Speech Enhancement Based on Teacher–Student Deep Learning Using Improved Speech Presence Probability for Noise-Robust Speech Recognition

Yan-Hui Tu, Jun Du, and Chin-Hui Lee, *Fellow, IEEE*

***Abstract*— In this paper, we propose a novel teacher-student learning framework for the preprocessing of a speech recognizer, leveraging the online noise tracking capabilities of improved minima controlled recursive averaging (IMCRA) and deep learning of nonlinear interactions between speech and noise. First, a teacher model with deep architectures is built to learn the target of ideal ratio masks (IRMs) using simulated training pairs of clean and noisy speech data. Next, a student model is trained to learn an improved speech presence probability by incorporating the estimated IRMs from the teacher model into the IMCRA approach. The student model can be compactly designed in a causal processing mode having no latency with the guidance of a complex and noncausal teacher model. Moreover, the clean speech requirement, which is difficult to meet in real-world adverse environments, can be relaxed for training the student model, implying that noisy speech data can be directly used to adapt the regression-based enhancement model to further improve speech recognition accuracies for noisy speech collected in such conditions. Experiments on the CHiME-4 challenge task show that our best student model with bidirectional gated recurrent units (BGRUs) can achieve a relative word error rate (WER) reduction of 18.85% for the real test set when compared to unprocessed system without acoustic model retraining. However, the traditional teacher model degrades the performance of the unprocessed system in this case. In addition, the student model with a deep neural network (DNN) in causal mode having no latency yields a relative WER reduction of 7.94% over the unprocessed system with 670 times less computing cycles when compared to the BGRU-equipped student model. Finally, the conventional speech enhancement and IRM-based deep learning method destroyed the ASR performance when the recognition system became more powerful. While our proposed approach could still improve the ASR performance even in the more powerful recognition system.**

***Index Terms*—Teacher-student learning, improved minima controlled recursive averaging, improved speech presence probability, deep learning based speech enhancement, noise-robust speech recognition.**

Manuscript received March 5, 2019; revised July 20, 2019; accepted August 30, 2019. Date of publication September 12, 2019; date of current version September 20, 2019. This work was supported in part by the National Key R&D Program of China under Grant 2017YFB1002202, in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the Key Science and Technology Project of Anhui Province under Grant 17030901005, and in part by the Huawei Noah's Ark Lab. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Xiaodong Cui. *(Corresponding author: Jun Du.)*

Y.-H. Tu and J. Du are with the University of Science and Technology of China, Hefei, Anhui 230052, China (e-mail: tuyanhui@mail.ustc.edu.cn; jundu@ustc.edu.cn).

C.-H. Lee is with the Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: chl@ece.gatech.edu).

Digital Object Identifier 10.1109/TASLP.2019.2940662

I. Introduction

SINGLE-CHANNEL speech enhancement (SE) [1] aims to suppress the background noise and interference from the observed noisy speech based on a single microphone setting, which is helpful to improve speech quality and the performance of automatic speech recognition (ASR) [2]. The classic noise suppressor is based on statistical signal processing and typically works in the frequency domain. The input signal is broken into overlapping frames and weighted and converted to the frequency domain, a process denoted as short-time Fourier transform (STFT). The noise suppressor applies a time-varying real-valued suppression gain to each frequency bin, based on the estimated presence of speech signal—close to zero if there is mostly noise and close to one if there is mostly speech. To estimate the suppression gain, most approaches assume that the noise changes slower than the speech signal and that Gaussian distributions for the noise and speech signal magnitudes. They build a noise model with noise variances for each frequency bin, typically using voice activity detector (VAD). The suppression rule is a function of the *a priori* and *a posterior* signal-to-noise ratios (SNRs). The oldest and still most commonly used rule is the Wiener suppression rule [3], which is optimal in the mean square error sense. Other frequently used suppression rules are the spectral magnitude estimator [4], maximum likelihood amplitude estimator [5], short-term minimum mean square error (MMSE) estimator [6] and the log-spectral minimum mean square error (log-MMSE) estimator [7]. These conventional techniques adapt to the noise level and perform well with quasi-stationary noises, but impulse nonspeech signals cannot be well suppressed. In [8], minima controlled recursive averaging (MCRA) was introduced by a noise estimation approach that combines the robustness of minimum tracking with the simplicity of recursive averaging. In [9], improved minima controlled recursive averaging (IMCRA) was proposed. The first iteration provides rough voice activity detection in each frequency band. Then, smoothing in the second iteration excludes relatively strong speech components, which makes minimum tracking during speech activity robust. This facilitates larger smoothing windows and thus a decreased variance of the minima values. The above mentioned methods are considered as unsupervised techniques that have been studied extensively in the past several decades.

However, recent advances in computational auditory scene analysis (CASA) [10], [11] and machine learning have inspired

new approaches, e.g., support vector machine (SVM) [12], nonnegative matrix factorization (NMF) [13]–[18] and deep neural network (DNN) [19]–[24]-based techniques, aiming at estimating either a clean speech feature at each time-frequency (T-F) bin directly or a T-F mask that is applied to the T-F bin of noisy speech to recover clean speech. For several potential future applications, e.g., DNN-based SE algorithms for hearing aids or mobile communications, the range of possible acoustic situations that can realistically occur is virtually endless. Specifically, in [20], [23], DNN was proposed as a nonlinear spectral regression model to map the log-power spectra (LPS) features of noisy speech [25] to those of clean speech. In [19], DNN was adopted to estimate the ideal masks including the ideal binary mask (IBM) [26] of one T-F bin and the ideal ratio mask (IRM) [22] of one T-F bin. [22] also demonstrated that IRM as the learning target led to better speech enhancement performance than that of IBM. The above mentioned methods are based on the DNN model, where the relationship between the neighboring frames is not explicitly modeled. Recurrent neural networks (RNNs) [27] may solve this problem using recursive structures between the previous frame and the current frame to capture the long-term contextual information and make better predictions. In [28], [29], a long short-term memory recurrent neural network (LSTM-RNN) was proposed for speech enhancement compared to DNN-based speech enhancement, yielding superior noise reduction performance at low SNRs.

Based on the above introduction, unsupervised and deep learning based single-channel speech enhancement approaches have demonstrated different strengths and weaknesses. For unsupervised method, e.g., IMCRA-based approach, it is an online adaptive algorithm of a few parameters to the test conditions, while a tradeoff in reducing speech distortion and residual noise needs to be made due to the sophisticated statistical properties of the interactions between speech and noise signals. Most of these unsupervised methods are based on either the additive nature of the background noise or the statistical properties of speech and noise. However, they often fail to track nonstationary noises for real-world scenarios in unexpected acoustic conditions. On the other hand, for deep learning methods, nonstationary background noise maybe handled well, with a large amount of training data simulated by different noise levels and types. However, when a mismatch exists between the training and test conditions, the quality of the estimated speech is usually degraded. Recently, [30] presented a novel architecture in which the general structure of a conventional noise suppressor was preserved, but the subtasks (VAD, noise variance estimation, and IRM estimation) were independently learned and carried out by separate neural networks. In [31], a student-teacher learning paradigm for single-channel speech enhancement was proposed. The teacher network is adopted to estimate the T-F masks from the beamformed speech obtained by multichannel enhancement. Then, the estimated masks are employed as the learning target of a student network with only single-channel input. Although experiments on the single-channel track of the CHiME-4 challenge showed some ASR performance improvements, the training data for the student-teacher model must come from multichannel recordings.

In this study, a novel teacher-student learning framework is proposed, which utilizes the advantages of both IMCRA with a well-designed online noise tracking procedure and deep learning approaches, providing strong prior information of the interactions between speech and noise. First, a teacher model with deep architectures is built to learn the target of IRMs using the simulated training pairs of clean and noisy speech data. Next, a better learning target, namely, the improved speech presence probability (ISPP), is designed by incorporating the estimated IRMs from the teacher model into the procedure of IMCRA approach. Then, the student model with deep architectures is trained to estimate the ISPP. Using the teacher-student learning framework, the student model can be compactly designed in the causal processing mode having no latency with the guidance of a complex and noncausal teacher model, which better meets the run-time requirements of realistic applications. Moreover, the stereo-data constraint requiring target clean speech is relaxed for student model training, implying that the realistic noisy speech data without the underlying clean speech can be directly used to further improve the recognition accuracy by adapting the enhancement model to potentially adverse environments.

In our proposed teacher-student learning framework, the IRMs accurately estimated by the teacher model have a great influence on the performance of the student model. Therefore, powerful neural networks with a large amount of training data pairs are necessary. However, in terms of recognition accuracy of ASR systems, speech enhancement based on the traditional teacher model using bidirectional LSTM (BLSTM) could not improve ASR performances, as shown in [31], [32]. This result is because, in highly mismatched testing conditions, IRMs estimated from the teacher model might misclassify the T-F regions dominated by nonspeech and noise. Meanwhile, IMCRA-based mask estimation can alleviate this problem by conservative noise reduction, but the estimation still cannot lead to significant ASR performance gains due to a large amount of residue noises. The ISPP-based student model simultaneously performs aggressive noise reductions and less speech distortions with the collaboration between deep learning-based IRM estimation and IMCRA-based mask estimation. Experiments on the CHiME-4 challenge task show that our best student model with bidirectional gated recurrent units (BGRUs) [33] can achieve a relative word error rate (WER) reduction of 18.85% on the real test set when compared to an unprocessed system without acoustic model restraint; however, the conventional teacher model degrades the unprocessed system performance in this case. In addition, the student model with a compact DNN in causal mode having no latency yields a relative WER reduction of 7.94% over the unprocessed system, with 670 times less computing cycles when compared to the BGRU-equipped student model.

This study is comprehensively extended from our previous work in [34], [35] with the following contributions. First, ISPP is adopted as a new learning target of deep models at the training stage, while ISPP in [34] is generated at the recognition stage requiring complex and time consuming computation. Second, the clean speech requirement is relaxed for the student model training, which can utilize a large amount of real-world noisy speech to fine-tune the enhancement model. Third, the proposed

learning framework facilitates a compact student model design. Finally, more experiments are designed with more detailed result analysis.

The remainder of this paper is organized as follows. In Section II, we briefly introduce the conventional IMCRA approach as preliminaries. In Section III, we present an overview of the proposed teacher-student learning framework and give a detailed description of teacher and student model training. Section IV discusses the experiments on the CHiME-4 challenge. Finally, we summarize our findings in Section V.

II. Prior Art: Improved Minima Controlled Recursive Averaging

In this section, the key principle of the IMCRA approach is briefly introduced as the preliminaries of calculating ISPP in Section III-B. First, we consider the problem of recovering a desired signal $s(n)$, when the observed signal $x(n)$ is its noisy version corrupted by additive background noise, i.e.,

$$x(n) = s(n) + d(n), \tag{1}$$

where n is a discrete-time index. The desired signal and background noise are assumed to be zero mean and mutually uncorrelated. The observed signal $x(n)$ is divided into overlapping frames by the application of a window function and analyzed using STFT. Specifically,

$$X(k,l) = S(k,l) + D(k,l), \tag{2}$$

where k denotes the frequency bin index, and l denotes the frame index. $S(k,l)$, $D(k,l)$ and $X(k,l)$ denote the STFT of desired clean speech, noise and noisy speech signals, respectively. To obtain an estimation of the desired clean signal, a specific gain function was applied to each spectral component of the noisy speech signal as follows:

$$\widehat{X}(k,l) = G(k,l)X(k,l), \tag{3}$$

$$G(k,l) = \frac{\xi(k,l)}{1+\xi(k,l)} \exp\left(\frac{1}{2}\int_{v(k,l)}^{\infty} \frac{e^{-t}}{t}\mathrm{d}t\right), \tag{4}$$

$$v(k,l) \triangleq \frac{\gamma(k,l)\xi(k,l)}{1+\xi(k,l)} \tag{5}$$

where $G(k,l)$ is the gain function. $\gamma(k,l)$ and $\xi(k,l)$ denote the *a posterior* SNR and *a priori* SNR, respectively. The log-spectral amplitude (LSA) estimator [7] is utilized among many speech enhancement methods due to its superiority in reducing musical noise. Clearly, the key point here is an accurate estimation of the *a priori* and *a posterior* SNRs. To achieve this, IMCRA is based on two hypotheses, $H_0(k,l)$ and $H_1(k,l)$, which indicate, respectively, speech absence and speech presence in the k-th frequency bin of the l-th frame as follows:

$$\begin{aligned} H_0(k,l) &: X(k,l) = D(k,l), \\ H_1(k,l) &: X(k,l) = S(k,l) + D(k,l). \end{aligned} \tag{6}$$

Accordingly, the *a priori* SNR $\xi(k,l)$ and *a posterior* SNR $\gamma(k,l)$ can be defined as follows:

$$\xi(k,l) \triangleq \frac{\lambda_s(k,l)}{\lambda_d(k,l)} \tag{7}$$

$$\gamma(k,l) \triangleq \frac{|X(k,l)|^2}{\lambda_d(k,l)} \tag{8}$$

where $\lambda_s(k,l) = E[|S(k,l)|^2 \big| H_1(k,l)]$ and $\lambda_d(k,l) = E[|D(k,l)|^2]$ denote the variances of desired speech and noise, respectively. For estimating *a posterior* SNR, only the noise is necessary to be estimated, by initializing $\lambda_d(k,l)$ at the first frame with $\lambda_d(k,0) = |X(k,0)|^2$. Then, $\lambda_d(k,l+1)$ is calculated by a recursive averaging between $\lambda_d(k,l)$ and $|X(k,l)|^2$. The corresponding smoothing factor is estimated using the minima controlled algorithm, which is related to $\xi(k,l)$ and $\gamma(k,l)$. The *a priori* SNR is estimated as follows:

$$\begin{aligned} \xi(k,l) = {} & \alpha G^2(k,l-1)\gamma(k,l-1) \\ & + (1-\alpha)\max\{\gamma(k,l-1) - 1, 0\} \end{aligned} \tag{9}$$

where α is a weighting factor that controls the tradeoff between noise reduction and speech distortion [6], [36]. More details can be found in [9].

III. Overall Teacher-Student Learning

The proposed teacher-student learning framework is illustrated in Fig. 1, consisting of two stages, namely, the teacher model training and the student model training, shown respectively in the left and right of the vertical dashed line in the middle of the figure. The dotted lines represent the process of obtaining the learning targets of the teacher and student models, while the solid lines denote the training process for teacher and student models. In the training stage of the teacher model, as shown in the left part of Fig. 1, a deep model (e.g., DNN, BLSTM, or BGRU) is employed to learn the mapping relationship between the simulated noisy training data and the IRM calculated by the training data pairs. The role of the teacher model is to calculate the ISPP target for student model training by incorporating with the IMCRA process. As shown in the right part of Fig. 1, only the noisy speech data are necessary to train the student model with the help of the teacher model, which relaxes the needs for clean speech data. The details of training the teacher and student models and their corresponding motivations are elaborated in the following.

In Section II, the estimation of the gain function at the current T-F bin is only based on the statistics of history frames, which is an online adaptive algorithm to testing environments. However, due to the strong model assumptions of speech and noise signals, IMCRA is not always robust in adverse environments, particularly when there are nonstationary noises. Therefore, a deep learning-based approach is a strong complementarity of the IMCRA approach. Following the framework in Fig. 1, the details of proposed teacher and student model training are elaborated in the following sections.

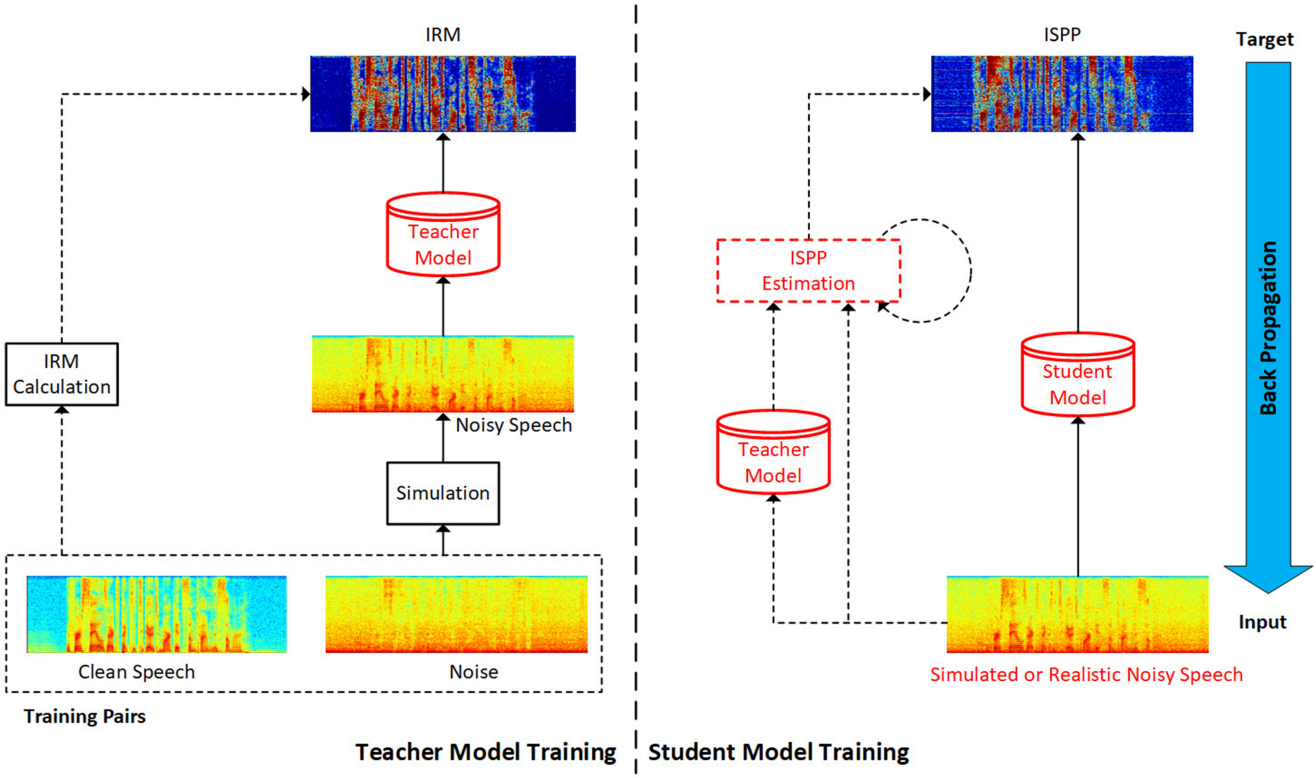


Fig. 1. Illustration of the proposed teacher-student learning framework.

A. *Teacher Model Training*

The teacher model adopts the neural network with deep architectures to estimate widely used IRMs from the noisy log-power spectra (LPS) features. Acoustic context information along both the time axis (with multiple neighboring frames) and the frequency axis (with full frequency bins) can be fully exploited by the neural network to obtain a good mask estimate in adverse environments, which is strongly complementary with the conventional IMCRA-based approach to retaining 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 at each T-F bin. The IRM as the learning target is defined as follows:

$$M_{\text{Ref}}(k,l) = S_{\text{PS}}(k,l)/\left[S_{\text{PS}}(k,l) + D_{\text{PS}}(k,l)\right], \qquad (10)$$

where $S_{\text{PS}}(k,l)$ and $D_{\text{PS}}(k,l)$ are power spectral features of clean speech and noise at the T-F unit (k,l). Training of the teacher model requires a large amount of time-synchronized stereo-data; thus, the simulation data are often synthesized by adding different types of noises 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 realistic speech data. To train the teacher model with a random initialization, supervised fine-tuning is used to minimize the mean squared error between the output of teacher neural network $\hat{M}_{\text{TeNN}}(k,l)$ and the reference IRM $M_{\text{Ref}}(k,l)$, defined as follows:

$$E_{\text{TeNN}} = \sum_{k,l} (\hat{M}_{\text{TeNN}}(k,l) - M_{\text{Ref}}(k,l))^2. \qquad (11)$$

A Adam-based backpropagation method [37] can then be adopted to update the parameters of a neural network in a mini-batch mode. Each mini-batch is one training utterance. At the test stage, the teacher model is directly utilized for decoding and generating the enhanced speech.

B. *Improved Speech Presence Probability (ISPP)*

Before introducing student model training, the proposed ISPP as the new learning target is elaborated. In IMCRA, the tradeoff in reducing speech distortion and residual noise is made due to the sophisticated statistical properties of the interactions between speech and noise signals. Most of these unsupervised methods such as IMCRA are based on either the additive nature of the background noise or the statistical model assumptions of speech and noise signals. They often fail to track nonstationary noises in real-world scenarios with unexpected acoustic conditions. For the deep learning method, it can often well handle the nonstationary background noise based on the prior knowledge learned from a large amount of training data simulated by different noise types and levels. However, when there exists a high mismatch between training data and test data, large speech distortion is usually generated with an aggressive noise reduction. Accordingly, ISPP aims to fully utilize the advantages

of the gain function in the IMCRA approach and the IRM in deep learning-based LPS regression.

From Eq. (4), the gain function $G(k,l)$ mainly relies on the *a posterior* SNR $\gamma(k,l)$ and *a priori* SNR $\xi(k,l)$. According to description of Section II, $\gamma(k,l)$ is related to $\gamma(k,l-1)$ and $\xi(k,l-1)$, while $\xi(k,l)$ depends on $G(k,l-1)$ and $\gamma(k,l-1)$. Therefore, the gain function, *a priori* SNR, and *a posterior* SNR are recursively coupled between consecutive frames. To improve the accuracy of these three estimations in adverse environments, we incorporate neural network-based mask estimation $\hat{M}_{\text{TeNN}}(k,l-1)$ to define an intermediate item $\hat{G}(k,l-1)$:

$$\hat{G}(k,l-1) = \delta \hat{M}_{\text{TeNN}}(k,l-1) + (1-\delta) G_{\text{ISPP}}(k,l-1), \tag{12}$$

where $G_{\text{ISPP}}(k,l-1)$ denotes ISPP-based gain function at T-F bin $(k,l-1)$ and δ is a weighting factor empirically set to 0.9 in our experiments. For the l-th frame, we first compute the noise estimation using the same algorithm as in IMCRA [9] with the statistics of previous frames, namely, ISPP-based *a posterior* SNR $\gamma_{\text{ISPP}}(k,l-1)$ and *a priori* SNR $\xi_{\text{ISPP}}(k,l-1)$. Then, we compute $\gamma_{\text{ISPP}}(k,l)$ using Eq. (8). Next, $\xi_{\text{ISPP}}(k,l)$ is calculated by modifying Eq. (9):

$$\begin{aligned}\xi_{\text{ISPP}}(k,l) = {} & \alpha \hat{G}^2(k,l-1)\gamma_{\text{ISPP}}(k,l-1) \\ & + (1-\alpha)\max\{\gamma_{\text{ISPP}}(k,l-1) - 1, 0\}.\end{aligned} \tag{13}$$

Finally, the new gain function, namely, improved speech presence probability, at the l-th frame is computed according to Eqs. (14) and (15) as follows:

$$G_{\text{ISPP}}(k,l) = \frac{\xi_{\text{ISPP}}(k,l)}{1+\xi_{\text{ISPP}}(k,l)} \exp\left(\frac{1}{2}\int_{v_{\text{ISPP}}(k,l)}^{\infty} \frac{e^{-t}}{t} \mathrm{d}t\right), \tag{14}$$

$$v_{\text{ISPP}}(k,l) \triangleq \frac{\gamma_{\text{ISPP}}(k,l)\xi_{\text{ISPP}}(k,l)}{1+\xi_{\text{ISPP}}(k,l)}. \tag{15}$$

The procedure of ISPP estimation is presented as Algorithm 1. Obviously, the calculation of ISPP is a recursive process, which simultaneously possesses a strong online tracking/adaption capability and an accurate estimation capability for related statistics by incorporating the strong prior information of the speech and noise signals from the teacher model. By comparing Eqs. (4), (10) and (14), although $G(k,l)$, $M_{\text{Ref}}(k,l)$, and $G_{\text{ISPP}}(k,l)$ have different definitions, their values are all in the range of [0, 1].

Algorithm 1: The Procedure of ISPP Estimation.

Input: One noisy speech utterance and the teacher model.
Output: ISPP estimation, denoted as $G_{\text{ISPP}}(k,l)$.

```
Initialize the statistics at the first frame for all frequency bins: ξ_ISPP(k,0) = 0; γ_ISPP(k,0) = 1; G_ISPP(k,0) = 1.
for all time frames l do
    for all frequency bins k do
        Compute the mask estimation M̂_TeNN(k,l−1) using the teacher model of Section III-A with the input of noisy LPS features centered at T-F bin (k,l−1).
        Compute Ĝ(k,l−1) by combining M̂_TeNN(k,l−1) and G_ISPP(k,l−1) according to Eq. (12).
        Compute the a posterior SNR γ_ISPP(k,l) according to Eq. (8) by using the noise estimation algorithm in [9] with γ_ISPP(k,l−1) and ξ_ISPP(k,l−1).
        Compute the a priori SNR ξ_ISPP(k,l) with Ĝ(k,l−1) and γ_ISPP(k,l−1) according to Eq. (13).
        Compute the gain function or ISPP G_ISPP(k,l) with γ_ISPP(k,l) and ξ_ISPP(k,l) according to Eq. (14).
    end for
end for
```

C. Student Model Training

Based on the above analysis, the main problem of the conventional teacher model using IRM as the learning target is that the large speech distortions might be generated in the enhanced speech when there is a high mismatch between the training and test data, especially in realistic applications. [38] demonstrated that the direct mapping from the noisy speech features to clean speech features using DNN regression model led to the performance degradation in low SNR cases. Even though the regression BLSTM model is used to estimate the IRM, it cannot improve the ASR performance on realistic test data as reported in [31], [32]. To alleviate this problem, ISPP is designed as a better learning target for the student model, which is elaborated in Section III-B.

As illustrated in the right part of Fig. 1, the dotted line represents the process of obtaining ISPP, namely, the training target of the student model. The input to the student model is the same as that of the teacher model, namely, noisy LPS features. To train the student model with a random initialization, supervised fine-tuning is used to minimize the mean squared error between the student model output $\hat{G}_{\text{StNN}}(k,l)$ and the estimated ISPP $G_{\text{ISPP}}(k,l)$ from Algorithm 1, defined as:

$$E_{\text{StNN}} = \sum_{k,l} (\hat{G}_{\text{StNN}}(k,l) - G_{\text{ISPP}}(k,l))^2. \tag{16}$$

The Adam-based backpropagation method is adopted to update the parameters of neural networks in mini-batch mode. Each training utterance is treated as one mini-batch. At the test stage, the student model is directly utilized for decoding and generating enhanced speech according to Eq. (3).

One advantage of our teacher-student learning is that the stereo-data constraint is relaxed for student model training. For the conventional IRM-based teacher model, only the simulation training data can be used due to this constraint, which is one main reason leading to the mismatch with the test data of realistic applications. However, for the student model, as the learning target ISPP is calculated via the input noisy speech spectra and the teacher model, the underlying clean speech signals are not necessary. Accordingly, both simulated and realistic noisy speech data could be adopted for training the student model.

Algorithm 2: The Procedure of Student Model Training.

Input: Simulated/realistic training data and the teacher model.

Output: The parameter set of student model.

1: Randomly initialize the parameters of the student model.
2: **for** each mini-batch **do**
3: Compute noisy speech spectra features and LPS features for all T-F bins in this mini-batch.
4: Compute the learning target $G_{\text{ISPP}}(k,l)$ for all T-F bins in this mini-batch using the teacher model and noisy speech spectra features according to Algorithm 1.
5: Accumulate gradients using noisy speech LPS features and $G_{\text{ISPP}}(k,l)$ in this mini-batch via Eq. (16).
6: Update the parameters of the student model using Adam.
7: **end for**

This is quite valuable because the realistic noisy speech data can potentially reduce the mismatch between the training and test data. More discussion will be presented in Section IV-E. Our proposed procedure of student model training is summarized in Algorithm 2.

Another advantage of our teacher-student learning is that the student model can be compactly designed. The IMCRA approach is performed efficiently and inherently in causal mode having no latency; thus, it is also crucial to design a compact student model to work in causal mode having no latency for the applications with a high demand of run-time efficiency. In the proposed framework, the neural network architecture of the student model is not necessarily the same as that of the teacher model. We can use a complicated teacher model such as BGRU to guarantee a decent performance of a simple student model such as DNN in a causal mode. We will discuss this topic more in Section IV-F.

To further illustrate the motivation of the proposed ISPP-based student model, Fig. 2 gives an utterance example from the real test set of CHiME-4. Fig. 2(a) and (b) plot the spectrograms from Channel 0 (the close-talking microphone to record the reference "clean" speech) and one corresponding channel with noisy speech. Fig. 2(c) plots the mask or gain function estimated by IMCRA, while Fig. 2(d) plots IRM from the teacher model with DNN. We observed that the estimated IRM by the teacher model might misclassify the T-F regions dominated by speech to nonspeech/noise, while the IMCRA method could alleviate this problem by generating the estimated mask with much higher values in the black rectangle. The masks estimated by the IMCRA method were often noncontinuous among consecutive speech frames. Clearly, there exists a strong complementarity between these two types of methods. In Fig. 2(e), the mask or ISPP estimated by the student model with DNN could fully utilize the complementarity IMCRA approach and IRM-based deep learning approach, achieving both better speech preservation and speech continuity.

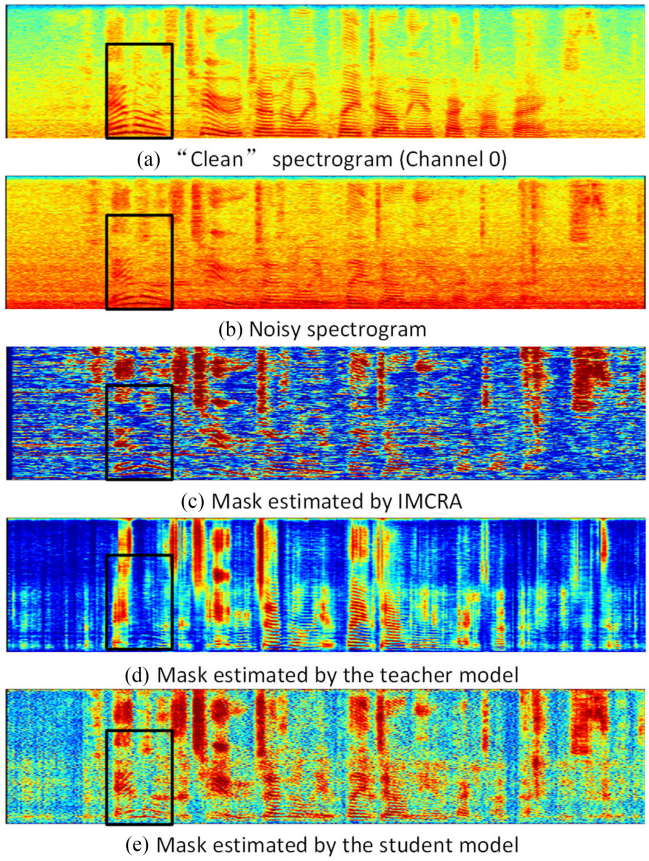


Fig. 2. The comparison of estimated masks of different approaches for an utterance from the real test set of CHiME-4.

IV. Experimental Evaluation

A. Data Corpus

We present the experimental evaluation of our framework in the CHiME-4 speech recognition task [39], which was designed to study real-world ASR 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 as follows: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). For each case, two types of noisy speech data were provided as follows: RealData and SimData. RealData were collected from talkers reading the same sentences from the WSJ0 corpus [40] in the four conditions. SimData, on the other hand, were constructed by mixing clean utterances with environmental noise recordings using the techniques described in [41]. The CHiME-4 offered three tasks including single-channel (1-channel) and multi-channel (2-channel, and 6-channel) tasks, and this study focused on the single-channel speech enhancement. So we used the 1-channel task to evaluate our algorithm. The readers can refer to [39] for more detailed information regarding CHiME-4.

B. Implementation Details

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. The STFT analysis is used to compute the DFT of each overlapping windowed frame. To train the regression model, the 257-dimensional feature vector was used for IRM and ISPP targets. PyTorch was used for neural network training [42]. The learning rate for the first 5 epochs was initialized as 0.25 and then decreased by 90% after each epoch, and the number of epochs was 10. The CHiME-4 challenge [43] training set was used as our training data. Specifically, we used simulated training data from Channel 1, Channel 3 and Channel 5 with 7138 utterances (about 12 hours) for each channel to train the teacher and student models.

The default configurations of different neural network architectures are as follows. DNN was fixed at 3 hidden layers, 2048 units for each hidden layer, and 1799-dimensional input LPS feature vector with 7-frame expansion. LSTM, BLSTM and BGRU were fixed at 2 hidden layers, 1024 units for each hidden layer, and 257-dimensional input with no frame expansion. For the IMCRA and ISPP approaches, all the tuning parameters in Algorithm 1 were set according to [9].

The ASR system officially provided in [39] was adopted to evaluate the recognition performance of different enhancement methods. The acoustic model is a DNN-HMM (hybrid hidden Markov model with DNN to estimate state posterior probability) discriminatively trained with the sMBR criterion [44]. The input of the DNN-HMM is a 440-dimensional feature vector extracted from Channel 5, consisting of a 40-dimensional fMLLR [45] with an 11-frame expansion. The language models are 5-gram with Kneser-Ney (KN) smoothing [46] for the first-pass decoding and the simple RNN-based language model [47] for rescoring. The model is trained according to the scripts downloaded from the official GitHub website[1] using Kaldi toolkit [48]. Note that all enhancement methods are only applied to the utterances in the recognition stage without retraining the acoustic model.

C. Motivation Experiments

First, we would provide one set of the recognition experiments to show the main motivation of our proposed approach. Table I shows a WER(%) comparison of the conventional IMCRA approach, IRM-based deep learning approaches using different neural network architectures and proposed ISPP approach for single-channel enhancement on the real test set. "Noisy" denotes the recognition of original noisy speech randomly selected from Channels 1-6 (except Channel 2), namely, 1-channel case. "IMCRA" denotes the recognition of enhanced speech obtained by IMCRA-based enhancement. "DNN-IRM," "LSTM-IRM," "BLSTM-IRM" and "BGRU-IRM" denote the recognition of enhanced speech obtained by the IRM-based teacher models using the DNN, LSTM, BLSTM and BGRU architectures, respectively. We observed that the IMCRA method slightly improved the ASR performance with an average WER reduced from 23.56% to 23.12%. Additionally, mixed results of the IMCRA method were demonstrated for different environments, e.g., effective for BUS and STR but ineffective for CAF and PED. Meanwhile, all IRM-based deep learning approaches degraded the ASR performances, e.g., an average WER of 26.61% for "DNN-IRM". Furthermore, more powerful deep architectures led to better recognition results. "BGRU-IRM" achieved the best results among all deep models with an average WER of 23.59%, which is used as the default teacher model in the following experiments. Finally, for our proposed ISPP method by combining BGRU-IRM and IMCRA, it can directly improve the performance, e.g., an average WER reduced from 23.56% to 19.77% comparing to "Noisy". In the following experiments, the ISPP is utilized as the learning target of student model.

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/chime4

TABLE I
WER (%) COMPARISON OF CONVENTIONAL IMCRA APPROACH, IRM-BASED DEEP LEARNING APPROACH USING DIFFERENT NEURAL NETWORK ARCHITECTURES AND PROPOSED ISPP APPROACH FOR SINGLE-CHANNEL SPEECH ENHANCEMENT ON THE REAL TEST SET

Enhancement	BUS	CAF	PED	STR	AVG
Noisy	36.10	24.45	19.39	14.29	23.56
IMCRA	32.59	25.96	20.91	13.02	23.12
DNN-IRM	40.53	29.27	20.98	15.65	26.61
LSTM-IRM	42.18	26.52	19.21	15.15	25.76
BLSTM-IRM	39.86	24.26	18.07	14.14	24.08
BGRU-IRM	39.37	23.86	17.35	13.78	23.59
ISPP	28.34	22.06	17.17	11.52	19.77

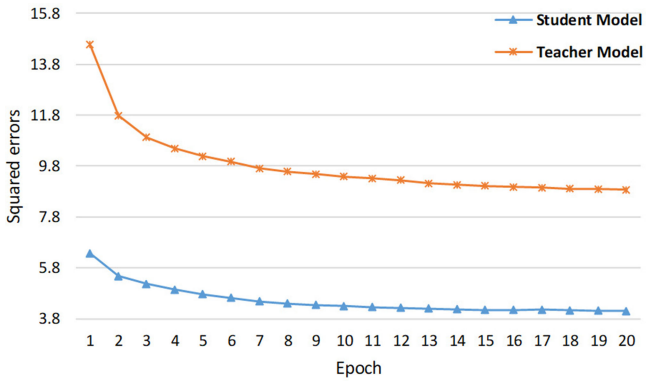


Fig. 3. Learning curves of the student and teacher models using BGRU architecture on the simulated development set.

D. Experiments on BGRU Teacher/Student Models

In this section, we would show the experiments on teacher-student learning by using the same BGRU architecture for both teacher and student models. Fig. 3 illustrates a comparison of learning curves between the student and teacher models using the averaged squared errors on the simulated development set. Clearly, the learning curve of the student models with the ISPP targets could achieve more stable and better convergence than those of the teacher model with the IRM targets. More interestingly, the initial point of the learning curve of the student model was much lower than that of the teacher model, which

TABLE II
WER (%) COMPARISON OF THE TEACHER MODELS (BGRU-IRM) AND STUDENT MODELS (PROPOSED) USING DIFFERENT SETTINGS OF (N_L, N_U) ON THE REAL TEST SET. N_L AND N_U DENOTED THE NUMBER OF HIDDEN LAYERS AND HIDDEN UNITS FOR BGRU, RESPECTIVELY

Enhancement	(N_L, N_U)	BUS	CAF	PED	STR	AVG
BGRU-IRM	(2,200)	42.18	26.52	19.21	15.15	25.76
	(2,512)	41.12	24.08	17.79	14.22	24.30
	(3,512)	41.04	24.13	17.65	14.25	24.27
	(2,1024)	39.37	23.86	17.35	13.78	**23.59**
Proposed	(2,200)	29.39	23.27	17.66	11.95	20.57
	(2,512)	29.33	22.34	17.40	11.80	20.22
	(3,512)	29.21	22.56	17.21	11.69	20.17
	(2,1024)	29.91	21.48	16.33	12.20	**19.98**
TeOracle-StBGRU	(2,1024)	28.14	21.06	16.21	11.75	**19.29**

demonstrated that the ISPP targets were more easily optimized than the IRM targets.

Table II shows WER (%) comparison of the teacher models (denoted as "BGRU-IRM") and the student models (denoted as "Proposed") using different settings of (N_L, N_U) on the real test set. N_L and N_U denote the number of hidden layers and hidden units for BGRU, respectively. There are three blocks in Table II for the teacher and student models.

For the first block of Table II, "BGRU-IRM" denotes the recognition of enhanced speech by the estimated IRM of the teacher models using different BGRU settings of (N_L, N_U). WERs of "BGRU-IRM" were deceased by increasing the number of hidden cells, e.g., a relative WER reduction of 8.42% from "BGRU-IRM(2,200)" to "BGRU-IRM(2,1024)" on average. Moreover, the number of hidden layers has no significant effect on the recognition performance of "BGRU-IRM," e.g., WER from 24.30% of "BGRU-IRM(2,512)" to 24.27% of "BGRU-IRM(3,512)". Accordingly, the teacher model was fixed at "BGRU-IRM(2,1024)" for the subsequent experiments.

For the second block, "Proposed" denotes the recognition of enhanced speech by the estimated ISPP of the student models using different BGRU settings of (N_L, N_U). We observed that ISPP estimated by all student models could directly improve the ASR performances without acoustic model retraining, while the best teacher model only achieved a comparable performance of "Noisy" in Table I. For example, "Proposed(2,1024)" yielded a relative WER reduction of 15.20% in average, when compared to "Noisy" in Table I. Second, the performance gaps among the student models with different architectures were much smaller than those among different teacher models. For example, WER was reduced from 25.76% of "BGRU-IRM(2,200)" to 23.59% of "BGRU-IRM(2,1024)," while only an absolute gain of 0.59% was generated from "Proposed(2,200)" to "Proposed(2,1024)". This also shows that the proposed teacher-student learning method is easier to optimize and more robust, even with a simple architecture.

For the third block, "TeOracle-StBGRU" denotes the student model trained with the oracle ISPP estimated by the combination of the calculated IRM used to train the teacher model and IMCRA method. We observed that the student model trained with the oracle ISPP was better than the proposed model. For example, "TeOracle-StBGRU" yielded a relative WER reduction of 3.45% in average, when compared to "Proposed(2,1024)," which could be considered as the performance gap due to the limitations of the teacher model. But the oracle ISPP can't be utilized for realistic training data due to the absence of clean speech, in the following experiments the realistic training data will be utilized in our learning framework to make up the gap.

TABLE III
WER (%) COMPARISON OF REALISTIC SPEECH DATA AUGMENTATION FOR STUDENT MODEL TRAINING ON THE REAL TEST SET

Model	Training Data	BUS	CAF	PED	STR	AVG
Student	SimData	29.91	21.48	16.33	12.20	19.98
Student	SimData+RealData	27.73	21.13	16.07	11.56	**19.12**

E. Experiments on Realistic Training Data

As discussed in Section III-C, real-world noisy speech could be adopted in our learning framework to train the student model. Table III shows WER (%) comparisons of realistic speech data augmentation for student model training on the real test set. The realistic training set was from Channel 1, Channel 3, and Channel 5 with 1600 utterances for each channel to train the teacher and student models. By adding the realistic training data, we achieved consistent and remarkable improvements for all noisy environments with a relative WER reduction of 4.30% on average over the best configured student model built with only simulated training data in Table III. The utilization of realistic speech data is quite important because an unlimited amount of speech data can be potentially collected from real-world applications and used to largely reduce the mismatch between the training and test environments.

F. Experiments on Compact Student Model

In addition to recognition accuracy, computing cycles and model size are also crucial for deep learning based speech enhancement methods in the applications with a high demand of efficiency. As discussed in Section III-C, our proposed teacher-student learning framework facilitates a compact model design. In this section, we conduct an overall comparison of both the recognition accuracy and practical issues among different neural network architectures for the teacher and student models.

Table IV shows WER (%) comparisons of the student models using different settings of (τ, N_L, N_U) on the real test set. τ, N_L, N_U denotes the number of expansion frames in the input layer, the number of hidden layers and the number of hidden units, respectively. "TeBGRU-StDNN" and "TeBGRU-StBGRU" represent the student models using DNN and BGRU models guided by the same teacher model, 'BGRU-IRM(1,2,2048)," respectively. It is well known that the computing cycles of DNN is much less than that of BGRU, and the number of expansion frames τ determines the latency of deep models. τ plays an important role in DNN modeling for achieving a decent performance, as illustrated in Table IV, while $\tau = 1$ is

TABLE IV
WER (%) COMPARISON OF THE TEACHER AND STUDENT MODELS USING DIFFERENT SETTINGS OF (τ, N_L, N_U) ON THE REAL TEST SET. τ, N_L, N_U DENOTE THE NUMBER OF EXPANSION FRAMES IN THE INPUT LAYER, THE NUMBER OF HIDDEN LAYERS AND UNITS, RESPECTIVELY. N_M AND N_T DENOTE THE MODEL SIZE AND COMPUTING CYCLES NORMALIZED BY THOSE OF THE DNN-IRM(1,3,2048) MODEL, RESPECTIVELY

Enhancement	(τ, N_L, N_U)	BUS	CAF	PED	STR	AVG	N_{M}	N_{T}
Noisy		36.10	24.45	19.39	14.29	23.56		
DNN-IRM	(1,3,2048)	41.23	30.12	22.13	16.21	27.42	1.00	1.00
	(5,3,2048)	40.97	29.89	21.43	15.97	27.06	1.22	1.13
	(7,3,2048)	40.53	29.27	20.98	15.65	26.61	1.33	1.34
BGRU-IRM	(1,2,1024)	39.37	23.86	17.35	13.78	23.59	2.89	673
TeBGRU-StDNN	(1,3,2048)	31.62	24.17	17.94	13.04	21.69	**1.00**	**1.00**
	(5,3,2048)	31.43	23.96	17.67	12.94	21.50	1.22	1.13
	(7,3,2048)	31.13	23.78	17.56	12.85	21.33	1.33	1.34
TeBGRU-StBGRU	(1,2,1024)	27.73	21.13	16.07	11.56	**19.12**	2.89	673

set for BGRU, as its structure can inherently capture the temporal constraints. $\tau = 1$ used only the central frame with no hard delay from the input, while $\tau = 5, 7$ employed both $2, 3$ history and future frames. For "BGRU-IRM" with the setting of (1,2,1024), although it outperformed "DNN-IRM," e.g., a relative WER reduction of 13.97% on average from "DNN-IRM(1,3,2048)" to "BGRU-IRM(1,2,1024)," there existed an utterance delay at the decoding stage. By using our proposed teacher-student learning framework, a better tradeoff between recognition accuracy and computing cycles could be made. For example, based on the same DNN architecture, a relative WER reduction of 1.66% in average from "TeBGRU-StDNN(1,3,2048)" to "TeBGRU-StDNN(7,3,2048)" was less than a relative WER reduction of 2.95% from "DNN-IRM(1,3,2048)" to "DNN-IRM(7,3,2048)". This implies that the proposed student model with ISPP target is easier to be optimized and more robust with small architectures. Moreover, even "TeBGRU-StDNN(1,3,2048)" with no hard delay ($\tau = 1$) could achieve significant improvements for all noise environments over best DNN-based and BGRU-based teacher models ("DNN-IRM(7,3,2048)" and "BGRU-IRM(1,2,1024)") with relative WER reductions of 18.49% and 8.05% on average, respectively.

In Table IV, the practical issues are also compared. N_M and N_T denote the model size and computing cycles normalized by those of the "DNN-IRM(1,3,2048)" model, respectively. The model size of "TeBGRU-StDNN(1,3,2048)" is about one-third of that of "TeBGRU-StBGRU(1,2,1024)". As for the computing cycles, "TeBGRU-StDNN(1,3,2048)" is 673 times faster than "TeBGRU-StBGRU(1,2,1024)". In summary, the "TeBGRU-StDNN(1,3,2048)" model could yield a relative WER reduction of 7.94% over "Noisy" with a much smaller model size and lower computing cycles than "TeBGRU-StBGRU(1,2,1024)".

Table V lists average WER (%) comparison of different teacher and student models for single-channel speech enhancement on the development and test sets across four environments. First, "IMCRA" and "BGRU-IRM" obtained comparable performance to "Noisy" for both simulation data (SimData) and realistic data (RealData) of both development and test sets, while "DNN-IRM" significantly degraded the recognition performance. Second, the proposed student model with ISPP as the learning targets could consistently improve the recognition accuracy compared with the conventional teacher model with IRM as the learning targets, e.g, a relative WER reduction of 20.90% from the "DNN-IRM" to "TeBGRU-StDNN," both with the setting (1,3,2048) in Table IV and a relative WER reduction of 18.95% from the "BGRU-IRM" to "TeBGRU-StBGRU" on RealData of the test set. Third, by comparing "TeBGRU-StDNN" and "TeDNN-StDNN" with the same DNN structure for the student model, we could observe that the more powerful teacher model led to a better performance of the student model due to more accurate estimation of IRM to calculate ISPP. Finally, the improvements in the proposed student model on the real set were larger than those on the simulation set, e.g., relative WER reductions of 7.94% and 4.27% from the "TeBGRU-StDNN" to "Noisy" on real and simulation test sets, respectively. Additionally, performance gains on the test set were more significant than those on the development set. This result demonstrates that our proposed approach is more effective on realistic data under adverse environments. In real applications, we can select the student model according to different priorities. For example, the compact "TeBGRU-StDNN" model could be adopted for the scenario with high demand of efficiency, while "TeBGRU-StBGRU" could be adopted for the application requiring a high recognition accuracy.

TABLE V
AVERAGE WER (%) COMPARISON OF DIFFERENT TEACHER AND STUDENT MODELS FOR SINGLE-CHANNEL SPEECH ENHANCEMENT ON THE DEVELOPMENT AND TEST SETS ACROSS FOUR ENVIRONMENTS

Enhancement	**Development Set**		**Test Set**	
	SimData	RealData	SimData	RealData
Noisy	12.98	11.57	20.84	23.56
IMCRA	13.04	11.91	21.10	23.12
DNN-IRM	15.06	13.78	23.98	27.42
BGRU-IRM	12.53	11.86	20.49	23.59
TeBGRU-StBGRU	**11.08**	**10.40**	**18.29**	**19.12**
TeDNN-StDNN	12.67	11.23	20.35	22.31
TeBGRU-StDNN	12.34	10.83	19.95	21.69

TABLE VI
AVERAGE WER (%) ON DIFFERENT ENHANCEMENTS USING TDNN-BASED ACOUSTIC MODEL AND LSTM-BASED LANGUAGE MODEL ON THE DEVELOPMENT AND TEST SETS ACROSS FOUR ENVIRONMENTS

Acoustic Model (Training Data)	Language Model	Enhancement	**Development Set** SimData	**Development Set** RealData	**Test Set** SimData	**Test Set** RealData
DNN (Channel 5)	RNN	Noisy	12.98	11.57	20.84	23.56
		IMCRA	13.04	11.91	21.10	23.12
		BGRU-IRM	12.53	11.86	20.49	23.59
		TeBGRU-StBGRU	**11.08**	**10.40**	**18.29**	**19.12**
TDNN (Channel 1-6)	RNN	Noisy	8.06	6.64	14.42	13.85
		IMCRA	11.60	8.80	16.53	17.49
		BGRU-IRM	10.76	6.90	19.60	13.67
		TeBGRU-StBGRU	**7.30**	**5.86**	**12.53**	**11.56**
TDNN (Channel 1-6)	LSTM	Noisy	6.55	5.29	12.50	12.14
		IMCRA	9.47	7.22	14.21	15.26
		BGRU-IRM	9.18	5.53	17.39	11.97
		TeBGRU-StBGRU	**5.78**	**4.47**	**10.66**	**9.81**

G. Experiments on Powerful Back-End

In this section, we will explore the ASR performance of our proposed enhancement with more powerful back-end system, including TDNN-based acoustic model and LSTM-based language model. For the acoustic model, the TDNN with LF-MMI training [49] instead of DNN with sMBR-based discriminative training [44]. The architecture of TDNN is similar to those described in [50]. For the language model, the LSTM model is trained by Kaldi-RNNLM [51] tools, and n-best re-scoring is utilized to improve the performance.

Table VI lists average WER (%) on different enhancements using TDNN-based acoustic model and LSTM-based language model on the development and test sets across four environments. First, comparing the first line of the three blocks of Table VI, more powerful back-end system using TDNN-based acoustic model and LSTM-based language model can directly improve the ASR performance for both simulation data and realistic data of both development and test sets, e.g., a relative WER reduction of 41.21% from the "Noisy" in the first block of Table VI to the "Noisy" in the second block on RealData of the test set. Second, "IMCRA" and "BGRU-IRM" destroyed the ASR performance comparing to "Noisy" when the recognition system became more powerful using TDNN-based acoustic model and LSTM-based language model showed in the second and third blocks of Table VI, while they could obtain the comparable performance to "Noisy" when the acoustic model was DNN-HMM shown in the first block of Table VI. Third, our proposed "TeBGRU-StBGRU" model could still consistently improve the ASR performance even in more powerful recognition system shown in the second and third blocks of Table VI, e.g., the relative WER reduction of 16.53% and 19.19% from "Noisy" to "TeBGRU-StBGRU" in second and third blocks of Table VI on RealData of test set.

V. CONCLUSION

In this study, a novel teacher-student learning framework is proposed, which combines unsupervised speech enhancement, e.g., IMCRA, and deep learning techniques at the training stage. The trained student model can be directly utilized for speech enhancement as the preprocessor of ASR systems in the recognition stage. By experimental analysis, we find that the regression model cannot perform well by learning the nonlinear relationship between the noisy LPS features and target IRM under adverse environments, where it cannot improve the ASR performance comparing to unprocessed noisy speech data. Conversely, the proposed student model with the ISPP as the new learning target, which is calculated with the help of the teacher model with IRM as the learning target, can improve the recognition accuracy. The experimental results on the CHiME-4 challenge show that the proposed approach yields consistent improvements over both BGRU-IRM and IMCRA-based techniques for ASR performance. In addition, the student model with DNN in causal mode having no latency achieves a relative WER reduction of 7.94% on the real test set comparing unprocessed speech with 670 times less computing cycles in processing when compared to the best student model with BGRU. Finally, the conventional speech enhancement and IRM-based deep learning method destroyed the ASR performance when the recognition system became more powerful. While our proposed approach could still improve the ASR performance even in the more powerful recognition system. For the future work, we will extend previously proposed iterative mask estimation (IME)-based multi-channel speech enhancement [52] into joint optimizing the neural network with conventional multi-channel speech enhancement algorithm inspired by the proposed single-channel speech enhancement based on teacher-student learning framework.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013.
[2] S. E. Levinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *Bell Syst. Tech. J.*, vol. 62, no. 4, pp. 1035–1074, 1983.
[3] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. Cambridge, MA, USA: MIT Press, 1964.

[4] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[5] R. McAulay and M. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[6] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[8] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[10] G. Brown and M. Cooke, "Computational auditory scene analysis," *Comput. Speech Lang.*, vol. 8, no. 4, pp. 297–336, 1994.

[11] D. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*. New York, NY, USA: Wiley-IEEE Press, 2006.

[12] J.-H. Chang, Q.-H. Jo, D. K. Kim, and N. S. Kim, "Global soft decision employing support vector machine for speech enhancement," *IEEE Signal Process. Lett.*, vol. 16, no. 1, pp. 57–60, Jan. 2009.

[13] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2008, pp. 4029–4032.

[14] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 411–414.

[15] M. N. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *Proc. IEEE Mach. Learn. Signal Process.* IEEE. 2007, pp. 431–436.

[16] G. J. Mysore and P. Smaragdis, "A non-negative approach to semi-supervised separation of speech from noise with the use of temporal dynamics," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2011, pp. 17–20.

[17] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in nonstationary noise environments," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, 2012, pp. 595–598.

[18] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[19] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[21] Y.-H. Tu, J. Du, Y. Xu, L.-R. Dai, and C.-H. Lee, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *Proc. IEEE 9th Int. Symp. Chin. Spoken Lang. Process.* IEEE, 2014, pp. 250–254.

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

[23] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[24] J. Du, Y.-H. Tu, L.-R. Dai, and C.-H. Lee, "A regression approach to single-channel speech separation via high-resolution deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1424–1437, Apr. 2016.

[25] J. Du and Q. Huo, "A speech enhancement approach using piecewise linear approximation of an explicit model of environmental distortions," in *Proc. 9th Annu. Conf. Int. Speech Commun. Assoc.*, 2008, pp. 569–572.

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

[27] B. A. Pearlmutter, "Gradient calculations for dynamic recurrent neural networks: A survey," *IEEE Trans. Neural Netw.*, vol. 6, no. 5, pp. 1212–1228, Sep. 1995.

[28] F. Weninger, F. Eyben, and B. Schuller, "Single-channel speech separation with memory-enhanced recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.* IEEE, 2014, pp. 3709–3713.

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

[30] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. IEEE, 2016, pp. 2870–2874.

[31] A. S. Subramanian, S. Chen, and S. Watanabe, "Student-teacher learning for blstm mask-based speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. IEEE, 2018, pp. 3249–3253.

[32] S. Chen, A. S. Subramanian, H. Xu, and S. Watanabe, "Building state-of-the-art distant speech recognition using the chime-4 challenge with a setup of speech enhancement baseline," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*. IEEE, 2018, pp. 1571–1575.

[33] K. Cho, B. Van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," *Empirical Methods Natural Lang. Process.*, pp. 103–111, 2014.

[34] Y.-H. Tu, I. Tashev, S. Zarar, and C.-H. Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. IEEE, 2018, pp. 2531–2535.

[35] Y.-H. Tu, J. Du, and C.-H. Lee, "DNN training based on classic gain function for single-channel speech enhancement and recognition," in *Proc. IEEE Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2019, pp. 910–914.

[36] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[38] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "SNR-based progressive learning of deep neural network for speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH*), 2016, pp. 3713–3717.

[39] 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," *Comput. Speech Lang.*, vol. 46, pp. 535–557, 2017.

[40] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "CSRI (WSJ0) complete," *Linguistic Data Consortium*, Philadelphia, PA, USA: Philadelphia, 2007.

[41] E. Vincent, R. Gribonval, and M. D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Process.*, vol. 87, no. 8, pp. 1933–1950, 2007.

[42] A. Paszke *et al.*, "Automatic differentiation in pytorch," 2017.

[43] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[44] K. Veselý, A. Ghoshal, L. Burget, and D. Povey, "Sequence-discriminative training of deep neural networks," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (IINTERSPEECH)*, vol. 2013, 2013, pp. 2345–2349.

[45] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[46] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, IEEE. 1995, pp. 181–184.

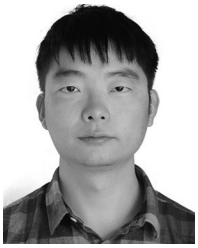
[47] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. speech Commun. Assoc.*, 2010, pp. 1045–1048.

[48] D. Povey *et al.*, "The Kaldi speech recognition toolkit," IEEE Signal Processing Society, Tech. Rep., 2011.

[49] D. Povey *et al.*, "Purely sequence-trained neural networks for asr based on lattice-free mmi," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (INTERSPEECH)*, 2016, pp. 2751–2755.

[50] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3214–3218.

[51] H. Xu *et al.*, "Neural network language modeling with letter-based features and importance sampling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*. IEEE, 2018, pp. 6109–6113.
[52] Y.-H. Tu *et al.*, "An iterative mask estimation approach to deep learning based multi-channel speech recognition," *Speech Commun.*, vol. 106, pp. 31–43, 2019.



Yan-Hui Tu received the B.S. degree from the Department of Electronic Information Engineering, Yunnan University, in 2013 and Ph.D degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), in 2019. He is currently a Post-Doctor with USTC). His current research interests include speech separation, microphone arrays, single-channel speech enhancement, and robust speech recognition.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), in 2004 and 2009, respectively. From July 2009 to June 2010, he worked with iFlytek Research on speech recognition. From July 2010 to January 2013, he worked with Microsoft Research Asia as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing, USTC.

Chin-Hui Lee (F'97) is a Professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001, he had 20 years of industrial experience, ending at Bell Laboratories, Murray Hill, NJ, USA, as a Distinguished Member of Technical Staff and the Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of ISCA. He has published more than 500 papers and holds 30 patents, and has been cited more than 34 000 times for his original contributions with an h-index of 80 on Google Scholar. He has received numerous awards, including the Bell Labs President's Gold Award in 1998. He also won SPS's 2006 Technical Achievement Award for "Exceptional Contributions to the Field of Automatic Speech Recognition." In 2012, he was invited by ICASSP to give a plenary talk on the future of speech recognition. In the same year, he was awarded the ISCA Medal in scientific achievement for pioneering and seminal contributions to the principles and practice of automatic speech and speaker recognition.