A Multi-Target SNR-Progressive Learning Approach to Regression Based Speech Enhancement

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

***Abstract*—We propose a multi-target, signal-to-noise-ratio (SNR)-progressive learning (SNR-PL) framework for regression based speech enhancement (SE). At low SNR levels, it is often not easy to directly learn the complicated regression required in SE. We therefore decompose the original SE problem of mapping noisy to clean speech features, with a large SNR gap, into a series of subproblems, each with a small SNR increment and presumably easier to learn. In our configurations, each hidden layer of the proposed regression neural network is guided to explicitly learn an intermediate target with a specified but small SNR gain. Tested on both deep neural network (DNN) and long short-term memory (LSTM) architectures, SNR-PL consistently outperforms the conventional "black box" DNN framework in terms of both objective measure superiority and network model compactness. Furthermore, with the best configured LSTM-based SNR-PL model, we often observe that the performance is easily saturated or even degraded when increasing the number of intermediate targets, due to the fact that useful information is lost in dimension reduction when involving more target layers. Accordingly, to address this information loss issue, we explore densely connected networks on top of the LSTM structure where the input and the preceding intermediate targets are concatenated together to learn the next target. Finally, to fully utilize the rich and complementary information of intermediate targets, a simple post-processing strategy is adopted to further improve the performance. Evaluated on the simulation speech data, experimental results in unseen noises cases demonstrate that the proposed approach consistently performs better than the conventional LSTM approach in terms of objective speech enhancement measures for speech intelligibility and quality. Furthermore, when evaluated on real data provided by the CHiME-4 Challenge for automatic speech recognition (ASR) of noisy microphone array speech, we show that the proposed approach with intermediate outputs can directly improve the ASR performance, while the conventional LSTM approach increases the word error rate.**

***Index Terms*—SNR-progressive learning, speech enhancement, neural network, dense structure, post-processing.**

I. Introduction

SPEECH enhancement has been an open research problem for a long time. A key goal of speech enhancement is to improve speech intelligibility and quality in the presence of noise signal. The background noise can cause the performance degradation of voice communication, speech recognition and hearing aids [1]. Numerous algorithms have been proposed over the past several decades to solve this problem. The conventional algorithms include spectral subtraction [2], Wiener filtering [3], minimum mean squared error (MMSE) estimation [4] and optimally-modified log-spectral amplitude (OM-LSA) speech estimator [5]. Spectral subtraction is one of the first algorithms proposed for noise reduction. An estimate of the clean speech spectrum can be obtained by subtracting an estimated of the noise spectrum from the noisy speech spectrum. However, the resulting enhanced speech often suffers from an annoying artifact called musical noise [6]. OM-LSA utilizes a minima controlled recursive averaging (MCRA) noise estimation [7] approach to avoid the musical noise. One limitation of the conventional speech enhancement algorithms is that they can improve the quality of speech but often cannot effectively improve the speech intelligibility [1], [8].

For learning based methods, nonnegative matrix factorization (NMF) was investigated in supervised and unsupervised manners for speech enhancement [9], [10]. The basic idea is to decompose noisy speech into bases and weight matrices for speech and noise, respectively. NMF based speech enhancement assumes that the subspaces of speech and noise are almost orthogonal with each other, which actually overlap and thus it leads to degraded performances for the estimated speech or noise components.

Recently, with the introduction of deep learning [11], speech enhancement has made great progress. The supervised deep learning approaches have been investigated from the aspects of learning targets, neural network structures, input features, etc. Xu *et al.* [12], [13] proposed a deep neural network (DNN) based regression framework to predict clean log-power spectra (LPS) features [14] from noisy LPS features. Such a mapping has the advantage that it makes no assumptions about the statistical properties of the signals, and it can also handle non-linear and highly non-stationary noises effectively.

Besides direct mapping, masking techniques were used to make classification on time-frequency (T-F) units for speech enhancement [15], [16], such as estimating the ideal binary mask (IBM) or smoothed ideal ratio mask (IRM). Speech enhancement by binary masks can be formulated as a binary classification task, where each time-frequency bin is to be classified as to whether the desired source. The hard IBM target is effective to improve speech intelligibility, but predicting the soft IRM

Manuscript received November 12, 2019; revised May 2, 2020; accepted May 11, 2020. Date of publication May 21, 2020; date of current version June 5, 2020. 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 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.)*

Yan-Hui Tu, Jun Du, and Tian Gao are with the University of Science and Technology of China, Hefei 230027, China (e-mail: tuyanhui@ustc.edu.cn; jundu@ustc.edu.cn; gtian09@mail.ustc.edu.cn).

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

Digital Object Identifier 10.1109/TASLP.2020.2996503

target is especially beneficial for improving objective speech quality. IRM is in the range of [0, 1], which can be considered as a suppression gain at each time-frequency unit. The final enhanced features are obtained as the element-wise product of estimated IRM and noisy features. In addition to the direct prediction of mask, Huang *et al.* [17], [18] investigated joint optimization of masking functions and neural networks with an extra masking layer. DNN is a very generic model and has been applied successfully for speech enhancement. Due to the sequential nature of speech, recurrent neural networks (RNNs) with long short-term memory (LSTM) [19] have been verified more suitable for speech enhancement [20]–[25]. In order to utilize the structure information of speech, convolutional neural networks (CNNs) [26] were investigated for speech enhancement in the frequency domain [27], [28] and the time domain [29]. Recently, WaveNet [30], [31] were proposed to model the raw clean audio waveforms, and this method could avoid the performance loss caused by the reconstruction of clean speech.

One key point of the deep learning approaches is the generalization capability to unseen noises, unseen speaking styles and low SNR conditions. To enhance the capability, a set of noise types and dynamic noise aware training approach were investigated [13], [32]. Kim *et al.* aimed at a fine-tuning scheme at the test stage to improve the performance of a well-trained DNN [33]. Meanwhile, multi-task learning (MTL) has also been adopted in speech enhancement. In [34], a multi-objective framework was proposed to improve the generalization capability of regression DNN. Based on MTL method, Jiang *et al.* [35] employed DNN-based speech denoising with IBM as the targets at different time-frequency scales simultaneously and collaboratively.

Focusing on the challenges of speech enhancement in low SNR conditions, a joint framework combining speech enhancement with voice activity detection (VAD) was proposed in [8], [36] to increase the speech intelligibility. In this framework, first two DNNs for speech enhancement were trained to process speech segments and non-speech segments, respectively. Then a VAD DNN was employed to integrate the results of two sub-DNNs, which could be considered as an implementation of ensemble learning. In [37], Zhang and Wang proposed a deep ensemble network for monaural speech enhancement. They used multi-context networks to integrate temporal information at different resolutions. Multiple modules were stacked to construct an ensemble, each performing multi-context masking or mapping.

Similar to ensemble learning, another notable machine learning strategy is the curriculum learning [38] originated from cognitive science. The basic idea is to start small, learn easier aspects of the task or easier sub-tasks, and then gradually increase the difficulty level. Curriculum learning is related to MTL where the initial tasks are boosted to guide the learner for the better achievement on the final task. However the motivation of MTL is to improve the generalization of the target task by leveraging on other tasks.

In this study, inspired by curriculum learning, we propose a novel SNR-progressive learning (PL) framework to improve the speech intelligibility of neural network based speech enhancement especially in low SNR environments. The whole training process is decomposed into multiple sub-training stages with different training targets corresponding to different SNRs. At each sub-training stage, the new sub-network consists of an input layer, a hidden layer and an output layer. The input of the new sub-network are the training target in the preceding layer, and the output is the training target in the current layer with a specific SNR gain. The subproblem solving in each stage can boost the subsequent learning of the next stage. We apply PL to two commonly used neural networks DNN and LSTM for speech enhancement, namely DNN-PL and LSTM-PL. It is observed that PL consistently outperforms the conventional "black box" framework in terms of both objective measure and compact model design. Furthermore, on the best configured LSTM-PL model, we observe that the performance is easily saturated or even degraded by increasing the number of intermediate targets, which can be explained as that the useful information is lost due to the dimension reduction when involving more target layers. Accordingly, we explore the densely connected structure on top of LSTM-PL where the input and the preceding intermediate targets are spliced together to learn the next target, which can alleviate the information loss problem in original PL structure. Finally, to fully utilize the rich and complementary information of intermediate targets, a simple post-processing strategy is adopted to further improve the performance. Evaluated on WSJ0 corpus with read speech, experimental results on unseen noises demonstrate that the proposed approach can consistently and significantly outperform the conventional LSTM approach in terms of objective measure for speech intelligibility, and also yield remarkable gains for other measures. Moreover, we design a highly mismatched test set involving the AMI corpus with conversational speech, where the conventional LSTM approach can even generate worse speech intelligibility performance than the unprocessed noisy speech. Our approach shows the strong and stable generalization ability.

This work is extended from our previously and recently disclosed versions [39], [40] with the new contributions as follows. First, DNN-PL in [39] and LSTM-PL [40] are unified into a general neural network framework and compared theoretically and experimentally. Second, the motivations of SNR-progressive learning, dense structure, and post-processing are elaborated in more technical detail. Moreover, a compact version for densely connected progress learning is newly proposed. Third, more experimental analyses on why PL leads to good performances and compact network structures are given. Finally, we design a new set of experiments on a highly mismatched test set to show the strong generalization capability of the proposed approach.

II. Multi-Target SNR-Progressive Learning

In this following, we describe neural network based SNR-progressive learning in detail. As we all know, the purpose of speech enhancement is to estimate clean speech signals from the observed noisy speech signals. Specific to the deep learning based algorithms, neural networks can be adopted to implement this process [12]. As shown on the left of Fig. 1, neural networks are usually guided to map the input noisy speech to the output

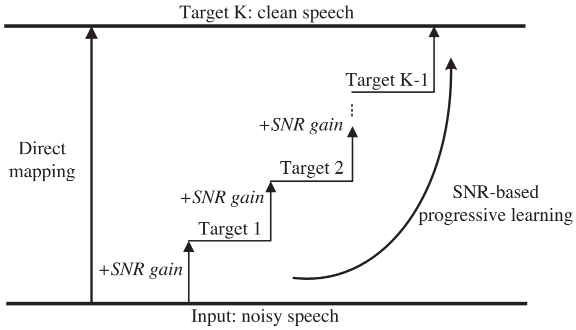


Fig. 1. An illustration of SNR-progressive learning.

clean speech, which is denoted as direct mapping. However, the direct mapping often leads to performance degradation in low SNR conditions [8] as the relationship between the high-dimensional input and output speech features are quite complicated to be learned as a black box. To address this issue, we propose a novel neural network based SNR-progressive learning as shown on the right of Fig. 1. The basic idea is to start small, learn easier aspects of the task or easier sub-tasks, and then gradually increase the difficulty level. Specific to neural network training, the direct mapping process from noisy speech to clean speech is decomposed into multiple stages with a specific SNR gain achieved in each stage. The SNR gains in each stage can boost the subsequent learning of the next stage. For example, if the input SNR of noisy speech is 0 dB, the learning target of the direct mapping system is clean speech (infinity dB). As for SNR-progressive learning, the intermediate learning targets with higher SNR (e.g., 10 dB or 20 dB) will be inserted as new layers. Meanwhile, the progressive concept has also been investigated by other researchers on reinforcement learning tasks [41], where the progressive networks retain a pool of pre-trained models throughout training and learn lateral connections from these to extract useful features for the new task. Unlike the lateral connection used in [41], SNR-progressive learning for speech enhancement is applied in a straightforward manner. The features from the preceding networks have clear definitions, and are conveyed to the next task via vertical connection.

For the implementation of SNR-progressive learning, a general neural network architecture is illustrated in Fig. 2. The activation function is linear in the target layers and non-linear in the other hidden layers (e.g., DNN or LSTM layers). All the target layers are designed to learn intermediate speech with higher SNRs (from Target 1 to Target $K-1$) or clean speech (Target K). This stacking-style neural network can learn multiple targets progressively and efficiently. In our previous work [39], we have applied SNR-progressive learning on DNN architecture with fully connected hidden layers successfully. Experimental results demonstrated that SNR-progressive learning could effectively improve speech intelligibility especially in low SNR environments. Nonetheless, the DNN-PL in [39] only considers the frame expansion in the input layer while all the target layers just use one central frame. The fully connected architecture can not well utilize the important temporal information from the intermediate targets.

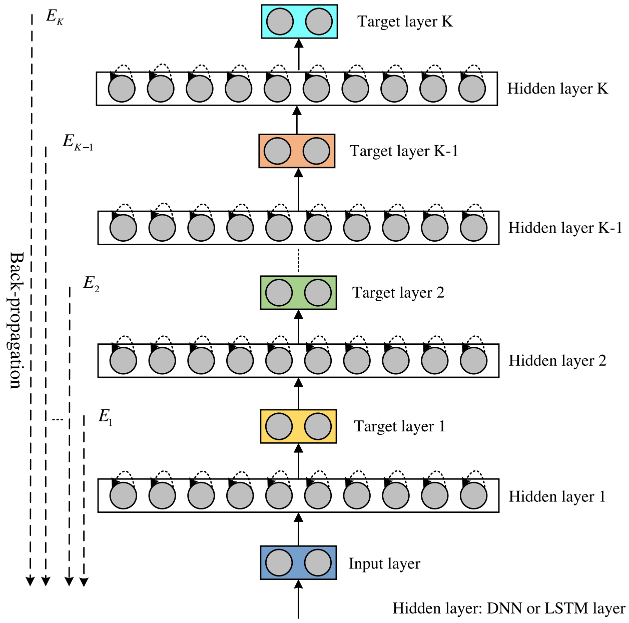


Fig. 2. Neural network based SNR-progressive learning for speech enhancement.

First, the learning target of intermediate layer k in the waveform domain can be expressed as:

$$x_k(t) = s(t) + n_k(t) = s(t) + g_k n_0(t) \tag{1}$$

where $s(t)$ is the t-th sample of clean speech in time domain. $n_k(t)$ and $x_k(t)$ are the t-th sample of noise signal and reference target speech in the time domain for k^{th} target layer, respectively ($k > 0$). g_k is an utterance-level coefficient to control the SNR for k-th target layer while $n_0(t)$ is the noise signal for input layer. We can define that $\mathbf{x}_n^k$ is the n^{th} D-dimensional LPS feature vector of learning target in intermediate layer k extracting from $x_k(t)$ in time domain.

Then, as for optimizing the parameters in Fig. 2, an MTL-based weighted MMSE criterion with K target layers is designed to update the randomly initialized parameters.

$$E^{\text{PL}} = \sum_{k=1}^{K} \alpha_k E_k^{\text{PL}} \tag{2}$$

$$E_k^{\text{PL}} = \frac{1}{N}\sum_{n=1}^{N} \|\mathcal{F}_k^{\text{PL}}(\hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\text{PL}}) - \mathbf{x}_n^k\|_2^2 \tag{3}$$

where α_k is the weighting factor of objective function for k^{th} target layer. $\mathcal{F}_k^{\text{PL}}(\hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\text{PL}})$ is the neural network function for k^{th} target using the previously learned intermediate target $\hat{\mathbf{x}}_n^{k-1}$, and $\mathbf{\Lambda}_k^{\text{PL}}$ represents the parameter set of the weight matrices and bias vectors before k^{th} target layer, which are optimized in the manner of BP (for DNN) or BPTT (for LSTM) with stochastic gradient descent, with N representing the mini-batch size. The output of target layer k is expressed in a nested way as:

$$\begin{aligned}\hat{\mathbf{x}}_n^k &= \mathcal{F}_k^{\text{PL}}(\hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\text{PL}})\\ &= \mathcal{F}_k^{\text{PL}}(\mathcal{F}_{k-1}^{\text{PL}}(...\mathcal{F}_1^{\text{PL}}(\hat{\mathbf{x}}_n^0, \mathbf{\Lambda}_1^{\text{PL}}), ...\mathbf{\Lambda}_{k-1}^{\text{PL}}), \mathbf{\Lambda}_k^{\text{PL}})\end{aligned} \tag{4}$$

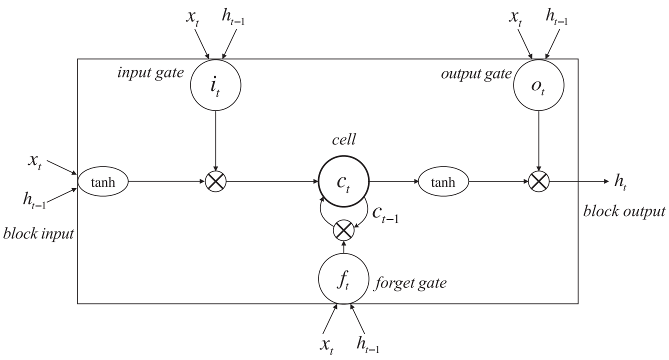


Fig. 3. An illustration of the LSTM block.

To improve the model capability of the speech feature sequence, LSTM [19], [42] seems to have a natural advantage to capture the long-term contextual information by using recursive structures between the previous frames and the current frame. The central idea of LSTM is to introduce a cell state variable alongside the RNN hidden activation which contains a series of gates to dynamically control the information flow. Fig. 3 illustrates a single LSTM memory block.

When SNR-progressive learning is introduced to LSTM case, the whole network architecture for speech enhancement is also illustrated in Fig. 2. For the input and multiple targets, LSTM layers with the dotted arrow on the units are used to link between each other. SNR-progressive learning and LSTM network seem like a perfect match because the sub-network from the Input layer to Target layer 1 and the sub-networks from the intermediate target to the next target have the same recursive structure which can fully utilize the temporal information from the input and intermediate learning targets by LSTM layers automatically.

III. Improvements to LSTM-Based SNR-PL

Through the above analysis, we believe LSTM is more suitable than DNN for SNR-progressive learning. In the following we focus our discussion only on improvements of SNR-progressive learning over the LSTM-PL architecture. Although the preliminary experiments show that LSTM-PL is superior to the conventional LSTM in terms of both objective measures and compact design, we observe that the performance of LSTM-PL architecture is easily saturated and even degraded by increasing the number of intermediate targets, which can be explained as that the useful information is lost due to the dimension reduction (the dimension of target layers is much smaller than that of hidden layers) when involving more target layers. Accordingly we present the densely connected SNR-progressive learning where the input and the estimations of intermediate targets are spliced together to learn the next target for making full use of the rich set of information from the multiple learning targets. Furthermore, we also propose post-processing of multiple targets to leverage upon the set of enhanced signals in the enhancement stage.

A. Densely Connected SNR-PL

Recently, the densely connected structure has also been investigated in convolutional network architecture, namely dense convolutional network (DenseNet) [43], which has shown excellent results on image classification tasks. It introduces direct connections between any two layers with the same feature-map size. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. Takahashi *et al.* [44] has extended the DenseNet to tackle the audio source separation problem by introducing multi-scale DenseNet with block skip connection and transposed convolution, and applying it to each frequency band. Such a dense connectivity enables all layers to receive the gradient directly and to reuse features computed in the preceding layers.

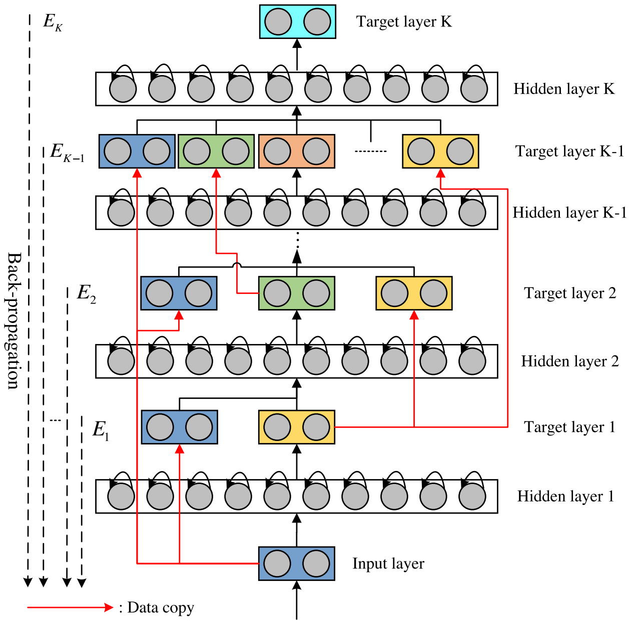


Fig. 4. LSTM-based densely connected SNR-progressive learning.

Different from the CNN-based DenseNet descried above, the proposed densely connected SNR-progressive learning structure in this study is implemented on the LSTM network and the features computed in the preceding layers have a definite physical meaning as illustrated in Fig. 4. For instance, in the learning process of Target K, the input and the estimates of Target 1, Target 2, ..., Target $K-1$ are concatenated together and then passed up to the next sub-network which can simultaneously see many forms of expressions for input and enhanced speech. In this way, the information loss problem discussed above can presumably be alleviated.

Since multiple outputs are estimated in densely connected SNR-PL, a weighted MMSE criterion in terms of MTL with K target layers is also designed to optimize all network parameters randomly initialized as follows:

$$E^{\mathrm{PLD}} = \sum_{k=1}^{K} \alpha_k E_k^{\mathrm{PLD}} \tag{5}$$

$$E_k^{\mathrm{PLD}} = \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_k^{\mathrm{PLD}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, \ldots, \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\mathrm{PLD}}) - \mathbf{x}_n^k \|_2^2 \tag{6}$$

where $\hat{\mathbf{x}}_n^k$ and $\mathbf{x}_n^k$ are the n^{th} D-dimensional vectors of estimated and reference target LPS feature vectors for k^{th} target layer, respectively ($k > 0$), with N representing the mini-batch size. $\hat{\mathbf{x}}_n^0$ denotes the n^{th} D-dimensional vector of input noisy LPS features. $\mathcal{F}_k^{\text{PLD}}(\hat{\mathbf{x}}_n^0, \hat{\mathbf{x}}_n^1, \ldots, \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\text{PLD}})$ is the neural network function for k^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\mathbf{x}}_n^0$ to $\hat{\mathbf{x}}_n^{k-1}$, and $\mathbf{\Lambda}_k^{\text{PLD}}$ represents the parameter set of the weight matrices and bias vectors before k^{th} target layer, which are optimized in the manner of BPTT with gradient descent.

The experiments demonstrate that the proposed dense structures indeed improve the performance of LSTM-PL by using more intermediate targets. But the computational complexity and the model size will be also dramatically increased with a large K. To design a compact version for densely connected SNR-PL, we only concatenate the two latest intermediate targets to learn the next target. The corresponding weighted MMSE criterion is defined as:

$$E^{\text{PLDC}} = \sum_{k=1}^{K} \alpha_k E_k^{\text{PLDC}} \tag{7}$$

$$E_k^{\text{PLDC}} = \frac{1}{N} \sum_{n=1}^{N} \| \mathcal{F}_k^{\text{PLDC}}(\hat{\mathbf{x}}_n^{k-2}, \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\text{PLDC}}) - \mathbf{x}_n^k \|_2^2 \tag{8}$$

where the main difference from Eq. (5) and Eq. (6) is that the network function $\mathcal{F}_k^{\text{PLDC}}(\hat{\mathbf{x}}_n^{k-2}, \hat{\mathbf{x}}_n^{k-1}, \mathbf{\Lambda}_k^{\text{PLDC}})$ is only dependent on two targets $\hat{\mathbf{x}}_n^{k-2}$ and $\hat{\mathbf{x}}_n^{k-1}$. For $k = 1$, $\mathcal{F}_k^{\text{PLDC}}$ refers to $\mathcal{F}_1^{\text{PLDC}}(\hat{\mathbf{x}}_n^0, \mathbf{\Lambda}_1^{\text{PLDC}})$.

B. *Why SNR-Progressive Learning?*

Recently, the ResNet [45] which aims to address the degradation problem with the network depth increasing is widely applied. And the authors also demonstrate that it is easier to optimize the residual mapping than to optimize the original mapping. And more detailed description about it can be found in [45]–[47]. Based on the above analysis, the whole optimized process can be seen as implicit learning. The intermediate targets just only represent residuals and have no physical meanings. The advantage of SNR-progressive learning is that multiple intermediate targets with different SNRs are provided. We examine that whether the SNR-PL network can achieve the original goal in Fig. 1, namely generating the SNR gain from each hidden layer. In Fig. 5, the average SNR gain with the variance generated from each hidden layer across the utterances of the cross validation (CV) set and the test set at −5 dB input SNR. 100 utterances with the seen noise types of the training stage are randomly selected from the CV set while another 100 utterances with unseen noise types (factory and white noises) are also randomly selected from the test set. The densely connected LSTM-PL network with $K = 5$ target/hidden layers listed in Table I is used. Accordingly 5 sets of SNR gains are generated from 5 hidden/LSTM layers as shown in Fig. 5. For both CV and test sets, the generated SNR gains in the enhancement stage are closer to the oracle ones (5 dB) setting in the training stage when the hidden/target layers are closer to the input layer. For example, the average SNR gain from the hidden layer 1 (H1) on the CV set is much larger than those from other hidden layers and the corresponding variance of the SNR gains is much smaller which indicates the stability across different utterances. Furthermore, the SNR gains on the test set are consistently smaller than those on the CV set, which is reasonable as there is the generalization issue for the unseen noise types of the test set. Overall, the LSTM-PL network can indeed gradually improve the SNR of input noisy speech as we intended in our design. And the intermediate targets can be utilized for improving the speech enhancement and speech recognition performance.

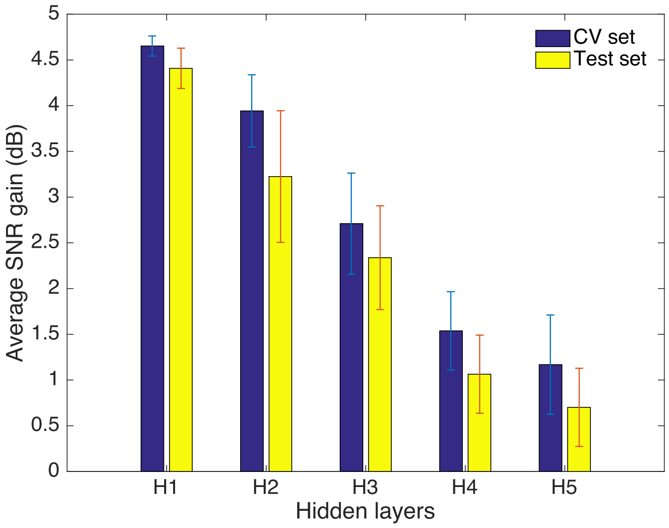


Fig. 5. The average SNR gain with the variance generated from each hidden layer across all the utterances of the cross validation (CV) set and the test set (with factory and white noises) at −5 dB input SNR. The densely connected LSTM-PL network with $K = 5$ target/hidden layers listed in Table I is used.

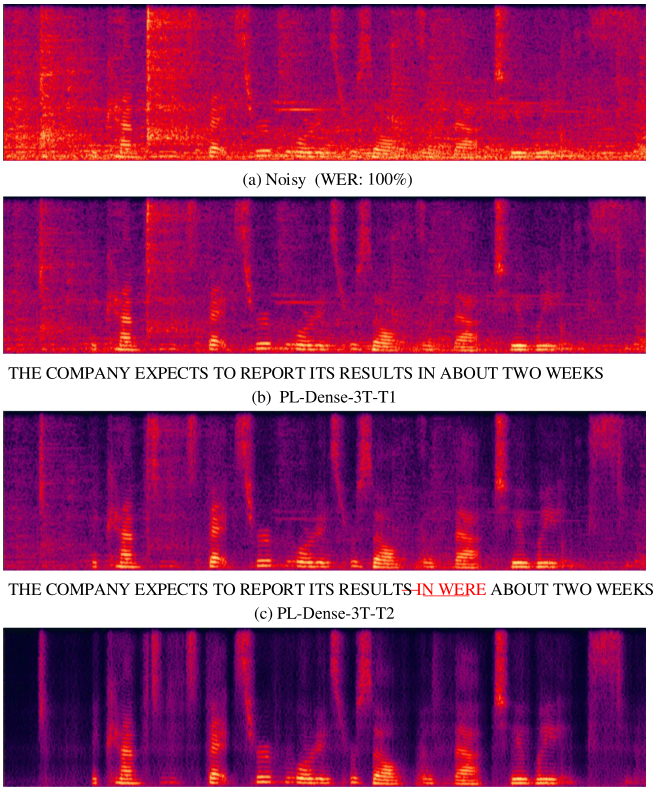
TABLE I
TARGET SNR GAIN CONFIGURATIONS OF SNR-PL SYSTEMS WITH DIFFERENT LEARNING TARGETS

K	SNR Gain for the intermediate target
2	10dB (Target 1)
3	10dB (Target 1-2)
5	5dB (Target 1-4)
7	2.5dB (Target 1-4), 5dB (Target 5-6)

C. *Post-Processing*

In this section, we aim at further improving the performance via post-processing of multiple targets with rich information in the enhancement stage. On the test set we have more interesting observations, especially for the low SNR cases. Although the final target K has the highest SNR, the aggressive noise reduction goal often leads to more speech distortions. Other targets close to target K with relatively lower SNRs can achieve better speech preservation. This complementarity motivates us to design a simple post-processing strategy to fully utilize the rich information of multiple learning targets as follows:

$$\hat{\mathbf{x}}_n = \begin{cases} (\hat{\mathbf{x}}_n^K + \hat{\mathbf{x}}_n^{K-1})/2, & K = 2 \\ (\hat{\mathbf{x}}_n^K + \hat{\mathbf{x}}_n^{K-1} + \hat{\mathbf{x}}_n^{K-2})/3, & K > 2 \end{cases} \tag{9}$$



(a) Noisy (WER: 100%)

THE COMPANY EXPECTS TO REPORT ITS RESULTS IN ABOUT TWO WEEKS
(b) PL-Dense-3T-T1

THE COMPANY EXPECTS TO REPORT ITS RESULTS IN WERE ABOUT TWO WEEKS
(c) PL-Dense-3T-T2

THE COMPANY EXPECTS TO REPORT ITS RESULTS IN WERE ABOUT TWO WEEKS
(d) PL-Dense-3T-T3

Fig. 6. Spectrograms with the recognition results of an utterance from the real test set of CHiME-4 challenge with different target outputs.

where $\hat{\mathbf{x}}_n^k (k = K-2, K-1, K)$ are the n^{th} D-dimensional LPS feature vectors of the top three target layers and $\hat{\mathbf{x}}_n$ is the post-processing result. The reason why we select the top three results is they have sufficiently high SNRs according to the average SNR gains (less than 1 dB) of H4 and H5 on the test set in Fig. 5. Finally a good tradeoff of noise reduction and speech preservation can be made.

D. Application to Robust ASR

Based on the description of Section III-C, the intermediate targets can not eliminate the background noise completely, but it can achieve better speech preservation comparing to the final target K. So for the different intermediate targets, they can achieve different tradeoffs between the noise suppression and speech preservation. For the popular recognition system under multi-condition training, it is robust to noise in relative high-SNR cases. So it is more important to feed the recognition system with less distorted speech from the enhancement model.

Fig. 6 plots the spectrograms with recognition results of an utterance from the real test set of CHiME-4 challenge with different target outputs. Fig. 6(a) presents the recognition of original noisy speech. The recognition error rate of noisy is 100% because of the real background noise. The noise reduction effectiveness of SNR-PL relies on the different target layer outputs, as shown in Fig. 6(b)–(d). Although the output of Target layer 1 (PL-Dense-3T-T1) can not eliminate much noise, it can better preserve target speech comparing to Fig. 6(c) and (d). Both high-level background noise of noisy speech and speech distortions introduced by enhancement algorithm can lead to recognition error, shown in Fig. 6(a), and Fig. 6(c)–(d), respectively. It seem that the key point of enhancement algorithm for ASR is to find a best trade-off between noise suppression and speech preservation, shown in Fig. 6(b) yielding a totally correct recognition result.

IV. Experiment on Speech Enhancement

We demonstrate the effectiveness of SNR-progressive learning with dense structure and post-processing based on speech enhancement metrics, short-time objective intelligibility (STOI, in %) [48], perceptual evaluation of speech quality (PESQ) [49] and source-to-distortion ratio (SDR, in dB) [50].

A. Implementation Details

115 noise types used in [39] were chosen as our noise database. Clean speech was derived from the WSJ0 corpus [51]. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, denoted as SI-84 training set, are corrupted with the above-mentioned 115 noise types at three SNR levels (−5 dB, 0 dB and 5 dB) to build a 36-hour training set, consisting of pairs of clean and noisy utterances. The 330 utterances from 12 other speakers, namely the Nov92 WSJ evaluation set, were used to construct the test set for each combination of noise types and SNR levels (−5 dB, 0 dB, 5 dB, 10 dB). Five unseen noises from the NOISEX-92 corpus [52], namely babble, factory (factory1, factory2), destroyer engine, m109 and white noises were adopted for testing.

As for the front-end, 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 256 samples. A short-time Fourier analysis is adopted to compute the DFT of each overlapping windowed frame. Then the 257-dimensional LPS features normalized by global mean and variance are employed to train neural networks. The results in the study are based on using MSE loss between enhanced and clean signal log-power-spectra and potentially the results may be different using other loss functions such as magnitude spectrogram domain losses. 2048 hidden nodes are used for each DNN layer while 1024 cells are used for each LSTM layer. For the ResNet, one LSTM layer with 1024 cells is used to connection a residual connections layer and the number of the hidden layers is 5.

For SNR-PL systems, one LSTM or DNN layer is used to connect each pair of the input/target or target/target layers. The parameter α_k in Eqs. (2), (5) and (7) is set as follows: $\alpha_K = 1.0; \alpha_k = 0.1, (k = 1, \ldots, K-1)$. For configuring the target layers and SNR gains, we investigate several systems as shown in Table I in this study. The Microsoft Computational Network Toolkit (CNTK) [53] is employed for neural network training.

B. DNN-PL vs. LSTM-PL

Table II gives the average STOI comparison of DNN/LSTM (all with 3 hidden layers) systems on the test set across five

TABLE II
THE AVERAGE STOI COMPARISON OF DNN/LSTM (ALL WITH 3 HIDDEN LAYERS) SYSTEMS ON THE TEST SET ACROSS FIVE UNSEEN NOISES. N_M IS THE MODEL SIZE NORMALIZED BY DNN BASELINE SYSTEM

System	N_M	-5dB	0dB	5dB	10dB
Noisy	-	64.7	76.5	86.7	93.2
DNN Baseline	1	65.5	79.2	87.6	91.9
DNN-PL	0.5	68.0	80.4	87.8	91.7
LSTM Baseline	1.8	67.3	81.1	89.0	93.1
LSTM-PL	1.3	69.2	82.0	89.4	93.3

unseen noises at different SNR levels. Noisy and DNN/LSTM Baseline represent the systems of unprocessed noisy speech and the conventional DNN/LSTM for speech enhancement, respectively. DNN-PL and LSTM-PL are the systems of DNN and LSTM based SNR-PL, respectively. From Table II, several observations could be made. First, DNN-PL improved STOI effectively at −5 dB and 0 dB when compared with DNN Baseline. However, both the DNN Baseline and DNN-PL underperformed the unprocessed system at relatively high SNR, e.g., 10 dB. When LSTM is employed, the degradation at 10 dB has been reversed. LSTM Baseline consistently yielded the average STOI gain of more than 1 over DNN Baseline for all SNRs. On top of the LSTM Baseline, LSTM-PL still obtained remarkable gains especially at low SNRs. These results validated our assumption in Section II: LSTM is more suitable than DNN for SNR-PL. The main reason is that the temporal information in DNN-PL training is only considered via frame expansion in the input layer but the important temporal structure of the intermediate learning targets cannot be well utilized. In the following experiments, SNR-PL is implemented with LSTM by default.

Based on Table II, SNR-PL not only improves the STOI performance, but also achieves the compact model design for both DNN and LSTM networks. For example, the model size of DNN-PL is about 50% of the conventional DNN model. To give the reader a better understanding, we list the comparison of weight distributions between DNN Baseline and DNN-PL as shown in Fig. 7. The weights linking the hidden layers (H1 to H3) and succeeding target layers are used. We can observe that the weight distribution of the first hidden layer (H1) closest to the input layer was relatively smooth while the weight distributions of both H2 and H3 had a sharp peak near the zero value for the conventional DNN. Compared with the conventional DNN, the weight distributions in DNN-PL were more balance across different hidden layers, which can be explained as each hidden layer of SNR-PL network is designed with an explicit learning objective, namely generating a specific SNR gain, and the scales of all the hidden layers are comparable.

C. Dense Structure and Post-Processing

To further improve the speech intelligibility performance of LSTM-PL, Fig. 8 shows the average STOI comparison of SNR-PL, densely connected SNR-progressive learning (PL+Dense) and PL+Dense with post-processing (PL+Dense+PP) along with different learning targets at −5 dB, 0 dB, 5 dB and 10 dB. It should be noted that, when the number of learning targets is 1, it refers to the conventional LSTM system with two hidden layers. The configurations of other learning targets are listed in Table I. First, by focusing on the blue line, we found that PL achieved significant STOI improvements from LSTM Baseline to PL with two learning targets. However, using more learning targets did not make additional gains and even led to the performance degradation, which might be due to the dimensional reduction and information loss when involving more target layers.

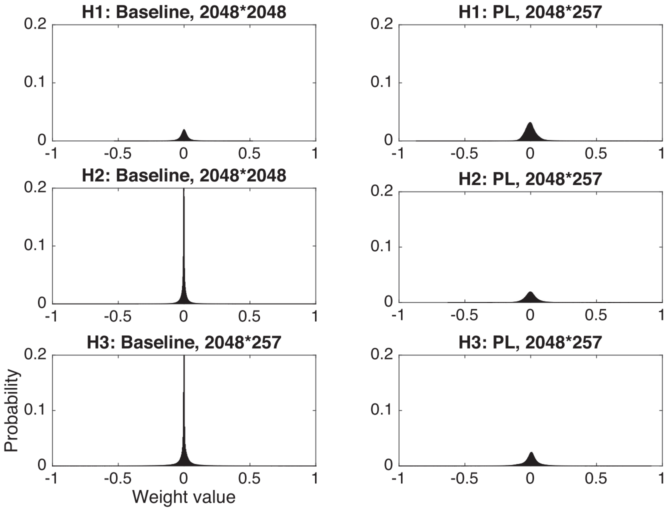
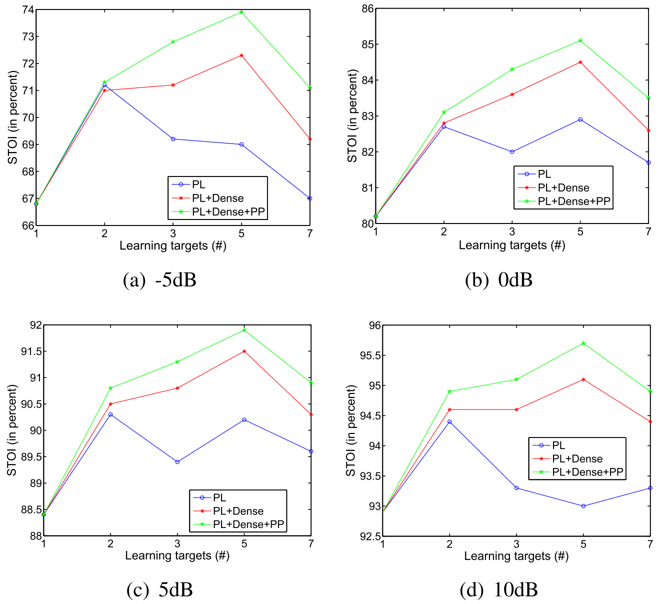


Fig. 7. The comparison of weight distributions between DNN Baseline (left column) and DNN-PL (right column) as shown in Table II. The weights linking the hidden layers (H1 to H3) and succeeding target layers are used.



(a) -5dB (b) 0dB

(c) 5dB (d) 10dB

Fig. 8. The average STOI comparison of SNR-PL systems on the test set along with different learning targets across five unseen noises at −5 dB, 0 dB, 5 dB and 10 dB.

Then by employing the dense structure (the red line in Fig. 8), PL+Dense had a different performance trend. To give a reasonable explanation, the analysis could be made based on Fig. 5 and the corresponding discussion in Section III-C. For the SNR-PL network, the intermediate targets were with notable pros and cons. The target close to the input layer had a lower SNR corresponding to good speech preservation and insufficient noise reduction while the target close to output layer had a higher SNR yielding good noise reduction and large speech distortions. Accordingly, the dense structure could make full use of the rich and complementarity information of different learning targets to enhance the robustness of the model. The best configuration for the target number in PL+Dense was 5 which was larger than that in PL implying that dense structure can accommodate more learning targets for SNR-PL. This setting is also used in the subsequent experiments by default. When using seven learning targets, the STOI performance was sharply degraded which might be explained as the over-fitting problem especially by the top high-dimensional target layer by concatenating many preceding targets. Finally, PL+Dense+PP denoted by the green line followed a similar performance trend as the PL+Dense. More learning targets produced more information available for post-processing. When the number of learning targets was larger than two, PP can further bring significant STOI gains to improve the speech intelligibility.

Table III lists the average STOI/SDR/PESQ results of different systems on the test set across five unseen noises at -5 dB, 0 dB, 5 dB and 10 dB. First, with the increase of the number of hidden layers, the performance of all three measures for LSTM Baseline seemed to be saturated at the setting of four hidden layers when the input SNR was above 0 dB. Second, PL system with a slightly smaller model size than LSTM Baseline with four hidden layers could obtain better STOI and SDR results when the input SNR was below 5 dB. But the observation on PESQ measure between them was mixed for different SNRs and there was no performance gain for both STOI and SDR at 10 dB input SNR. Furthermore, the PL+Dense using the dense structure yielded significant gains over PL, e.g., the improvements of 3.3 STOI and 0.55 dB SDR at -5 dB. More interestingly, for the 10 dB SNR case, STOI could be improved from 93.0 to 95.1 while SDR could be improved from 10.64 dB to 13.24 dB. So PL+Dense was quite robust for both low and high SNRs unlike PL and LSTM Baseline which still underperformed the unprocessed system at 10 dB in terms of STOI measure. Finally, PL+Dense+PP using post-processing generated consistently additional gains over PL+Dense for STOI and SDR measures and achieved the best overall performance for all three measures. And PP was more effective for low SNRs, e.g., 1.6 STOI gain and 0.1 PESQ gain over PL+Dense at -5 dB.

TABLE III
THE AVERAGE STOI/SDR/PESQ COMPARISON OF DIFFERENT SYSTEMS ON THE TEST SET ACROSS FIVE UNSEEN NOISES. N_{L} REPRESENTS THE NUMBER OF HIDDEN/LSTM LAYERS. N_{M} IS THE MODEL SIZE NORMALIZED BY LSTM BASELINE SYSTEM WITH TWO HIDDEN LAYERS

STOI						
System	N_{L}	N_{M}	-5dB	0dB	5dB	10dB
Noisy	-	-	64.7	76.5	86.7	93.2
LSTM Baseline	2	1	66.8	80.2	88.4	92.9
	3	1.6	67.3	81.1	89.0	93.1
	4	2.2	67.8	81.1	88.9	93.0
PL	5	2.0	69.0	82.9	90.2	93.0
PL+Dense	5	2.7	72.3	84.5	91.5	95.1
PL+Dense+PP	5	2.7	**73.9**	**85.1**	**91.9**	**95.7**

SDR						
System	N_{L}	N_{M}	-5dB	0dB	5dB	10dB
Noisy	-	-	-6.26	-1.32	3.66	8.65
LSTM Baseline	2	1	2.01	5.76	8.59	10.66
	3	1.6	2.32	6.05	8.81	10.80
	4	2.2	2.31	6.00	8.71	10.67
PL	5	2.0	2.69	6.56	9.24	10.64
PL+Dense	5	2.7	3.24	7.42	10.66	13.24
PL+Dense+PP	5	2.7	**3.52**	**7.95**	**11.57**	**14.81**

PESQ						
System	N_{L}	N_{M}	-5dB	0dB	5dB	10dB
Noisy	-	-	1.42	1.70	2.01	2.35
LSTM Baseline	2	1	1.68	2.23	2.67	3.00
	3	1.6	1.67	2.26	2.71	3.04
	4	2.2	1.77	2.27	2.62	3.01
PL	5	2.0	1.70	2.32	2.75	3.06
PL+Dense	5	2.7	1.76	2.33	**2.77**	3.10
PL+Dense+PP	5	2.7	**1.86**	**2.33**	2.71	**3.13**

D. Experiment on Generalization Capability

In the above experiments, we demonstrate the effectiveness of SNR-PL with dense structure and post-processing for unseen noises. However, the speaking style of test speech data is similar to that of the training set, namely the read style. To simulate more realistic case and further verify the generalization ability of the proposed method in speech part, we design a highly mismatched test set with both unseen speaking styles and unseen five noise types. The clean testing speech is conversational and from the AMI corpus [54] which consists more than one hundred meetings of 4-5 participants. The conversations are recorded in parallel on multiple devices, including a tabletop array of 8 microphones and head-mounted microphones for each meeting participant. The head-mounted microphone recorded speeches are noiseless and 200 high quality speech clips are captured to form our clean testing speech. Each speech clip is about 3-5 seconds long and then mixed with five unseen noises at -5 dB, 0 dB, 5 dB and 10 dB to construct the highly mismatched test set.

Table IV shows the average STOI/SDR/PESQ results of different systems on the highly mismatched test set across five unseen noise types at -5 dB, 0 dB, 5 dB and 10 dB. One difference from Table III was that LSTM Baseline system could not improve the STOI measure on the highly mismatched test set especially for low SNRs. This observation indicates that the generalization ability of LSTM Baseline in highly mismatched case is not good. The other difference from Table III was that the popular network of ResNet was adopted for comparison. We could find that ResNet could only improve the STOI measure in high SNRs cases slightly, e.g., 1.1 STOI gain at 0 dB and 1.7 STOI gain at 5 dB comparing to Noisy.

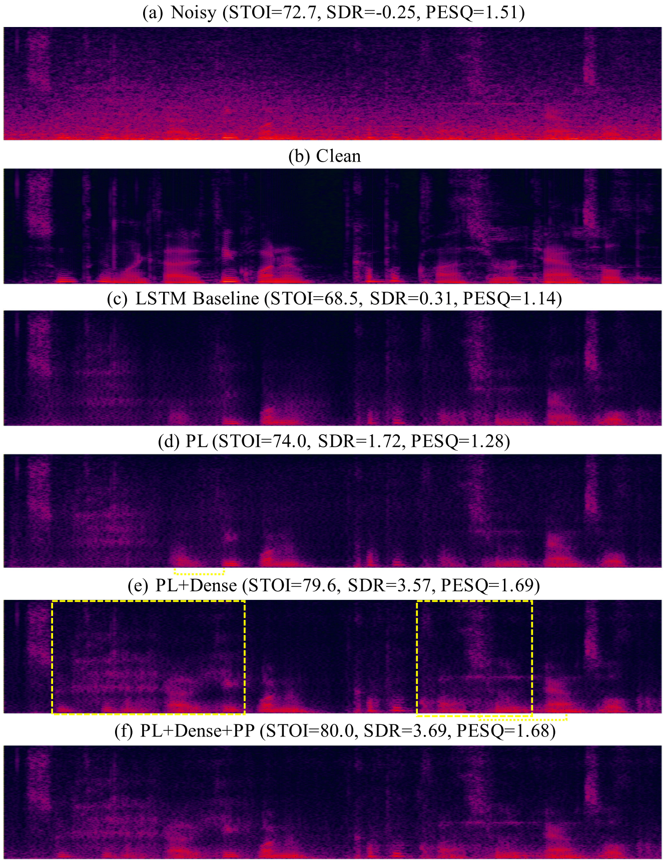
TABLE IV
THE AVERAGE STOI/SDR/PESQ COMPARISON OF DIFFERENT SYSTEMS ON THE HIGHLY MISMATCHED TEST SET ACROSS FIVE UNSEEN NOISES. N_{L} REPRESENTS THE NUMBER OF HIDDEN/LSTM LAYERS. N_{M} IS THE MODEL SIZE NORMALIZED BY 4-LAYER LSTM BASELINE SYSTEM

STOI						
System	N_{L}	N_{M}	-5dB	0dB	5dB	10dB
Noisy	-	-	52.1	63.3	74.4	83.7
LSTM Baseline	4	1	49.3	63.0	74.7	83.1
ResNet	5	0.9	50.2	64.4	76.1	84.1
PL	5	0.9	51.0	66.0	78.1	85.3
PL+Dense	5	1.2	53.4	67.5	78.2	85.9
PL+Dense+PP	5	1.2	**56.2**	**69.3**	**79.5**	**86.7**

SDR						
System	N_{L}	N_{M}	-5dB	0dB	5dB	10dB
Noisy	-	-	-5.67	-0.72	4.26	9.26
LSTM Baseline	4	1	-0.04	4.22	7.25	9.50
ResNet	5	0.9	0.30	4.52	7.58	9.81
PL	5	0.9	0.43	4.99	8.10	10.07
PL+Dense	5	1.2	1.41	5.98	9.29	11.92
PL+Dense+PP	5	1.2	**1.47**	**6.19**	**9.67**	**12.44**

PESQ						
System	N_{L}	N_{M}	-5dB	0dB	5dB	10dB
Noisy	-	-	1.48	1.64	1.86	2.14
LSTM Baseline	4	1	1.38	1.78	2.21	2.56
ResNet	5	0.9	1.35	1.78	2.21	2.57
PL	5	0.9	1.38	1.78	2.22	2.58
PL+Dense	5	1.2	1.35	1.80	2.22	2.59
PL+Dense+PP	5	1.2	**1.50**	**1.89**	**2.28**	**2.62**

When SNR-PL was implemented, the STOI measure was significantly improved over LSTM Baseline for all SNRs, e.g., 3.0 STOI gain at 0 dB and 3.4 STOI gain at 5 dB. Besides STOI, PL also achieved improvements on the SDR measure. Clearly PL has stronger ability of generation in comparison with LSTM Baseline. However, PL still underperformed the unprocessed noisy speech in terms of STOI at -5 dB. By using dense structure, the further improvements of STOI were yielded by PL+Dense especially for -5 dB case (2.4 STOI gain). The best results for all measures were consistently achieved by PL+Dense+PP. One interesting observation was post-processing in Table IV could bring more significant gains of STOI than Table III for quite low SNRs (e.g., 2.8 STOI gain vs. 1.6 STOI gain at -5 dB), which might be explained as that larger speech distortions existed in highly mismatch test set and post-processing could play a more important role in reducing the distortions. For SDR and PESQ measures, the observations were mostly similar to those in Table III, namely, PL+Dense+PP yielded remarkable performances gains over LSTM Baseline. Overall, SNR-PL with densely connected structure and post-processing showed much stronger generalization ability of unseen speaking styles and unseen noises over the conventional LSTM structures, which was quite important in realistic scenarios.


(a) Noisy (STOI=72.7, SDR=-0.25, PESQ=1.51)

(b) Clean

(c) LSTM Baseline (STOI=68.5, SDR=0.31, PESQ=1.14)

(d) PL (STOI=74.0, SDR=1.72, PESQ=1.28)

(e) PL+Dense (STOI=79.6, SDR=3.57, PESQ=1.69)

(f) PL+Dense+PP (STOI=80.0, SDR=3.69, PESQ=1.68)

Fig. 9. Spectrograms of an utterance corrupted by factory noise at 0 dB SNR: (a) Noisy speech, (b) Clean speech, (c) Baseline with four LSTM layers, (d) PL, (e) PL+Dense, (f) PL+Dense+PP.

Fig. 9 shows spectrograms of an utterance from the highly mismatched test set corrupted by factory noise at 0 dB SNR and enhanced by LSTM Baseline, PL, PL+Dense and PL+Dense+PP. The LSTM Baseline could achieve a good noise reduction but with severe speech distortion and speech loss, yielding the STOI and PESQ degradation over the unprocessed noisy speech. Meanwhile, the individual PL did not seem to solve these problems well. By using the dense structure, the speech distortion and speech loss problems in LSTM Baseline and PL were largely solved by PL+Dense, as shown in the yellow dotted box areas of Fig. 9(e). The post-processing slightly improved the enhanced spectrogram as shown in Fig. 9(f), with marginal performance gains in STOI and SDR measures.

Finally, we list an overall comparison of different methods on the highly mismatched test set across five unseen noises and two SNRs (-5 dB and 0 dB) in Table V. PL+CDense+PP is a compact version of PL+Dense+PP using the formulations in Eq. (7) and Eq. (8), namely concatenating only two latest intermediate targets to learn the next target. From Table V, although PL+Dense+PP with 5 hidden layers and 1024 LSTM cells for each layer improved all measures all over LSTM Baseline, a larger model size and a higher run-time latency were required. If we directly reduce the number of LSTM cells from 1024 to 512, the model size is half of that in LSTM Baseline but run-time latency is the same as LSTM Baseline. Subsequently, remarkable performance degradations for three measures were

TABLE V
THE OVERALL COMPARISON OF DIFFERENT METHODS ON THE HIGHLY MISMATCHED TEST SET ACROSS FIVE UNSEEN NOISES AND TWO SNRS (-5 dB AND 0 dB). N_{L} AND N_{C} REPRESENT THE NUMBER OF HIDDEN LAYERS AND LSTM CELLS, RESPECTIVELY. N_{M} AND N_{T} ARE THE MODEL SIZE AND RUN-TIME LATENCY NORMALIZED BY LSTM BASELINE

System	N_{L}	N_{C}	N_{M}	N_{T}	STOI	SDR	PESQ
Noisy	-	-	-	-	57.7	-3.20	1.56
LSTM Baseline	4	1024	1	1	56.2	2.09	1.58
PL+Dense+PP	5	1024	1.2	1.1	62.8	3.83	1.69
	5	512	0.5	1.0	61.1	3.00	1.59
PL+CDense+PP	5	1024	1.0	0.8	61.2	3.45	1.66
	5	512	0.4	0.7	61.6	3.23	1.64

observed especially for SDR and STOI. By using the compact version of PL+Dense+PP, PL+CDense+PP with 512 LSTM cells for each layer achieved better performance for all three measures, a smaller model size, and a lower run-time latency in comparison to PL+Dense+PP with 512 LSTM cells. So the PL+CDense+PP is a recommended version which makes a good trade between performance and complexity in real applications.

E. Experiment on Speech Recognition After SNR-PL

Four noise types provided by CHiME-4 Challenge [55] were chosen as our noise database. Clean speech was derived from the WSJ0 corpus [51]. 7138 utterances (about 12 hours of reading style speech) from 83 speakers, denoted as SI-84 training set, are corrupted with the above-mentioned 4 noise types at three SNR levels (-5 dB, 0 dB and 5 dB) to build a 36-hour training set, with pairs of clean and noisy speech.

As for the front-end, the best configuration of PL+Dense mentioned in the above section was utilized, and the ideal ratio mask (IRM), which was better for speech recognition, was also used as learning target in final layer. And the training process was the same as in the Section IV-A. The ASR system officially provided in [55] 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 [56]. The input of the DNN-HMM is a 440-dimensional feature vector extracted from Channel 5, consisting of a 40-dimensional fMLLR [57] with an 11-frame expansion. The model is trained according to the scripts downloaded from the official GitHub website[1] using Kaldi toolkit [58]. Note that all enhancement methods are only applied to the utterances in the recognition stage without retraining the acoustic model.

Table VI shows average WER (%) comparison of different enhancements on the development and test sets across four environments of CHiME-4 test set. There are three blocks in Table VI for the different enhancement methods.

For the first two block of Table VI, "Noisy" denotes the recognition of original noisy speech randomly selected from Channels 1-6 (except Channel 2), namely, 1-channel case. "LSTM-IRM" denotes the recognition of enhanced speech obtained by the LSTM-based regression model using the IRM as learning target. "LSTM-LPS" denotes the recognition of enhanced speech obtained by the LSTM-based regression model using the LPS feature as learning target. "ResNet-LPS" denotes the recognition of enhanced speech obtained by the ResNet-based regression model using the LPS feature as learning target. We observed that all regression models which directly learned the targets degraded the ASR performances, e.g., average WERs of 56.11% and 53.26% for "LSTM-LPS" and "ResNet-LPS" on real evaluation set, respectively.

TABLE VI
AVERAGE WER (%) COMPARISON ON THE DEVELOPMENT AND TEST SETS ACROSS FOUR CHIME-4 ENVIRONMENTS AFTER ENHANCEMENT

	Dev		**Eval**	
Enhancement	simu	real	simu	real
Noisy	15.64	15.68	24.13	27.67
LSTM-IRM	24.92	21.82	33.98	35.30
LSTM-LPS	39.23	36.50	52.49	56.11
ResNet-LPS	37.12	32.46	48.87	53.26
PL+Dense-T1-LPS	15.08	13.74	23.52	26.08
PL+Dense-T2-LPS	15.38	13.24	23.86	25.27
PL+Dense-T3-LPS	18.24	15.51	27.75	29.42
PL+Dense-T4-LPS	22.96	19.29	33.13	35.12
PL+Dense-T5-LPS	36.80	34.67	48.16	52.38
PL+Dense-T5-IRM	24.02	21.49	33.53	36.51

For the third block, "PL+Dense-T1-LPS", "PL+Dense-T2-LPS", "PL+Dense-T3-LPS", and "PL+Dense-T4-LPS" denote the recognition of enhanced speech obtained by the PL+Dense regression model using the outputs of target layer 1-4, respectively. First, the intermediate layer of the ResNet-based regression model can not directly be used for enhancement, while the intermediate target of the proposed PL-based regression model can be utilized because of its specific physical meaning. For example, "PL+Dense-T1-LPS" and "PL+Dense-T2-LPS" , the intermediate target with +5 dB and +10 dB SNR gain, can improve the ASR performance without acoustic model retraining at all situation comparing to "Noisy". Second, "PL+Dense-T3-LPS", and "PL+Dense-T4-LPS" degraded the ASR performance, but they are better than "LSTM-LPS" and "ResNet-LPS" which directly learned the LPS target. Finally, "PL+Dense-T5-LPS" and "PL+Dense-T5-IRM" denote the recognition of enhanced speech obtained by the PL+Dense regression model using the outputs of target layer 5 with LPS and IRM, respectively. We can find that the IRM as learning target outperforms LPS feature for recognition metric.

V. CONCLUSION

In this study, we propose a novel SNR-progressive learning framework for neural network based speech enhancement to improve the speech intelligibility. Specific to SNR-PL based speech enhancement, the direct mapping from noisy to clean speech is decomposed into multiple stages with SNR increasing

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

progressively by guiding hidden layers in the neural network to learn targets explicitly. We implement SNR-PL with both DNN and LSTM architectures and find LSTM is more suitable. We next improve SNR-PL with the dense structure in which the input and the estimations of intermediate targets are spliced together to learn the next target. In view of the multiple outputs in the progressive network, post-processing is then used in the testing stage to make a full use of the rich set of information. Experimental results on unseen noise conditions demonstrate that SNR-PL with densely connected structure and post-processing can learn more targets and yield much better speech intelligibility for all SNR levels. We also find that SNR-PL has a much stronger generalization ability than the conventional LSTM approach in highly mismatched conditions. Finally, intermediate outputs of the SNR-PL model can attain decreased word error rates for ASR because it can make a good tradeoff between the noise suppression and target speech preservation.

References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.

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

[3] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proc. IEEE*, vol. 67, no. 12, pp. 1586–1604, Dec. 1979.

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

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.

[6] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2002, pp. IV–4164.

[7] 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.

[8] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "A unified DNN approach to speaker-dependent simultaneous speech enhancement and speech separation in low SNR environments," *Speech Commun.*, vol. 95, pp. 28–39, 2017.

[9] 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.

[10] H. T. Fan, J. Hung, X. Lu, S. S. Wang, and Y. Tsao, "Speech enhancement using segmental nonnegative matrix factorization," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 4483–4487.

[11] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.

[12] 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.

[13] 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.

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

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

[16] 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, Dec. 2014.

[17] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Deep learning for monaural speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2014, pp. 1562–1566.

[18] P. S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Joint optimization of masks and deep recurrent neural networks for monaural source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2136–2147, Dec. 2015.

[19] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Proc. 9th Int. Conf. Artif. Neural Netw.*, 1999.

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

[21] M. Wöllmer, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Feature enhancement by bidirectional LSTM networks for conversational speech recognition in highly non-stationary noise," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6822–6826.

[22] 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 Inf. Process.*, 2014, pp. 577–581.

[23] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 708–712.

[24] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Latent Variable Analysis and Signal Separation*. Berlin, Germany: Springer-Verlag, 2015, pp. 91–99.

[25] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," *J. Acoust. Soc. Amer.*, vol. 141, no. 6, pp. 4705–4714, 2017.

[26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[27] S. W. Fu, Y. Tsao, and X. Lu, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 3768–3772.

[28] T. Kounovsky and J. Malek, "Single channel speech enhancement using convolutional neural network," in *Proc. IEEE Int. Workshop Electron., Control, Meas., Signals Appl. Mechatronics*, 2017, pp. 1–5.

[29] S. W. Fu, T. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Tran. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[30] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florencio, and M. Hasegawa-johnson, "Speech enhancement using Bayesian wavenet," in *Proc. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 2013–2017.

[31] D. Rethage, J. Pons, and X. Serra, "A wavenet for speech denoising," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5069–5073.

[32] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. 15th Annu. Conf. Int. Speech Commun. Assoc.*, 2014, pp. 2670–2674.

[33] M. Kim and P. Smaragdis, "Adaptive denoising autoencoders: A fine-tuning scheme to learn from test mixtures," in *Latent Variable Analysis and Signal Separation*. Berlin, Germany: Springer-Verlag, 2015, pp. 100–107.

[34] Y. Xu, J. Du, Z. Huang, L.-R. Dai, and C.-H. Lee, "Multi-objective learning and mask-based post-processing for deep neural network based speech enhancement," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 1508–1512.

[35] W. Jiang, H. Zheng, S. Nie, and W. Liu, "Multiscale collaborative speech denoising based on deep stacking network," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, pp. 1–5.

[36] T. Gao, J. Du, Y. Xu, C. Liu, L.-R. Dai, and C.-H. Lee, "Improving deep neural network based speech enhancement in low SNR environments," in *Latent Variable Analysis and Signal Separation*. Berlin, Germany: Springer-Verlag, 2015, pp. 75–82.

[37] X.-L. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 24, no. 5, pp. 967–977, May 2016.

[38] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.

[39] 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.*, 2016, pp. 3713–3717.

[40] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progresssive learning for LSTM-based speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 5054–5058.

[41] A. A. Rusu *et al.*, “Progressive neural networks,”2016, *arXiv:1606.04671*.

[42] A. Graves, A. Mohamed, and G. Hinton, “Speech recognition with deep recurrent neural networks,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 6645–6649.

[43] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vision Pattern Recog. (CVPR)*, Honolulu, HI, 2017, pp. 2261–2269.

[44] N. Takahashi and Y. Mitsufuji, “Multi-scale multi-band DenseNets for audio source separation,” *IEEE Workshop Appl. Signal Process. Audio Acoustics (WASPAA)*, New Paltz, NY, 2017, pp. 21–25.

[45] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[46] J. F. Santos and T. H. Falk, “Speech dereverberation with context-aware recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 7, pp. 1236–1246, Jul. 2018.

[47] J. F. Santos and T. H. Falk, “Investigating the effect of residual and highway connections in speech enhancement models,” in *Proc. Conf. Neural Inf. Process. Syst.*, 2018.

[48] C. Taal, R. Hendriks, R. Heusdens, and J. Jensen, “A short-time objective intelligibility measure for time-frequency weighted noisy speech,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4214–4217.

[49] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, “Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2001, pp. 749–752.

[50] 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, Jul. 2006.

[51] J. Garofalo, D. Graff, D. Paul, and D. Pallett, “CSR-I (WSJ0) complete,” in *Proc. Linguistic Data Consortium*, Philadelphia, 2007.

[52] A. Varga and H. J. Steeneken, “Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems,” *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[53] D. Yu *et al.*, “An introduction to computational networks and the computational network toolkit,” , Microsoft, Redmond, WA, USA, Tech. Rep. MSR-TR-2014–112, 2014.

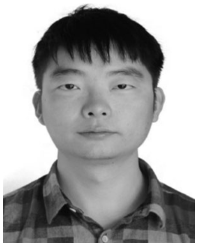
[54] J. Carletta *et al.*, “The AMI meeting corpus: A pre-announcement,” in *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, 2005, pp. 28–39.

[55] 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.

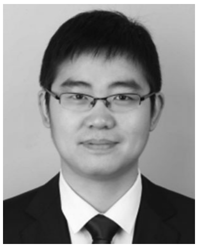
[56] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, “Sequence-discriminative training of deep neural networks.” in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, vol. 2013, pp. 2345–2349.

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

[58] D. Povey *et al.*, “The kaldi speech recognition toolkit,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understanding*, 2011.



Yan-Hui Tu received the B.S. degree from the Department of Electronic Information Engineering, Yunnan University in 2013 and Ph.D degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), in 2019, respectively. He is currently a Postdoctor at 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 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for nine months at Microsoft Research Asia (MSRA), Beijing. In 2007, he also worked as a Research Assistant for six months in the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he worked at iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA 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 (NEL-SLIP) of USTC.

Tian Gao received the B.S. degree from the Department of Communication Engineering, Hefei University of Technology in 2013 and received the Ph.D. degree from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2018. He is currently a Researcher at iFlytek AI research. His research interests include speech enhancement, robust speech recognition and speaker diarization.

Chin-Hui Lee (Fellow, IEEE) is a Professor in the School of Electrical, and Computer Engineering, Georgia Institute of Technology. Before joining academia in 2001 he had 20 years of industrial experience ending in Bell Laboratories, Murray Hill, New Jersey, as a Distinguished Member of Technical Staff, and Director of the Dialogue Systems Research Department. Dr. Lee is a Fellow of the IEEE, and a Fellow of ISCA. He has published over 400 papers, and 30 patents, and was highly cited for his original contributions with an h-index of 66. He received numerous awards, including the Bell Labs President’s Gold Award in 1998. He won the 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.