IKDMM: Iterative Knowledge Distillation Mask Model for Robust Acoustic Beamforming

Zhaoyi Liu

1701213615@sz.pku.edu.cn
Peking University Shenzhen Graduate School, Shenzhen, China

Yuexian Zou

zouyx@pku.edu.cn
Peking University Shenzhen Graduate School, Shenzhen, China

Peng Cheng Laboratory, Shenzhen, China

ABSTRACT

Microphone array beamforming has been approved to be an effective method for suppressing adverse interferences. Recently, acoustic beamformers that employ neural networks (NN) for estimating the time-frequency (T-F) mask, termed as TFMask-BF, receive tremendous attention and have shown great benefits as a front-end for noise-robust Automatic Speech Recognition (ASR). However, our preliminary experiments using TFMask-BF for ASR task show that the mask model trained with simulated data cannot perform well in the real environment since there is a data mismatch problem. In this study, we adopt the knowledge distillation learning framework to make use of real-recording data together with simulated data in the training phase to reduce the impact of the data mismatch. Moreover, a novel iterative knowledge distillation mask model (IKDMM) training scheme has been systematically developed. Specifically, two bi-directional long short-term memory (BLSTM) models, are designed as a teacher mask model (TMM) and a student mask model (SMM). The TMM is trained with simulated data at each iteration and then it is employed to separately generate the soft mask labels of both simulated and real-recording data. The simulated data and the real-recording data with their corresponding generated soft mask labels are formed into the new training data to train our SMM at each iteration. The proposed approach is evaluated as a front-end for ASR on the six-channel CHiME-4 corpus. Experimental results show that the data mismatch problem can be reduced by our IKDMM, leading to a 5% relative Word Error Rate (WER) reduction compared to conventional TFMask-BF for the real-recording data under noisy conditions.

CCS CONCEPTS

• Computing methodologies → Speech recognition.

KEYWORDS

Iterative knowledge distillation, mask model, acoustic beamforming, speech recognition

1 INTRODUCTION

For far-field speech applications, background noise and reverberation degrades speech quality as well as the performance of the Automatic Speech Recognition (ASR) system, especially under low signal-to-noise ratio (SNR) conditions. To solve the speech quality degradation problem, microphone array acoustic beamforming techniques have shown to be beneficial for robust ASR [1–3]. Delay-and-sum beamformers, minimum variance distortionless response (MVDR) beamformers, and generalized Eigenvalue (GEV) beamformers are often employed to reduce noise [4–6] and have been shown to improve the ASR performance in distant speech recognition [7].

Recently, researchers have proposed acoustic beamforming algorithms based on deep learning and T-F masking [3, 8], termed as TFMask-BF. The central idea of TFMask-BF is to learn an NN-based mask estimator, termed as the mask model [9]. The spatial covariance matrices of target speech and noise can then be estimated for constructing beamformers more accurately. A prominent feature of TFMask-BF is that it can suppress the adverse noise effectively without relying on prior knowledge such as the microphone array geometry or the acoustic conditions of the room (e.g., a plane wave condition). This is particularly advantageous for many practical applications of ASR tasks. For example, in CHiME-3 and CHiME-4 challenges [10, 11], the mask model has been developed for beamforming [7, 12], which achieves state-of-the-art performance. In the winning solution of CHiME-3 [13], a BLSTM mask model has been designed and trained. This approach treats the multi-channel signals separately where one speech mask and one noise mask are learned for each channel’s signal. These masks are then combined into the final masks using median pooling. The covariance matrices of speech and noise, which are computed by the estimated speech and noise masks, can be used to derive the beamforming weights.

A robust mask model is important to perform beamforming efficiently. However, the literature shows that the mask model inevitably degrades when there is a data mismatch between the training and testing conditions. The data mismatch problem appears because the mask model is trained with only the simulated speech corpus while it is applied for the real-recording data [11, 14]. This work aims to reduce the adverse effect of the data mismatch on mask model learning. In this study, we adopt the knowledge distillation (KD) learning framework [15–18] to train a better mask model by using extra real-recording data in the training phase. Specifically, two BLSTM models are respectively designed as a teacher mask model (TMM) and a student mask model (SMM). The TMM is trained with simulated training data. Then, it is fixed and used to separately generate the soft mask labels of both simulated and real-recording data. Finally, the simulated data and the real-recording data with the corresponding generated soft mask labels form the new training data to train the SMM. For presentation clarity, this resulting approach is conventional knowledge distillation mask model training scheme (CDKMM). Moreover, motivated by the cognitive process of human beings where knowledge and experience are learned gradually [18, 19], we propose a novel iterative knowledge distillation mask model (IKDMM) training scheme. In
Figure 1: The block diagram of the baseline beamformer (TFMask-BF [13])

This approach, a TMM is trained with simulated data and predicts soft mask labels of simulated and real-recording data at each iteration. Then, the SMM is trained by using both simulated data and real-recording data with their corresponding iterative soft mask labels. The well-trained SMM is then used to produce enhanced speech and mask out interference (such as background noise) which helps to derive a robust beamformer. We validate the effectiveness of our proposed IKDMM on the six-channel CHiME-4 corpus. The evaluation results show that our IKDMM outperforms the conventional TFMask-BF and CKDMM in the real-recording test set.

The remainder of this work is organized as follows. In section 2, we present the background of TFMask-BF. Section 3 describes the proposed approach in details. Detailed experimental corpus, setups, and results are shown in Section 4. Finally, the conclusion is in Section 5.

2 BACKGROUND

For presentation completeness, Fig. 1 shows the block diagram of the mask-based GEV beamformer method [13] used as the baseline in this paper, which is termed as TFMask-BF. It receives a set of noisy speech signals captured by a microphone array and generates a single enhanced speech signal. The TFMask-BF consists of two main blocks: GEV beamformer based on mask and mask model. The details are given in Section 2.1 and Section 2.2, respectively.

2.1 GEV Beamformer based on Mask

The model of microphone array in the short-time Fourier transform (STFT) domain is formulated as

\[ Y(\tau, \omega) = X(\tau, \omega) + N(\tau, \omega) \]  

where \( Y(\tau, \omega) \) is the STFT vector of the observed noisy speech, \( X(\tau, \omega) \) and \( N(\tau, \omega) \) represent the STFT vector of the received speech and noise at a specific T-F unit, respectively. \( \omega \) and \( \tau \) denote frequency bin and time frame index. In this work, we assume that there is only one target source and its position is fixed within each utterance.

The objective of the GEV beamformer is to maximize the SNR of the beamformer output in each frequency bin separately [5] with beamforming coefficients:

\[ w_{GEV}(\omega) = \text{argmax}_w \frac{w^H \Phi_{XX}(\omega) w}{w^H \Phi_{NN}(\omega) w} \]  

where \( \Phi_{XX}(\omega) \) and \( \Phi_{NN}(\omega) \) are the spatial covariance matrices of speech and noise respectively. Eq. (2) is known as the Rayleigh coefficient. The optimization problem of Eq. (2) leads to the well-known generalized eigenvalue problem:

\[ \Phi_{XX}(\omega) w_{GEV}(\omega) = \lambda \Phi_{NN}(\omega) w_{GEV}(\omega) \]  

where \( \lambda \) is the corresponding eigenvalue and \( w_{GEV}(\omega) \) is the eigenvector of \( \Phi_{NN}^{-1}(\omega) \Phi_{XX}(\omega) \).

Once we obtain the beamforming coefficients \( w_{GEV}(\omega) \), we can calculate the output of the GEV beamformer, \( \hat{s}(\tau, \omega) \), as follow:

\[ \hat{s}(\tau, \omega) = w_{GEV}(\omega) Y(\tau, \omega) \]  

where superscript \( H \) denotes conjugate transposition.

As we can see, the key to successful noise reduction with the GEV beamformer is the accurate estimation of the speech covariance matrix \( \Phi_{XX}(\omega) \) and the noise covariance matrix \( \Phi_{NN}(\omega) \). Let \( M_X(\tau, \omega) \) and \( M_N(\tau, \omega) \) be the estimated masks for speech and noise at each T-F bin \( (\tau, \omega) \), respectively. Following [13], the respective spatial covariance matrices of speech and noise can be estimated as:

\[ \Phi_{XX}(\omega) = \sum_{\tau=1}^{T} M_X(\tau, \omega) Y(\tau, \omega) Y(\tau, \omega)^H \]  

\[ \Phi_{NN}(\omega) = \sum_{\tau=1}^{T} M_N(\tau, \omega) Y(\tau, \omega) Y(\tau, \omega)^H \]  

where \( T \) is the total time indexes.

With the above method, the accuracy of the acoustic beamforming depends mainly on that of the masks estimated by the mask model.

2.2 Mask model

The mask model is trained in a fully supervised manner to estimate two masks: the speech mask and the noise mask. The magnitude spectrum of the noisy speech is taken as the input of the mask model. The ideal binary speech mask \( IBM_X \in \{0, 1\} \) and the ideal binary noise mask \( IBM_N \in \{0, 1\} \) are defined by:

\[ IBM_X = \begin{cases} 1, & ||X(\tau, \omega)|| < 10^{tb_X(\omega)}, \\ 0, & \text{else} \end{cases} \]  

\[ IBM_N = \begin{cases} 1, & ||N(\tau, \omega)|| > 10^{tb_N(\omega)}, \\ 0, & \text{else} \end{cases} \]  

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

Table 1 shows the configurations of the BLSTM mask model as our baseline method [13].

In the training phase, the BLSTM mask model utilizes the simulated training data, and it is trained by using the binary cross-entropy (BCE) loss function. Let’s define the total time indexes as \( T \) and the total number of frequency bins as \( W \). Then the BCE loss
IKDMM: Iterative Knowledge Distillation Mask Model for Robust Acoustic Beamforming

Figure 2: The framework of our proposed iterative knowledge distillation mask model (IKDMM) training scheme. At each training iteration: (a) Feature extraction: Obtain the short-time Fourier transforms (STFT) of the noisy signals and calculate their magnitude spectra $|Y(\tau, \omega)|$. Use the magnitude spectrum of $i$th channel $|Y_i(\tau, \omega)|$ as the input of the mask model. (b) Teacher mask model (TMM): A TMM is trained with simulated data and their corresponding hard mask label IBM. (c1) For simulated data, the TMM generates soft mask labels for speech $TMS_X$ and noise $TMS_N$ as additional labels to student mask model, respectively. (c2) Similarly, soft mask labels of speech $TMR_X$ and noise $TMR_N$ are produced for real-recording data by TMM. (d) Student mask model (SMM): The SMM is trained with both simulated data and real-recording data.

Table 1: Configurations of BLSTM Mask Model

<table>
<thead>
<tr>
<th>Layer</th>
<th>Units</th>
<th>Type</th>
<th>Activation</th>
<th>Dropout</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td>256</td>
<td>BLSTM</td>
<td>Tanh</td>
<td>0.5</td>
</tr>
<tr>
<td>L2</td>
<td>513</td>
<td>Feedforward 1</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>L3</td>
<td>513</td>
<td>Feedforward 2</td>
<td>ReLU</td>
<td>0.5</td>
</tr>
<tr>
<td>L4</td>
<td>1026</td>
<td>Feedforward 3</td>
<td>Sigmoid</td>
<td>0.0</td>
</tr>
</tbody>
</table>

The function is given by [14]:

$$
\text{Loss} = \text{BCE}(IBM_L, M_U)
$$

$$
def_{L} \frac{1}{W} \sum_{W=1, \tau=1}^{W} \log(M_U(\tau, \omega) - M_U(\tau, \omega)) + (1 - M_U(\tau, \omega)) \log(1 - M_U(\tau, \omega))
$$

where $IBM_N$ and $IBM_X$ are given in (7) and (8), respectively. $M_X(\tau, \omega)$ and $M_N(\tau, \omega)$ are the estimated masks of speech and noise, respectively.

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

3 Iterative Knowledge Distillation Mask Model (IKDMM)

Our preliminary experiments using TFMask-BF for ASR task show that the mask model trained with only simulated training data results in poor performance when the real-recording data is used in the prediction stage, where a data mismatch problem occurs. Motivated by the promising results brought by BLSTM mask model, this work focuses on improving the generalization capability of mask model by introducing the real-recording training data in the training phase to alleviate the data mismatch problem between the training data and the testing data. Therefore, we adopt the knowledge distillation learning framework, where the TMM can process both simulated and real-recording training data to generate the mask labels to the SMM. Furthermore, in this study, we propose a novel iterative knowledge distillation mask model (IKDMM) training scheme. Though it is possible to utilize the real-recording training data by using the conventional knowledge distillation mask model (CKDMM) training scheme, we find our IKDMM approach provides superior performance, as shown in our results. Fig. 2 shows an overview of the proposed framework. The details of IKDMM are given in the following context.
Table 2: WERs (%) obtained using different mask models for the six-channel CHiME-4 development set (dt_05) and evaluation set (et_05). Bold fonts indicate the best score for each condition.

<table>
<thead>
<tr>
<th>Mask Model</th>
<th>Parameters</th>
<th>Set</th>
<th>Simulated Data</th>
<th>Real-recording Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>dt_05</td>
<td>Ave BUS CAF PED STR</td>
<td>Ave BUS CAF PED STR</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.0</td>
<td></td>
<td>6.75 6.14 8.1 5.96 6.8</td>
<td>7.91 10.09 7.61 5.49 8.44</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.0</td>
<td></td>
<td>6.43 5.37 7.92 5.78 6.64</td>
<td>7.7 10.06 7.26 5.18 8.29</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.2</td>
<td></td>
<td>7.14 9.26 6.8 7.67</td>
<td>9.14 11.95 8.64 6.45 9.53</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.0</td>
<td></td>
<td>9.14 11.95 8.64 6.45 9.53</td>
<td>9.02 12.18 8.42 6.08 9.41</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.5</td>
<td>dt_05</td>
<td>7.63 6.37 9.53 6.68 7.94</td>
<td>9.21 12.26 8.73 6.17 9.7</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.5</td>
<td>dt_05</td>
<td>7.18 6.73 9.71 6.93 7.89</td>
<td>9.12 12.18 8.75 6.24 9.31</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.8</td>
<td></td>
<td>7.78 6.67 9.57 6.76 8.11</td>
<td>9.2 12.82 8.6 5.96 9.44</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.8</td>
<td></td>
<td>7.75 6.5 9.69 6.76 8.05</td>
<td>9.18 12.11 8.69 6.59 9.32</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 1.0</td>
<td></td>
<td>7.66 6.17 9.79 6.77 7.91</td>
<td>8.89 11.74 8.45 5.94 9.44</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 1.0</td>
<td></td>
<td>7.72 6.3 9.93 6.64 8.04</td>
<td>8.86 11.71 7.82 6.4 9.51</td>
</tr>
<tr>
<td>cBLSTM</td>
<td>−</td>
<td>et_05</td>
<td>11.16 11.34 11.09 11.43 10.76</td>
<td>17.53 32.32 14.62 12.87 10.33</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.0</td>
<td>et_05</td>
<td>8.37 8.14 8.65 8.39 8.29</td>
<td>13.39 23.39 12.4 9.17 8.59</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.0</td>
<td>et_05</td>
<td>8.24 7.7 8.07 8.11 8.44</td>
<td>12.55 21.63 11.77 8.67 8.14</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.2</td>
<td>et_05</td>
<td>10.43 10.68 10.24 10.74 10.07</td>
<td>17.28 30.66 15.52 12.5 10.44</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.2</td>
<td>et_05</td>
<td>10.53 11.02 10.2 10.61 10.29</td>
<td>16.88 29.78 14.23 12.76 10.74</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.5</td>
<td>et_05</td>
<td>10.91 11.3 11.02 10.65 10.66</td>
<td>17.17 30.53 14.92 12.65 10.59</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.5</td>
<td>et_05</td>
<td>9.77 9.19 10.16 10.01 9.71</td>
<td>17.06 31.07 14.42 12.22 10.55</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 0.8</td>
<td>et_05</td>
<td>10.92 11.43 11.08 10.72 10.46</td>
<td>17.4 30.79 14.42 12.65 11.75</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 0.8</td>
<td>et_05</td>
<td>10.79 11.39 10.57 11.02 10.16</td>
<td>17.37 31.13 14.7 12.61 11.04</td>
</tr>
<tr>
<td>CKDMM</td>
<td>π = 1.0</td>
<td>et_05</td>
<td>10.81 11.71 10.27 10.61 10.63</td>
<td>16.96 30.83 14.55 12.41 10.09</td>
</tr>
<tr>
<td>IKDMM</td>
<td>π = 1.0</td>
<td>et_05</td>
<td>10.35 11.19 10.18 10.29 9.75</td>
<td>16.86 30.42 14.01 12.24 10.93</td>
</tr>
</tbody>
</table>

3.1 Proposed Teacher mask model

The training process and configuration for the TMM is the same as for the baseline (BLSTM mask model) described in Section 2.2. As shown in Fig. 2, the TMM is trained with simulated training data. During the training stage, the TMM is used to generate soft mask labels of the real-recording data as well as the simulated data for the SMM at each iteration. Hence, for the simulated data, we can obtain two soft masks: the speech soft mask label $TMS_X(τ, ω) \in [0, 1]$, and the noise mask label $TMS_N(τ, ω) \in [0, 1]$. For the real-recording data, another two masks can be generated as well, which are denoted as $TMR_X(τ, ω) \in [0, 1]$, and $TMR_N(τ, ω) \in [0, 1]$. In the end, we can form three different sets of training data for the SMM at each training iteration: $D_1=$ (simulated noisy data, corresponding IBMs), $D_2=$(simulated noisy data, corresponding TMs), and $D_3=$(real-recording noisy data, corresponding TMs). The example of these soft mask labels are given in the (c1) and (c2) of Fig. 2. These training data will be used to train our SMM.

3.2 Proposed Student mask model

In this work, we use the same structure for the SMM as was used for the TMM (BLSTM mask model). During the SMM training stage, we use different loss functions $L_{SS}$ and $L_{SR}$ to train our SMM with the simulated data and the real-recording data, respectively.

At training iteration $ℓ$, for simulated training data $D_1$ and $D_2$, we consider the following loss function $L_{SS}^F$:

$$L_{SS}^F = πBCE(IBM_X(τ, ω), SMS_X(τ, ω)) + (1 − π)BCE(TMS_X(τ, ω), SMS_X(τ, ω)) \quad ∀v \in \{X, N\} \quad (10)$$

where $SMS_X$ and $SMS_N$ denote the estimated speech mask and noise mask for simulated data by SMM, respectively; and $π \in [0, 1]$ is the linear interpolation weight. The $IBM_X$ and $IBM_N$ are the hard mask labels of speech and noise, respectively. The $TMS_X$ and $TMS_N$ represent the soft mask labels at iteration $ℓ$, which are generated by TMM. The BCE is defined in Eq. (9).

For real-recording data, in our proposed IKDMM, the training data $D_3$ is used, where the soft mask labels of the real-recording data can be obtained from TMM. The loss function $L_{SR}$ on real-recording data at training iteration $ℓ$, is defined as:

$$L_{SR}^F = BCE(TMR_X(τ, ω), SMR_X(τ, ω)) \quad ∀v \in \{X, N\} \quad (11)$$
where the $SMR_X$ and $SMR_N$ are the estimated masks of speech and noise by SMM, respectively. The $TMR_X$ and $TMR_N$ represent the soft mask labels generated from TMM at training iteration $\ell$.

As displayed in (d) of Fig. 2, the SMM has been trained using the simulated data and real-recording data with the loss $L_{SS}$ and the loss $L_{SR}$, respectively. When the SMM predicts the speech mask and the noise mask for each channel of microphone, we calculate the beamforming coefficients by using the approach described in Section 2.1.

4 EXPERIMENTS

Our proposed IKDMM approach is evaluated on the six-channel CHiME-4 corpus [10]. Word Error Rate (WER), the common metric for ASR performance, is used. The IKDMM for acoustic beamforming is used as a front-end for the ASR system.

4.1 Corpus

The six-channel CHiME-4 corpus is created using the six microphones which are mounted on a tablet, with the second microphone in the rear and the other five microphones in the front. This corpus includes simulated data and real-recording data from four challenging daily noisy environments, i.e., public transport (BUS), pedestrian area (PED), cafe (CAF), and street junction (STR), as well as it features several male and female speakers, uttering the speaker-independent medium-vocabulary (5k) subset of the Wall Street Journal (WSJ) sentences. The corpus is divided into 3 individual subsets. The first one is the training set, containing 8738 (1600 real-recording + 7138 simulated) noisy utterances. The second one is the development set (dt_05), consisting of 3280 (1640 real-recording + 1640 simulated) noisy utterances. The third one is the evaluation set (et_05), including 2640 (1320 real-recording + 1320 simulated) noisy utterances. Each of the three real-recording subsets is recorded with four different speakers.

4.2 Metric

WER is the common metric to evaluate the performance of ASR system [20]. The WER compares a reference to an hypothesis and is defined as:

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

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 facilitate the comparisons, we use the original baseline back-end of CHiME-4 [10]. Specifically, the DNN-based acoustic model (AM) is used, which contains seven hidden layers, each with 2,048 exponential linear units. The AM is trained with all the noisy signals from all the six microphones. A standard WSJ 5k word 3-gram language model (LM) is used for decoding. Both of them are trained by using Kaldi speech recognition toolkit [21].

For the front-end of ASR, we compare the three different mask models, namely a conventional BLSTM mask model (cBLSTM) described in Section 2.2, a student mask model trained using CKDMM (described in Section 1), and our student mask model trained using IKDMM. For the CKDMM, the two BLSTM networks with the same structure as cBLSTM are designed as a teacher mask model (cTMM) and a student mask model (cSMM), respectively. The cSMM use the same form of loss function training as described in Section 3.2, but there is no iteration information in its loss functions. The cTMM is trained using all the 7138’s simulated utterances (~90h) in the training set, which is the same as the cBLSTM, and the TMM in our IKDMM. Hence, in our experiments, we only need to compare student mask models in CKDMM and IKDMM. For expression clarity, we represent the student mask model by using the name of their corresponding training scheme in Table 2. For the IKDMM, the number of iterations is set as 60. The same mask-based GEV beamformer (see Section 2) and the same ASR back-end (see the above paragraph) are used for comparison.

4.4 Results

Table 2 summarizes the ASR performance obtained in our experiments. From Table 2, the results reveal that all of the mask models, which adopt the knowledge distillation (KD) learning framework, outperform that of the cBLSTM for the simulated data and real-recording data, despite sharing the same back-end. Meanwhile, we can see that the ASR performance of our IKDMMs with different hyper-parameters $\pi$ are better than that of CKDMMs with the same parameters, as expected. Specifically, the student mask model with hyper-parameters $\pi = 0$ in proposed IKDMM approach achieves the best ASR performance, 12.55% WER, in the real-recording test set, and it can obtain relative 5% WER reduction compared to the cBLSTM for the real-recording evaluation data. This means that our proposed IKDMM for acoustic beamforming can offer good mitigation for the influence of the data mismatch problem, and it has a better generalize ability for real-life applications.

5 CONCLUSION

This work aims to reduce the adverse effect of the data mismatch on mask model learning. In this study, we adopt the knowledge distillation (KD) learning framework to the mask model for acoustic beamforming where the real-recording data can be pooled with the simulated data for mask models training. The mask models using KD training scheme are more robust against data mismatch problem. Moreover, we propose a novel iterative distillation mask model (IKDMM) training scheme where the teacher mask model generates the soft mask labels at each iteration to the student mask model. Experiments using the six-channel CHiME-4 corpus shows that our IKDMM significantly outperforms the conventional TFMaskBF and mask models using conventional knowledge distillation training scheme for beamforming in the real-recording test set. This means that the proposed IKDMM approach can alleviate the data mismatch problem and has generalization capability for applying to real-recording data.

ACKNOWLEDGMENT

This paper was partially supported by Shenzhen Science & Technology Fundamental Research Programs (No: JCYJ20170817160058246 & JCYJ20180507182908274).
REFERENCES


