A LSTM-Based Joint Progressive Learning Framework for Simultaneous Speech Dereverberation and Denoising

XinTang*, JunDu*, LiChai*, Yannan Wang†, Qing Wang†, Chin-Hui Lee‡
* University of Science and Technology of China, HeFei, China
E-mail: tangxin@mail.ustc.edu.cn, jundu@ustc.edu.cn, cl122@mail.ustc.edu.com
† Tencent Technology (Shenzhen) Co., Ltd, Shenzhen, China
E-mail: yannanwang@tencent.com, sarahqwang@tencent.com
‡ Georgia Institute of Technology, Atlanta, Georgia, USA
E-mail: chl@ece.gatech.edu

***Abstract*—We propose a joint progressive learning (JPL) framework of gradually mapping highly noisy and reverberant speech features to less noisy and less reverberant speech features in a layer-by-layer stacking scenario for simultaneous speech denoising and dereverberation. As such layers are easier to learn than mapping highly distorted speech features directly to clean and anechoic speech features, we adopt a divide-and-conquer learning strategy based on a long short-term memory (LSTM) architecture, and explicitly design multiple intermediate target layers. Each hidden layer of the LSTM network is guided by a step-by-step signal-to-noise-ratio (SNR) increase and reverberant time decrease. Moreover, post-processing is applied to further improve the enhancement performance by averaging the estimated intermediate targets. Experiments demonstrate that the proposed JPL approach not only improves objective measures for speech quality and intelligibility, but also achieves a more compact model design when compared to the direct mapping and two-stage, namely denoising followed dereverberation approaches.**

I. Introduction

The presence of background noise severely degrades speech quality and intelligibility in speech communication. Besides, reverberation caused by the reflection of signal from room walls and other objects further deteriorates the auditory environment, especially in indoor environments [1]. In addition, for many real-world tasks, such as speech recognition and speaker recognition, the performance is greatly reduced under such acoustic conditions. Therefore, to improve the quality and intelligibility of the target speech under the condition that noise and reverberation exist simultaneously is quite necessary.

In the past few decades, most studies on denoising and dereverberation were carried through as two separate tasks which are called as speech enhancement and speech dereverberation, respectively. Due to the limitations of traditional methods under noisy-reverberant conditions, it is difficult to deal with background noise and reverberation simultaneously. Recently, benefiting from the development of deep learning approach in various tasks [2] [3] [4], many methods have been proposed to promote the speech quality and intelligibility in adverse environments [5] [6] [7]. From the perspective of deep learning, speech enhancement is regarded as a regression problem in which the deep neural network (DNN) learns the mapping between the noisy speech spectrum and the corresponding clean speech spectrum [8] or different masks [9] [10] to recover target speech. For instance, Ref. [11] proposed a spectral mapping algorithm to perform denoising and dereverberation simultaneously using a single DNN. However, this implementation did not achieve improvements in terms of speech intelligibility. In response to this issue [12] explained that the different natures of noise and reverberation make it difficult for DNN to handle them together. In general, background noise is an additive signal to clean speech, while reverberation is a convolutional process with a room impulse response (RIR) [13]. In consideration of the difference between noise and reverberation, Ref. [14] proposed a two-stage strategy for noisy-reverberant speech enhancement which means the whole speech enhancement process is divided into denoising stage and dereverberation stage sequentially and the model of each stage was trained individually before joint training. Moreover, another work on noisy and reverberant speech enhancement was time-frequency masking in complex domain by Williamson and Wang [15]. They introduced a complex ideal ratio mask (cIRM) using clean-anechoic speech as the desired signal for DNN-based enhancement. Furthermore, [16] utilized the maximum a posteriori (MAP) method to solve the speech dereverberation and denoising problem jointly. A half quadratic splitting (HQS) method was adopted to solve the joint MAP problem in a DNN framework by splitting it into two minimization problems.

In this study, inspired by our previous work of progressive learning for speech denoising only [17], we proposed a joint progressive learning (JPL) framework based on signal-to-noise ratio (SNR) and reverb time -60dB (RT60) measures for joint speech denoising and dereverberation. In JPL, each hidden layer implemented by the long short-term memory (LSTM) architecture is guided to learn an intermediate target with a specific SNR gain and RT60 estimation. Different from the direct mapping approach [11] and the idea of two-stage design in [14], we utilize the idea of divide-and-conquer and explicitly

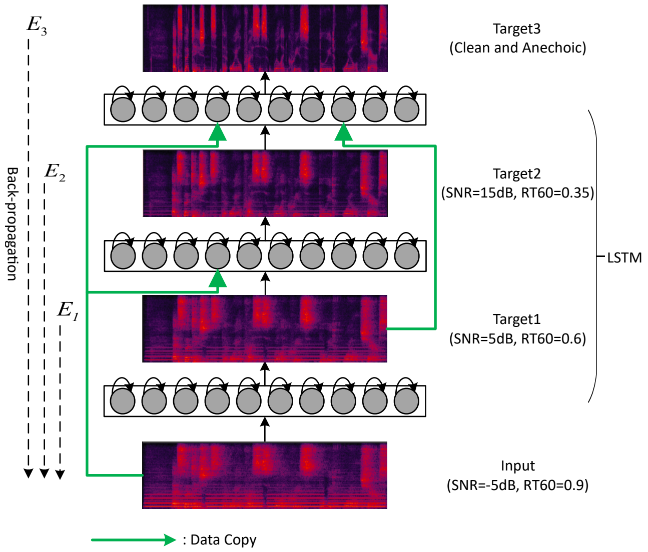


Fig. 1. An illustration of joint progressive learning

design multiple intermediate target layers with jointly denoised and dereverberated speech to some extent. Moreover, a post-processing is applied to further improve the enhancement performance by averaging the estimations of multiple intermediate targets. Experiments show that the proposed JPL approach not only significantly improves objective measures for speech quality and intelligibility, but also achieves a more compact model design compared with the direct mapping and two-stage approaches.

The rest of the paper is organized as follows. In Section II, we describe the proposed JPL approach. In Section III, we present the experiments. Finally, we conclude in Section IV.

II. Joint Progressive Learning (JPL)

A. Signal model

In time domain, background noise is an additive signal to clean speech while reverberation is a convolutional signal. In this way, the noisy-reverberant speech $x(t)$ can be formulated as:

$$x(t) = r(t) + g{\cdot}n(t) = s(t) * h(t) + g{\cdot}n(t) \qquad (1)$$

where $r(t)$, $s(t)$, $h(t)$, $n(t)$ denote reverberant speech, anechoic speech, room impulse response function and noise, respectively. Besides, g is an adjustable factor of utterance level employed to control the SNR level and $*$ refers to convolution operation. The training data in our work are created with different types of noise, SNRs, reverberation times and speakers according to this equation. In order to restore the anechoic speech $s(t)$ from noisy-reverberant speech $x(t)$, intuitively we can perform dereverberation after denoising as shown in [14], namely eliminating noise $n(t)$ from received speech $x(t)$ firstly and then recover target anechoic speech $s(t)$ from reverberant speech $r(t)$.

B. Motivation

Regarding to the abovementioned two-stage approach, the handling of denoising and dereverberation is independent using two separate networks. However, in practice the corruption process of $s(t)$ with background noises and reverberation could be coupled due to many complicated environmental factors. This motivates us to treat denoising and dereverberation simultaneously in a progressive way by using deep learning architectures to implement this process.

Based on the formulation of noisy-reverberant speech signal in (1), our JPL approach aims to progressively make g and $h(t)$ close to zero and unit impulse response respectively. Accordingly, we concretize the implementation of JPL by using the SNR measure for g and RT60 measure for $h(t)$.

C. Design of intermediate target layer and model training

The illustration of the proposed JPL is shown in Figure 1. The ”Data Copy” means copying the current 257-dimensional vector and then concat it with the next target layer. With clean speech corrupted by the noise and reverberation, we redefine the intermediate target layer to guide hidden layer to learn the corresponding target illuminated by curriculum training. For denoising, the design of each intermediate layer achieves a specific SNR gain. As for dereverberation, we adopt RT60 in the model training procedure which is defined as the time it takes for the sound pressure level to reduce by 60 dB according to [18]:

$$\mathrm{RT60} = \frac{24\ln 10}{c_{20}} \cdot \frac{V}{Sa} \qquad (2)$$

where c_{20} is the speed of sound in the room for 20 degrees Celsius, V is the volume of the room in m^3, S stands for total surface area of room in m^2 and a refers to the average absorption coefficient of room surfaces. In this study we utilize RT60 as an indicator to measure the level of reverberation and to guide the intermediate layer to learn the target speech under lower RT60 condition. Table I lists the SNR and RT60 configuration of our JPL approach in training stage. For instance, if the SNR of input noisy-reverberant speech is -5dB, 0dB, 5dB and the RT60 is 0.9s, 0.8s, 0.7s, the SNR and RT60 of first target layer to learn are set as 5dB, 10dB, 15dB and 0.6s, 0.5s, 0.4s, respectively. Similarly, for other top target layers, the SNR of target speech is increased while the corresponding RT60 is reduced as shown in Table I. For the final target layer, the clean-anechoic speech is set. In fact, for RT60-based progressive learning, as the position of source and microphone is unchanged when generating the corresponding RIR, gradually decreasing RT60 is equivalent to truncating the original RIR, so that the RIR progressively approaches the unit impulse response.

For the network design, LSTM-based densely connected multi-task learning is adopted as shown in Figure 1. All the target layers are designed to learn intermediate speech features with higher SNRs and lower RT60 simultaneously. LSTM layers are adopted as the hidden layers to learn each target. This stacking style network can learn multiple targets progressively and efficiently. As for objective function optimization, with

TABLE I
SNR AND RT60 CONFIGURATIONS OF JPL IN TRAINING STAGE

	SNR (dB)			**RT60 (s)**		
Input	-5,	0,	5	0.90,	0.80,	0.70
Target1	5,	10,	15	0.60,	0.50,	0.40
Target2	15,	20,	25	0.35,	0.25,	0.15
Target3		Clean			Anechoic	

prediction errors from each target layer, a weighted minimum mean squared error (MMSE) criterion is used to update all network parameters randomly initialized with K target layers as follows:

$$E = \sum_{k=1}^{K} \alpha_k E_k \tag{3}$$

$$E_k = \frac{1}{N} \sum_{n=1}^{N} \|F(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, ..., \hat{\boldsymbol{x}}_n^{k-1}, \mathbf{W}_k) - \boldsymbol{x}_n^k\|_2^2 \tag{4}$$

where $\hat{\boldsymbol{x}}_n^k$ and $\boldsymbol{x}_n^k$ are the estimated and reference target log-power spectra (LPS) feature vectors of D-dimensional for k^{th} target layer, respectively $(k > 0)$, with N representing the mini-batch size. $\hat{\boldsymbol{x}}_n^0$ denotes the n^{th} D-dimensional vector of input noisy LPS features. $F(\hat{\boldsymbol{x}}_n^0, \hat{\boldsymbol{x}}_n^1, ..., \hat{\boldsymbol{x}}_n^{k-1}, \mathbf{W}_k)$ is the neural network function for k^{th} target with the dense structure using the previously learned intermediate targets from $\hat{\boldsymbol{x}}_n^0$ to $\hat{\boldsymbol{x}}_n^{k-1}$, and $\mathbf{W}_k$ represents the parameter set of weight matrices and bias vectors before k^{th} target layer, which are optimized in the manner of the back propagation through time BPTT with gradient descent [19]. In this work, K is equal to 3 according Figure 1 and Table I. We set $\alpha_1 = \alpha_2 = 0.1$ and $\alpha_3 = 1$.

D. Post-processing

One benefit of joint progressive learning is that more than one estimated target are obtained with the network. Moreover, the estimated LPS features of different targets can provide rich information with different preferences to noise suppression or dereverberation in complex environments. Therefore, in the inference stage, a post-processing method to average the estimations of multiple targets can be adopted to further improve the overall performance similar to [20].

III. EXPERIMENTS AND RESULT ANALYSIS

A. Experimental setup

In our experiments, clean and anechoic speech data is derived from the WSJ0 corpus [21] and 115 noise types were selected as our noise database. We utilized an RIR generator [22] to generate the RIRs, which is based on the image model [23]. In addition, we kept the distance (2m) between the receiver and the speaker, so that the direct to reverberant ratio (DRR) did not change much under each RT60 condition. The RIRs in training set and test set are generated with different room sizes, which are 4m×6m×3m and 10m×7m×3m, respectively. For training set, firstly we convolved 7138 utterances (about 15 hours) from 83 speakers with the above mentioned RIRs at three RT60 values (0.9s, 0.8s, 0.7s) to generate reverberant utterances. And then we corrupted the obtained reverberant utterances with 115 noise types [24] at three SNR levels (-5dB, 0dB, 5dB) to build a 135-hour training set composed of pairs of clean-anechoic and noisy-reverberant utterances. Similarly, 330 utterances from 8 other speakers, namely the Nov92 WSJ evaluation set, 3 RIRs with three unseen RT60 values (0.75s, 0.85s, 0.95s), 5 unseen noises including babble, buccaneer, factory1, hfchannel and pink from NOISEX-92 corpus [25], were used to construct the test set. Perceptual evaluation of speech quality (PESQ) [26] and short-time objective intelligibility (STOI) [27] are adopted to evaluate the intelligibility and quality of enhanced speech.

As for feature extraction, first the speech waveform was sampled at 16kHz, and the corresponding frame length was set to 32 msec (512 samples) with a frame shift of 16 msec (256 samples). A short-time Fourier analysis was employed to calculate the spectra of each overlapping windowed frame. Thus, the 257-dimensional LPS features were produced and normalized by global mean and variance before feeding them into the neural network [28]

B. Implemented approaches for comparison

We built two different systems to compare with our JPL approach. For all deep learning based systems, LSTM was used with 1024 units for each layer. One competing system is based on direct mapping [11] which estimates the clean-anechoic speech features directly from noisy-reverberant speech features. We explored different settings of the LSTM architecture and the best configuration was achieved by 3 LSTM layers. We denote this system as **Direct Mapping** in the subsequent sections. The other competing system is based on the two-stage approach in which we perform denoising and dereverberation sequentially as in [14]. In our implementation, to perform a fair comparison, two separate networks were jointly trained via a similar multi-task learning as in our JPL approach. We also investigated different settings of LSTM architecture for each network and the best configuration was 3 LSTM layers for denoising network and 1 LSTM layer for dereverberation network. We denoted this system as **Two-Stage** in the subsequent sections.

C. Result analysis

Table II and Table III list the average STOI and PESQ results of different systems across 5 unseen noise types at different SNR levels and RT60 settings, respectively. **Noisy+Reverb** refers to the unprocessed system. **JPL-Target1**, **JPL-Target2**, and **JPL-Target3** are our PPL systems by using Target1, Target2, and Target3 for enhancement as shown in Figure 1, respectively. **JPL-PP** denotes our JPL system using the post-processing of Target2 and Target3 as introduced in Section II-D.

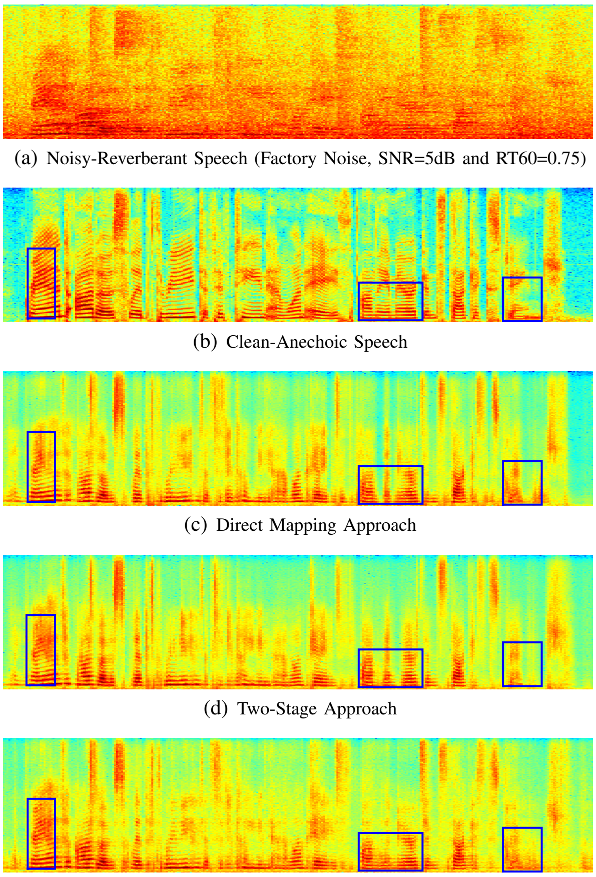
First, both direct mapping and two-stage approaches achieved similar STOI and PESQ improvements compared with noisy-reverberant speech under 0dB and positive SNR

TABLE II
THE AVERAGE STOI COMPARISON OF DIFFERENT SYSTEMS ACROSS 5 UNSEEN NOISES AT DIFFERENT SNR LEVELS, UNDER EACH OF RT60

	STOI															
RT60 (s)	**0.75**					**0.85**					**0.95**					N_{M}
SNR (dB)	**-5**	**0**	**5**	**10**	**15**	**-5**	**0**	**5**	**10**	**15**	**-5**	**0**	**5**	**10**	**15**	
Noisy+Reverb	0.483	0.554	0.621	0.671	0.700	0.471	0.536	0.597	0.642	0.669	0.459	0.521	0.579	0.621	0.646	-
Direct Mapping	0.490	0.610	0.695	0.745	0.773	0.469	0.588	0.676	0.727	0.755	0.457	0.573	0.661	0.714	0.744	85 MB
Two-Stage	0.496	0.622	0.707	0.757	0.783	0.476	0.598	0.685	0.737	0.764	0.463	0.585	0.672	0.725	0.753	106 MB
JPL-Target1	0.515	0.601	0.671	0.717	0.741	0.500	0.579	0.646	0.688	0.711	0.486	0.562	0.626	0.669	0.691	21 MB
JPL-Target2	0.545	0.647	0.722	0.766	0.790	0.527	0.625	0.698	0.742	0.765	0.532	0.609	0.682	0.727	0.750	46 MB
JPL-Target3	0.527	0.645	0.728	0.774	0.798	0.505	0.623	0.709	0.755	0.779	0.493	0.610	0.695	0.744	0.769	75 MB
JPL-PP	**0.548**	**0.661**	**0.742**	**0.788**	**0.811**	**0.528**	**0.639**	**0.721**	**0.767**	**0.791**	**0.515**	**0.625**	**0.706**	**0.755**	**0.780**	75 MB

TABLE III
THE AVERAGE PESQ COMPARISON OF DIFFERENT SYSTEMS ACROSS 5 UNSEEN NOISES AT DIFFERENT SNR LEVELS, UNDER EACH OF RT60

	PESQ															
RT60 (s)	**0.75**					**0.85**					**0.95**					N_{M}
SNR (dB)	**-5**	**0**	**5**	**10**	**15**	**-5**	**0**	**5**	**10**	**15**	**-5**	**0**	**5**	**10**	**15**	
Noisy+Reverb	1.23	1.35	1.53	1.67	1.76	1.18	1.31	1.46	1.60	1.67	1.24	1.33	1.47	1.60	1.67	-
Direct Mapping	1.13	1.46	1.75	1.93	2.06	1.08	1.39	1.68	1.86	1.98	1.09	1.37	1.64	1.82	1.92	85 MB
Two-Stage	1.12	1.47	1.77	1.97	2.09	1.07	1.39	1.68	1.89	2.00	1.08	1.36	1.64	1.83	1.94	106 MB
JPL-Target1	1.30	1.53	1.73	1.87	1.94	1.26	1.47	1.66	1.78	1.85	1.30	1.48	1.66	1.78	1.84	21 MB
JPL-Target2	1.39	1.70	1.95	2.11	2.20	1.34	1.63	1.88	2.03	2.11	1.41	1.62	1.85	1.99	2.06	46 MB
JPL-Target3	1.22	1.59	1.90	2.09	2.20	1.16	1.51	1.82	2.01	2.11	1.17	1.49	1.77	1.95	2.05	75 MB
JPL-PP	**1.34**	**1.70**	**1.99**	**2.17**	**2.27**	**1.28**	**1.62**	**1.91**	**2.09**	**2.18**	**1.30**	**1.60**	**1.87**	**2.04**	**2.13**	75 MB



(a) Noisy-Reverberant Speech (Factory Noise, SNR=5dB and RT60=0.75)

(b) Clean-Anechoic Speech

(c) Direct Mapping Approach

(d) Two-Stage Approach

(e) Our Approach

Fig. 2. Spectrograms of an utterance example

level conditions. For instance, the STOI gain was 0.095 at the input SNR of 15dB and RT60 of 0.85s while the PESQ gain was 0.33 for the two-stage approach. However, about more than 0.1 PESQ decline was observed for -5dB condition with all values of RT60 which indicated that these two comparing systems did not work in the extremely adverse environments with quite low SNR and large reverberation. Second, compared with direct mapping and two-stage approaches, **JPL-Target3** could achieve remarkable improvements for both STOI and PESQ, e.g., STOI increasing from 0.757 to 0.774 and PESQ increasing from 1.97 to 2.09 at SNR=10dB and RT60=0.75s. Moreover, **JPL-Target2** was strongly complementary with **JPL-Target3**, namely yielding better STOI at low SNRs and worse STOI at high SNRs. The reason is that these two targets make different tradeoffs between noise/reverberation reduction and introduced nonlinear distortions. Accordingly, the post-processing system **JPL-PP** by combining Target2 and Target3 achieved the best overall performance for both STOI and PESQ measures. Finally, our **JPL-PP** consistently and significantly outperformed the most competing system **Two-Stage** in terms of STOI and PESQ for all SNR levels and all RT60 settings, especially under adverse environments, for instance more than 0.2 PESQ gain and 0.05 STOI gain at SNR=-5dB and RT60=0.75.

Our proposed JPL approach not only yielded better STOI and PESQ results but also achieved more compact model design than direct mapping and two-stage approaches. As shown in the rightmost columns in Tables II and III, N_{M} denotes the size of deep models for enhancement in each system. Clearly, all JPL models were smaller than those in direct mapping and two-stage approaches. **JPL-Target1** are more compact than **JPL-Target3** in that it is not necessary to go through the whole network to compute Target2 and Target3 as shown in Figure 1. In real applications, **JPL-Target2** is a good solution as it makes a better tradeoff between the performance and efficiency.

Figure 2 shows spectrograms of an utterance corrupted by factory noise at SNR=5dB/RT60=0.75 and enhanced by different approaches. The direct mapping could achieve a good noise reduction but with severe speech distortions while the two-stage method reduced speech distortions as shown in blue rectangle boxes of Figure 2(c) and Figure 2(d). Compared with direct mapping and two-stage approaches, our JPL approach using post-processing achieved the best restoration of speech segments, leading to the improved speech quality and intelligibility.

IV. Conclusions

In this study, we propose a JPL framework for noisy-reverberant speech enhancement. In contrast to the direct mapping and two-stage methods, we attempt to recover clean and anechoic speech as a simultaneous elimination of noise and reverberation combined with progressive learning. Moreover, since each intermediate target provides complementary information, post-processing to integrate different targets can also be performed to further improve speech quality and intelligibility. Experiments demonstrate that the proposed JPL framework achieves good PESQ and STOI improvements across all noise levels and reverberation conditions. In the future, the generalization capabilities of JPL in real-world environments will be explored.

V. Acknowledgements

This work was supported in part by the National Key R&D Program of China under contract No. 2017YFB1002202, the National Natural Science Foundation of China under Grants No. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005, and Huawei Noah's Ark Lab.

References

[1] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation.* Springer Science & Business Media, 2010.

[2] Z. Zhang, J. Geiger, J. Pohjalainen, A. E.-D. Mousa, W. Jin, and B. Schuller, "Deep learning for environmentally robust speech recognition: An overview of recent developments," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 9, no. 5, p. 49, 2018.

[3] J. Rohdin, A. Silnova, M. Diez, O. Plchot, P. Matějka, and L. Burget, "End-to-end dnn based speaker recognition inspired by i-vector and plda," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 4874–4878.

[4] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[5] M. Cooke, V. Aubanel, and M. L. G. Lecumberri, "Combining spectral and temporal modification techniques for speech intelligibility enhancement," *Computer Speech & Language*, vol. 55, pp. 26–39, 2019.

[6] S. R. Chiluveru and M. Tripathy, "Low snr speech enhancement with dnn based phase estimation," *International Journal of Speech Technology*, pp. 1–10, 2019.

[7] D. Ribas, J. Llombart, A. Miguel, and L. Vicente, "Deep speech enhancement for reverberated and noisy signals using wide residual networks," *arXiv preprint arXiv:1901.00660*, 2019.

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

[9] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines.* Springer, 2005, pp. 181–197.

[10] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing.* IEEE, 2013, pp. 7092–7096.

[11] K. Han, Y. Wang, D. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[12] Y. Zhao, D. Wang, I. Merks, and T. Zhang, "Dnn-based enhancement of noisy and reverberant speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2016, pp. 6525–6529.

[13] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[14] Y. Zhao, Z.-Q. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 53–62, 2019.

[15] D. S. Williamson and D. Wang, "Time-frequency masking in the complex domain for speech dereverberation and denoising," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 25, no. 7, pp. 1492–1501, 2017.

[16] A. Raikar, S. Basu, and R. M. Hegde, "Single channel joint speech dereverberation and denoising using deep priors," in *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP).* IEEE, 2018, pp. 216–220.

[17] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Densely connected progressive learning for lstm-based speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2018, pp. 5054–5058.

[18] R. W. Young, "Sabine reverberation equation and sound power calculations," *The Journal of the Acoustical Society of America*, vol. 31, no. 7, pp. 912–921, 1959.

[19] P. J. Werbos *et al.*, "Backpropagation through time: what it does and how to do it," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[20] T. Gao, J. Du, L.-R. Dai, and C.-H. Lee, "Snr-based progressive learning of deep neural network for speech enhancement." in *INTERSPEECH*, 2016, pp. 3713–3717.

[21] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csr-i (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[22] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.

[23] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[24] G. Hu, "100 nonspeech environmental sounds,"[online] available: http://web. cse. ohio-state. edu/pnl/corpus/hunonspeech," *HuCorpus. html*, 2004.

[25] 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 communication*, vol. 12, no. 3, pp. 247–251, 1993.

[26] 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 *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

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

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