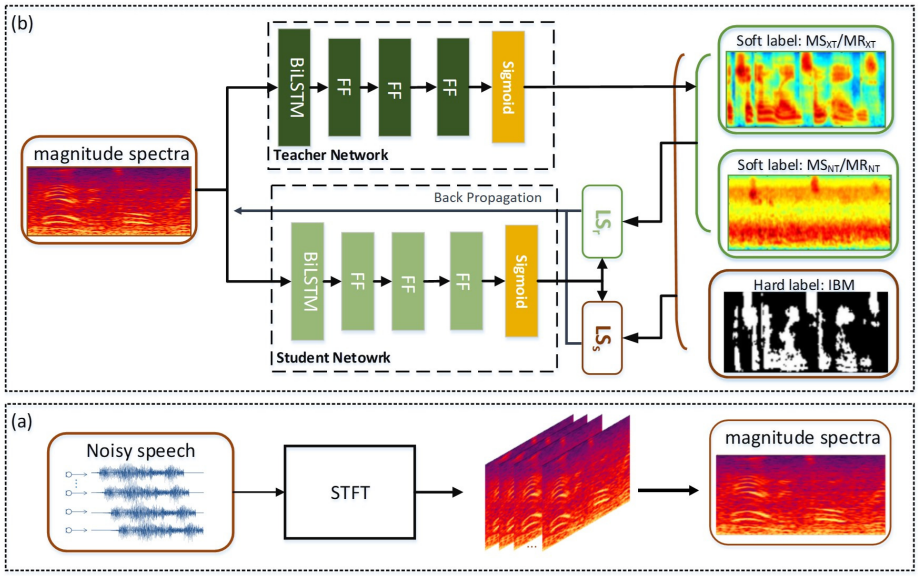

Fig. 2: The framework of our proposed BiLSTM teacher-student learning mask estimation (BiLSTM-TS). (a) **Feature extraction**. Obtain the short-time Fourier transforms (STFT) of the noisy signals and calculate their magnitude spectra $|\mathbf{Y}|_{\tau,\omega}$. (b) **Proposed BiLSTM-TS model**. Use magnitude spectrum of ith channel $Y_i^m(\tau,\omega)$ as the input of NN. The well-trained BiLSTM teacher network generates the estimated clean speech soft labels MS_{XT} and MR_{XT} as well as noise soft label MS_{NT} and MR_{NT} as additional labels to student network.

This work focuses on the data mismatch problem in NN-mask. Considering that the NN-mask is a supervised training which requires target labels in training stage. Hence, parallel speech corpus, such as original clean speech and simulated noisy speech, is required to prepare the corresponding labels. This means the conventional NN-mask can only be trained with simulated data, which may lead to a poor performance of the mask estimation network under the real data conditions where a data mismatch problem occurs. In order to reduce the impact of the data mismatch problem, our idea is quite intuitive that the real data can be pooled with the simulated data in the NN-mask training stage to train a better mask estimation network. In this work, we introduce teacher-student (T-S) scheme to reduce data mismatch problem. The training strategy is that the well-trained teacher network is used as the label generator which processes the original simulated and real data in order to predict soft labels. Then the student network can utilize real-data information and simulated data information to train a mask estimator. Fig. 2 illustrates the framework of proposed BiLSTM teacher-student learning (BiLSTM-TS).

Teacher network For teacher network, in the training stage, the magnitude spectrum of noisy signal in STFT-domain is given as the input of the teacher network. Note that the teacher network is only trained by using simulated training data. We employ the ideal binary mask (IBM) as the training target. There are two types of IBM are estimated: one is the clean speech mask $IBM_X(\tau,\omega) \in 0,1$, the other is the noise mask $IBM_N(\tau,\omega) \in 0,1$, which are defined as :

$$\mathbf{IBM_N} = \begin{cases} 1, & \frac{||X(\tau,\omega)||}{||N(\tau,\omega)||} < 10^{th_{\mathbf{N}}}, \\ 0, & else. \end{cases} \tag{6}$$

$$\mathbf{IBM_X} = \begin{cases} 1, & \frac{||X(\tau,\omega)||}{||N(\tau,\omega)||} > 10^{th_{\mathbf{X}}}, \\ 0, & else. \end{cases} \tag{7}$$

where $||X(\tau,\omega)|| \in \mathbb{R}_{\geqslant 0}$ and $||N(\tau,\omega)|| \in \mathbb{R}_{\geqslant 0}$ are power spectra of the clean speech signal and the noise signal at each T-F unit (τ,ω), respectively. To obtain the better reults, the two theresholds $th_{\mathbf{X}}$ and $th_{\mathbf{N}}$ are manually selected to be different from each other.

The teacher network is trained to predict the clean speech mask $MS_{XT}(\tau,\omega) \in [0,1]$ and $MR_{XT}(\tau,\omega) \in [0,1]$ as well as the noise mask $MS_{NT}(\tau,\omega) \in [0,1]$ and $MR_{NT}(\tau,\omega) \in [0,1]$ at each T-F bin (τ,ω). We use the NN proposed in [2] as the architecture of our teacher network, including a BiLSTM layer followed with three-feed forward layers. Table 1 shows the configurations of teacher network.

We use the binary cross-entropy (BCE) as the loss function of teacher network which is defined as:

$$\begin{aligned} Loss =& BCE(IBM_v, M_{vT}) \\ \stackrel{def}{=}& \frac{1}{T}\frac{1}{W} \sum_{v\in\{X,N\}} \sum_{\tau=1}^{T} \sum_{\omega=1}^{W} IBM_v(\tau,\omega)\, log(M_{vT}(\tau,\omega)) \\ &+ (1 - IBM_v(\tau,\omega)) log(1 - M_{vT}(\tau,\omega)) \end{aligned} \tag{8}$$

Table 1: Configurations of BiLSTM Teacher mask network.

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

As shown in Fig. 2, in our design, the well-trained teacher network is used as the soft label generator for the real data and the simulated data. Then the student network can utilize the generated masks of clean speech $M_{XT}(\tau,\omega)$ and noise $M_{NT}(\tau,\omega)$ as the soft labels.

Student network The structure of the student network is the same as the teacher network described in Section 3.1. In the training stage of student network, we use different loss functions to train our SMM with the simulated data and the real-recording data, respectively. For simulated data, we consider the following lost functions for speech LS_{sX} and noise LS_{sN}, as follow:

$$\begin{aligned} LS_{sX} =&(1-\pi)BCE(IBM_X(\tau,\omega), MS_{XS}(\tau,\omega)) \\ &+ \pi BCE(MS_{XT}(\tau,\omega), MS_{XS}(\tau,\omega)) \end{aligned} \tag{9}$$

$$\begin{aligned} LS_{sN} =&(1-\pi)BCE(IBM_N(\tau,\omega), MS_{NS}(\tau,\omega)) \\ &+ \pi BCE(MX_{NT}(\tau,\omega), MS_{NS}(\tau,\omega)) \end{aligned} \tag{10}$$

where $MS_{XS}(\tau,\omega)$ and $MS_{NS}(\tau,\omega)$ denotes the estimated clean speech mask and noise mask by student network, respectively. The hyper-parameter π is the imitation parameter adjusting the relative attention of two type of targets. The $IBM_X(\tau,\omega)$ and $IBM_N(\tau,\omega)$ are the hard mask labels of speech and noise, respectively.The final cost function of SMM for simulated data, termed as L_{Ss} is expressed as:

$$LS_s = (LS_{sX} + LS_{sN})/2 \tag{11}$$

For real data, the conventional NN-mask can only utilize parallel data combined by the clean speech and noise, which are not usually obtained in the practical application. However, the student network is able to obtain the soft labels of the real data generated by teacher network. Therefore, the loss function for the student network for real data, termed as LS_r, is defined as:

$$\begin{aligned} LS_r =&[BCE(MR_{XT}(\tau,\omega), MR_{XS}(\tau,\omega)) \\ &+ BCE(MR_{NT}(\tau,\omega), MR_{NS}(\tau,\omega))]/2 \end{aligned} \tag{12}$$

where $MR_{XS}(\tau,\omega)$ and $MR_{NS}(\tau,\omega)$ represent the estimated noise mask and clean speech mask by student network, respectively.

With this setup, the student network has been trained on the simulated data and real-recording data with loss LS_s and loss LS_r, respectively. When the student predicts the clean speech mask and noise mask for each microphone channel, we calculate the beamforming coefficients using the method shown in section 2.

3.2 Mask-based Post-processing

For the beamformer, the aim of the method is to improve the signal-to-noise ratio (SNR) without distorted the clean speech, but it is hard to completely eliminate the noise. There are many post-processing approaches can be utilized to eliminate the noise in the beamformed speech, and obtain the extra enhanced signals. However, the enhanced speech need to avoid being distorted to further improve the ASR performance. We explore three different mask-based post-processing for the beamformed speech:

1. Apply the estimated clean speech mask $M_X(\tau,\omega)$ directly (**direct-mask**), after beamforming as post-processing. After this post-processing, we can obtain enhanced speech $\tilde{x}_{\tau,\omega}$ as follow:

$$\tilde{x}_{\tau,\omega} = \tilde{s}_{\tau,\omega} \odot M_X(\tau,\omega) \tag{13}$$

   where $\odot$ presents dot multiplication. And the M_X is estimated by the mask estimation network.
2. In order to simultaneously control the noise reduction level and speech distortion, the beamformed speech $\tilde{s}_{\tau,\omega}$ can be conditionally used $M_X(\tau,\omega)$ by piecewise function (**condition-mask**) as follows:

$$\tilde{x}_{\tau,\omega} = \begin{cases} \tilde{s}_{\tau,\omega} & M_X(\tau,\omega) \geqslant 0.8 \\ \tilde{s}_{\tau,\omega} \odot M_X(\tau,\omega) & 0.2 \leqslant M_X(\tau,\omega) < 0.8 \\ \tilde{s}_{\tau,\omega} \odot 0.2 & otherwise \end{cases} \tag{14}$$

   Note that the value of $M_X(\tau,\omega)$ are real numbers within the range [0,1]. If the value of estimated $M_X(\tau,\omega)$ is very large indicating that it has very high SNR at certain T-F unit, it is not necessary to perform noise reduction which can potentially result in the speech distortion.
3. Apply the **threshold-mask** post-processing method. Firstly, we compute the global SNR for each frequency ω, termed as $gSNR(\omega)$, as follow:

$$gSNR(\omega) = 10log_{10}\frac{\sum_{\tau=1}^{T} M_X(\tau,\omega)\tilde{s}^2_{\tau,\omega}}{\sum_{\tau=1}^{T} M_N(\tau,\omega)\tilde{s}^2_{\tau,\omega}} \tag{15}$$

   Secondly, we use $gSNR(\omega)$ to calculate a threshold, $th(\omega)$, as:

$$th(\omega) = \frac{1}{1+e^{(\alpha gSNR-\beta)/\gamma}} \tag{16}$$

   The Eq. (16) is a sigmoidal function, hence the threshold is ranged [0, 1]. We use parameters α, β and γ to adjust the shape of the sigmoidal function. Through cross-validation, their values are set to 1.5, -5 and 2, respectively. Then, the threshold-mask can be obtained, termed as $M_{th}(\tau,\omega)$, as:

$$M_{th}(\tau,\omega) = M_X(\tau,\omega)^{th(\omega)} \tag{17}$$

From Eq. (16) and (17), we can find that when $gSNR(\omega)$ is high the value of $th(\omega)$ will be close 0. This makes the threshold-mask $M_{th}(\tau,\omega)$ be close to 1, which is independent of the value of clean speech mask $M_X(\tau,\omega)$. If not, $M_{th}(\tau,\omega)$ is close to $M_X(\tau,\omega)$ when $gSNR(\omega)$ is low. Finally, we can obtain the enhanced speech $\tilde{x}_{\tau,\omega}$ by:

$$\tilde{x}_{\tau,\omega} = \tilde{s}_{\tau,\omega} \odot M_{th}(\tau,\omega) \tag{18}$$

4 Experiments

In this work, we evaluate the proposed acoustic beamforming approach on ASR tasks using the CHiME-3 corpus [1]. The proposed algorithm is used as a frontend for ASR systems.

4.1 Corpus

CHiME-3 The CHiME-3 corpus includes real and simulated data generated by artificially mixing the incorporations of Wall Street Jounral (WSJ) corpus [5] sentences spoken with 4 different noisy environments which selected: cafe (CAFE), 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 subset:

- Training set: composing 8738 (1600 real + 7138 simulated) noisy utterances.
- Development set (dt_05): containing 3280 (1640 real + 1640 simulated) noisy utterances.
- Evaluation set (et_05): including 2640 (1320 real + 1320 simulated) noisy utterances.

4.2 Metric

Word Error Rate (WER) WER is a common metric to evaluate the performance of ASR system [6]. The WER compares a reference to an hypothesis and is defined as:

$$WER = \frac{S + D + I}{N} \tag{19}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and N is the number of total words in the reference. The lower the value of WER, the better the ASR performance.

4.3 Experimental Setups

To compare the performance of different masking models, the standard back-end ASR provided by the CHiME-3 challenge is directly used, which contains based on a relatively simple Gaussian Mixture Model (GMM) acoustic model [12] trained using Kaldi speech recognition toolkit [13]. For language model, A standard the Wall Street Journal (WSJ) speaker-independent medium-vocabulary (5K) word tri-gram language model is used for decoding in this work. We use a common metric word error rate (WER) to denote the performance of ASR.

4.4 Evaluation on BiLSTM-TS

As frontend processing, the mask-based beamforming approach proposed by Heymann *et al.* [2] which is described in Section 2 as well as set as our teacher network, is used for comparison with our proposed student models with different values of hyper-parameter π.

Table 2: Comparison of the performance (%WER) of different mask estimation networks for ASR systems on CHiME-3.

Parameters	BiLSTM Mask	DEV		EVAL	
		simu	*real*	*simu*	*real*
–	Baseline/Teacher	10.8	11.97	11.59	17.97
$\pi = 0.0$	Student	10.87	11.84	11.62	17.34
$\pi = 0.2$	Student	10.61	11.67	11.84	16.89
$\pi = 0.5$	Student	10.2	11.26	11.79	15.53
$\pi = 0.8$	Student	10.4	10.78	9.96	14.75
$\pi = 1.0$	Student	**8.4**	**9.37**	**8.89**	**13.79**

The results of these experiments are shown in Table 2. From the results, we can see that the performance of most student models with different configurations are better than that of the teacher network as expected, although the teacher model has already been robust. The results also reveal that except for the student model with hyper-parameter $\pi = 0.0$, the improvements are largely achieved not only on the real test condition but also on the simulated test condition. This is an interesting finding, since the data mismatch problem between the original simulation training and the test conditions is small, only adding the real data in the training actually increases the mismatch for the simulated test conditions. Specifically, adding the teacher-student (T-S) learning scheme results in a relative improvement rate of up to 4.1% and 2.7% for the real and simulated evaluation data, respectively. In contrast, utilizing T-S learning scheme can reduce the impact of the data mismatch problem of mask estimation for acoustic beamforming by pooling real data with simultaed data in the training stage with soft labels from teacher model, thus contributes to better performance for real applications.

4.5 Evaluation on different post-processing methods

We also compared the three different mask-based post-processing methods described in Section 4 on two BiLSTM mask estimation networks for beamforming by using same ASR back-end. In detail, the two BiLSTM mask models are the teacher model and the student model with hyper-parameter $\pi = 0.0$. And we use the BiLSTM teacher model and stdent model without the post-processing as the baselines.

Table 3: Comparison of the performance (%WER) of the BiLSTM teacher and student mask estimation network with direct-mask, condition mask, and threshold-mask methods as well as without post-processing for ASR systems on CHiME3.

BiLSTM MASK	Post-processing	DEV		EVAL	
		simu	*real*	*simu*	*real*
Baseline	None	10.8	11.97	11.59	17.97
Student	None	**8.4**	9.37	**8.89**	13.79
Baseline	Direct-mask	12.37	13.12	13.01	19.85
Student	Direct-mask	11.81	12.58	10.36	16.27
Baseline	Condition-mask	10.27	11.43	11.63	17.86
Student	Condition-mask	8.22	9.35	9.09	13.52
Baseline	Threshold-mask	10.97	12.08	11.27	17.42
Student	Threshold-mask	8.42	**8.96**	9.1	**13.46**

Table 3 lists the ASR performance of direct-mask, condition mask and threshold-mask approaches. First, for the direct-mask method, we can find that since directly applying the mask to the beamformed signal is very sensitive to the mask estimation error. And the performance of direct-mask method underperforms that of the baselines. Second, for the condition-mask method, the results of this method show the improvments on the real data, while the performances of condition-mask post-processing on the simulated data are slightly decreased. From the results of the threshod-mask method, we can find that threshod-mask post-processing for beamformed speech can suppress the noise, which can further improve the ASR performance for both real-recording data and simulated data.

5 Conclusion

In this work, motivated by the data mismatch problem for NN-based mask estimation acoustic beamforming results from training simulated data and testing real data, we propose BiLSTM teacher-student learning (BiLSTM-TS) approach. With the aim of utilizing the real record data in mask estimation training, the T-S is applied on the real record data to produce the soft labels, hence the real data can be combined with the simulated data for mask estimation. Experimental results show that, as a frontend, the student model of BiLSTM-TS improves ASR performance. Furthermore, through exploring the different mask-based post-processing methods, we find that the threshold-mask can further suppress the noise in the beamformed signal. For the future work, by using strong ASR back-end, we believe that the ASR performance can be further improved.

References

1. Barker, J., Marxer, R., Vincent, E., Watanabe, S.: The third 'chime' speech separation and recognition challenge: Analysis and outcomes. Computer Speech Language **46**, 605–626 (2017)
2. Heymann, J., Drude, L., Haebumbach, R.: Neural network based spectral mask estimation for acoustic beamforming. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) pp. 196–200 (2016)
3. Hinton, G.E., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network. arXiv: Machine Learning (2015)
4. Hoshen, Y., Weiss, R.J., Wilson, K.W.: Speech acoustic modeling from raw multichannel waveforms. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
5. John Garofalo, David Graff, D.P., Pallett, D.: "csr-i (wsj0) complete". Linguistic Data Consortium, Philadelphia (2007)
6. Klakow, D., Peters, J.: Testing the correlation of word error rate and perplexity. Speech Communication **38**(1), 19–28 (2002)
7. Kubo, Y., Nakatani, T., Delcroix, M., Kinoshita, K., Araki, S.: Mask-based MVDR Beamformer for Noisy Multisource Environments: Introduction of Time-varying Spatial Covariance Model. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
8. Kumatani, K., Mcdonough, J.W., Raj, B.: Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors. IEEE Signal Processing Magazine **29**(6), 127–140 (2012)
9. Mofrad, M.H., Mosse, D.: Speech recognition and voice separation for the internet of things. Proceedings of the 8th International Conference on the Internet of Things p. 8 (2018)
10. Mosner, L., Wu, M., Raju, A., Parthasarathi, S.H.K., Kumatani, K., Sundaram, S., Maas, R., Hoffmeister, B.: 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)
11. Pfeifenberger, L., Zohrer, M., Pernkopf, F.: Dnn-based speech mask estimation for eigenvector beamforming. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
12. Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Rastrow, A., et al.: The subspace Gaussian mixture model-A structured model for speech recognition. Computer Speech Language **25**(2), 404–439 (2011)
13. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The Kaldi Speech Recognition Toolkit. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (2011)
14. Subramanian, A.S., Chen, S.J., Watanabe, S.: Student-teacher learning for BLSTM mask-based speech enhancement. Interspeech pp. 3249–3253 (2018)
15. Tu, Y., Du, J., Sun, L., Ma, F., Lee, C.: On Design of Robust Deep Models for CHiME-4 Multi-Channel Speech Recognition with Multiple Configurations of Array Microphones. INTERSPEECH pp. 394–398 (2017)
16. Wang, D., Chen, J.: Supervised Speech Separation Based on Deep Learning: An Overview. IEEE Transactions on Audio, Speech, and Language Processing **26**(10), 1702–1726 (2018)
17. Yu, D., Li, J.: Recent progresses in deep learning based acoustic models. IEEE/CAA Journal of Automatica Sinica **4**(3), 396–409 (2017)
18. Zhou, Y., Qian, Y.: 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) pp. 536–540 (2018)