2D-TO-2D MASK ESTIMATION FOR SPEECH ENHANCEMENT BASED ON FULLY CONVOLUTIONAL NEURAL NETWORK

Yan-Hui Tu[1], Jun Du[1], Chin-Hui Lee[2]

[1]University of Science and Technology of China, Hefei, Anhui, P.R.China
[2]Georgia Institute of Technology, Atlanta, GA, USA
tuyanhui@mail.ustc.edu.cn, jundu@ustc.edu.cn, chl@ece.gatech.edu

ABSTRACT

In recent years, the deep learning-based approaches are popular in the field of singe-channel speech enhancement. Convolutional neural networks (CNNs) are a standard component of many current speech enhancement system. In this study, we design a new Fully CNN (FCNN)-based regression model, which can directly achieve the 2-dimensional (2D) noisy lpg-power spectra (LPS) input to 2-dimensional (2D) time-frequency mask output mapping, denoted as 2D-RFCNN. First, the whole 2D noisy LPS of one utterance is directly used as network input to make sure each convolutional filter can see more contextual information. Second, we only use the pooling operation on the frequency bin to ensure that the final dimension of frequency bin has a value of 1 and make the number of feature mapping same to frequency dimension, simultaneously. Finally, we also use the deep convolutional layers with a small size of filter, which is popularly used in speech recognition, for speech enhancement. Experiments of the CHiME-4 challenge task shows that our proposed 2D-RFCNN model not only improves the speech quality (PESQ) and intelligibility (STOI), but also reduces the recognition error rate on real test set.

Index Terms— speech enhancement, 2D-to-2D mapping, ideal ratio mask, deep learning, fully convolutional neural network

1. INTRODUCTION

Single channel speech enhancement is a widely researched problem in signal processing, which aims to suppress the background noise and interference from the observed noisy speech to improve the perceptual quality and the performance of automatic speech recognition (ASR) [1]. The problem of speech enhancement has been an attractive area of research in statistical signal processing for a rather long time, and short-time Fourier transform (STFT) based methods achieve relatively good performance in this field [2]. It is appropriate to further categorize this class of speech enhancement algorithms into the sub-categories of spectral subtraction [3], Wiener filtering [4], minimum mean-square error (MMSE) estimator [5], and the optimally modified log-spectral amplitude (OM-LSA) speech estimator [6]. These conventional methods are adaptive to the test signal, which is in general not robust enough in adverse environments, particularly when there are non-stationary noises.

Recently, a supervised learning framework has been proposed, where a deep neural network (DNN) is trained to map from input features to the output targets. In [7], a regression DNN is adopted using mapping-based method directly predicting the log-power spectra (LPS) of clean speech from LPS of the noisy speech. In [8], the authors use the regression DNN to learn the complex relationship between the noisy spectra and the reference mask. More complicated neural network architectures, for example long short-term memory (LSTM) based recurrent neural network (RNN) [9], convolutional neural network (CNN) [10], fully CNN (FCNN) [11] and CNN-BLSTM [12] with an expense of higher computational complexities and run-time latencies than the conventional DNN are applied to speech enhancement. But the input of existing CNN-based speech enhancement models is a 2-dimensional (2D) sliding window of the whole LPS for one utterance and the network output is a one-dimensional (1D) vector, so we can regard the whole regression process as 2D-to-1D mapping. For those regression tasks, the sizes of input and output feature maps are the same and the network aims to learn the complex relationship between noisy features and reference features at each time-frequency units.

Based on the above analysis, a better way to solve the regression problem is to find a new architecture to achieve the 2D input to 2D output feature mapping on the whole utterance level due to the strong contextual information in speech application. For the 2D-to-2D mapping problem, the information contained by 2D input on utterance level is much richer than 1D vector on frame level or 2D input with a narrow size on the time axis due to the frame expansion. Obviously, compared with the conventional 1D-to-1D mapping (e.g., DNN or LSTM) and 2D-to-1D mapping (e.g., CNN), it is much more challenging to perform a 2D-to-2D mapping due to the curse of dimensionality from the perspective of machine learning. In this study, we design a new CNN-based regression model, which can directly achieve the 2-dimensional (2D) noisy spectrogram input to 2-dimensional (2D) time-frequency mask output mapping, denoted as 2D-RFCNN. First, the whole 2D noisy LPS of one utterance is directly used as network input to make sure each convolutional filter can see more contextual information. Second, we only use the pooling operation on the frequency bin to ensure that the final dimension of frequency bin has a value of 1 and make the number of feature mapping same to frequency dimension, simultaneously. Finally, we also use the deep convolutional layers with small size of filter, which is popularly used in speech recognition [13, 14], for speech enhancement. Experiments of the CHiME-4 challenge task shows that our proposed 2DR-CNN model not only improves the speech quality, perceptual evaluation of speech quality (PESQ) [15] and intelligibility, short time bbjective intelligibility (STOI) [16], but also reduces the recognition error rate on real test set.

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 Grant Nos. 61671422 and U1613211, the Key Science and Technology Project of Anhui Province under Grant No. 17030901005. This work was also funded by Huawei Noah's Ark Lab.

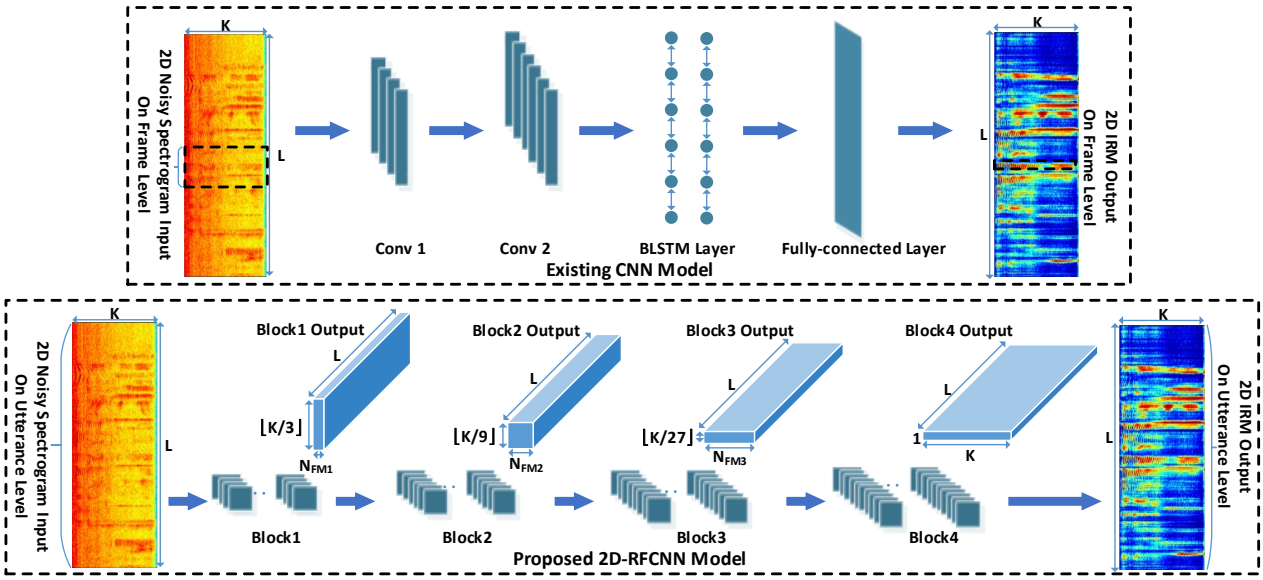

Fig. 1. The comparison of existing CNN model and proposed 2D-RFCNN model for speech enhancement. $\lfloor \cdot \rfloor$ denotes the floor function. $N_{\text{FM1}}, N_{\text{FM2}}, N_{\text{FM3}}$ denote the number of feature map at each block.

2. ARCHITECTURAL AND TRAINING NOVELTIES

For the speech enhancement problem, given an utterance of noisy log-power spectral (LPS), denoted as $\mathbf{X} \in \mathbb{R}^{K \times L}$, where K and L are the dimensions of frequency bin and time frame, respectively. The corresponding reference ideal ratio mask (IRM) [8] is denoted as $\mathbf{M}_{\text{ref}} \in \mathbb{R}^{K \times L}$ and the estimated mask by neural network is denoted as $\hat{\mathbf{M}} \in \mathbb{R}^{K \times L}$. The $\mathbf{x}^l \in \mathbb{R}^{K \times 1}$, $\mathbf{m}^l_{\text{ref}} \in \mathbb{R}^{K \times 1}$ and $\hat{\mathbf{m}}^l \in \mathbb{R}^{K \times 1}$ are the K-dimensional noisy LPS, reference IRM and estimated mask at l-th frame. A large amount of time-synchronized training set, denoted as $D = \{(\mathbf{X}^i, \mathbf{M}^i_{\text{ref}}) | i = 1, 2, ..., N\}$, with N pairs of noisy LPS, $\mathbf{X}$, and reference IRM, $\mathbf{M}_{\text{ref}}$ is built simultaneously. The training data are synthesized by adding different types of noise to the clean speech utterances with different SNR levels.

2.1. Review of 2D-to-1D CNN-based Mask Estimation

This section will give a review of a mask-based speech enhancement using deep learning-based regression model. For convenience, we first define the $\mathbf{X}^l_{\text{2D-F}} \in \mathbb{R}^{K \times (2\tau+1)}$ as noisy LPS input with frame expansion and τ is the number of frames in both of left and right context at l-th frame. The neural network predicts the mask $\hat{\mathbf{m}}^l$ at l-th frame as follow:

$$\hat{\mathbf{m}}^l = f_\theta(\mathbf{X}^l_{\text{2D-F}}; \theta) \tag{1}$$

where f_θ is parameterized by θ.

The top of Fig. 1 shows the popular CNN training strategy, and the network input usually is a processed 2D LPS with frame expansion. The size of each 2D LPS input is $K \times (2\tau+1)$, and the dimension of each output vector is $K \times 1$. So we can see the previous CNN-based regression task as two-dimensional LPS to one-dimensional vector mapping, denoted as 2D-1D. And the supervised fine-tuning is usually used to minimize the mean squared error (MSE) between the neural network output $\hat{\mathbf{m}}^l$ and the reference IRM $\mathbf{m}^l_{\text{ref}}$, which is defined as

$$\min_\theta \sum_{i=1}^{N_{\text{mini\_F}}} \left\| \hat{\mathbf{m}}^i - \mathbf{m}^i_{\text{ref}} \right\|_2^2 = \min_\theta \sum_{i=1}^{N_{\text{mini\_F}}} \left\| f_\theta(\mathbf{X}^i_{\text{2D-F}}; \theta) - \mathbf{m}^i_{\text{ref}} \right\|_2^2 \tag{2}$$

where $N_{\text{mini\_F}}$ denote the number of frames in each mini-batch and $\|\cdot\|_2$ denotes the L2 norm of a vector. A Adam-based back-propagation method [17] is adopted to update the parameters of a neural network in a mini-batch mode.

2.2. The Proposed 2D-to-2D FCNN-based Mask Estimation

In this study, we design a new CNN-based regression model, which can directly archive the 2-dimensional (2D) noisy LPS input to 2-dimensional (2D) reference IRM output mapping. In the following, we will give a detailed description of our proposed 2D-to-2D CNN-based regression model, denoted as 2D-RFCNN.

2.2.1. 2D Noisy LPS Input

For speech enhancement task, the CNN is usually utilized to solve the two-dimensional LPS input to one-dimensional vector output mapping problem with a fully-connected layer or recurrent layer connecting to output layer, as shown in the top of Fig. 1. On the one hand, it will produce a plenty of redundant use of training data and computations due to the frame expansion and frame shift. On the other hand, the size of each feature map along the time dimension is $2\tau + 1$, where the τ usually is 5, so it is hard for the CNN to take advantage of weights sharing in convolutional networks to see more contextual information along the time axis. In [11], the authors also find 1-D convolutional filter is better than 2-D convolutional filter for speech enhancement in the 2D-to-1D regression mapping case. In this study, the whole noisy LPS of one utterance is adopted as the input to reduce the redundant use of training data. Another advantage of the proposed model is weights sharing in convolutional networks, so it can make full use of the contextual information along the time axis.

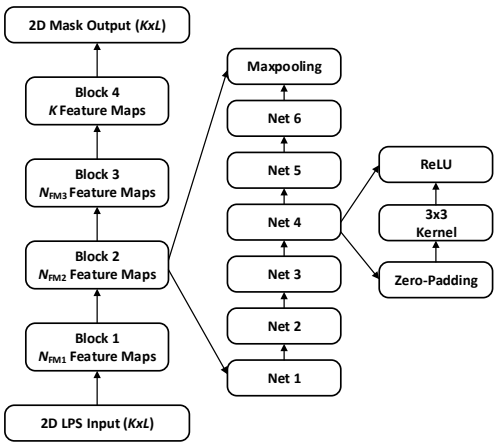

Fig. 2. The detailed architecture of proposed 2D-RFCNN model.

2.2.2. 2D Mask Output

As illustrated in the bottom of Fig.1, the light blue cuboids show the size of each block output. In order to achieve the final dimension of frequency bin with a value of 1, we only use one pooling operation on the frequency bin in each block. The number of feature maps is also gradually increased after each block, and we set the number of feature maps in last block is the same as the frequency dimension K.

Let the convolutional kernel of size is $b \times w$. As opposite to the traditionally used lager size of filter in speech enhancements, we only use filters of 3×3 with pooling size constrained to 3×1, which is popularly applied in speech recognition [18]. After a convolution operation, the input size $(K \times L)$ of each feature map will be changed without padding, e.g., the output size of each feature map is $(K-b+1)\times(L-w+1)$ with the $b\times w$ convolutional kernel. To ensure that the final output of our network has the same length in the time dimension as the input LPS, the zero-padding operation is applied before each convolutional layer shown in Fig. 2. And for the last two nets in block4, the zero-padding operation on frequency bin is not utilized to ensure the final dimension of frequency bin with a value of 1. The neural network predicts the mask $\hat{\mathbf{M}}$ for each utterance as follow:

$$\hat{\mathbf{M}} = g_\theta(\mathbf{X};\theta) \tag{3}$$

where g_θ is parameterized by θ.

2.2.3. 2D-RFCNN Optimization on Utterance Level

Based on the above detailed expansion, our proposed FCNN-based model can directly output the estimated 2D mask with the same size of input LPS to achieve the 2D-to-2D mapping for regression task on the utterance level. The supervised fine-tuning is used to minimize the MSE between the neural network output $g_\theta(\mathbf{X})$ and the reference IRM $\mathbf{M}_{\text{ref}}(k,l)$, which is defined as follow:

$$\min_\theta \sum_{i=1}^{N_{\text{mini\_U}}} \left\| \hat{\mathbf{M}}^i - \mathbf{M}^i_{\text{ref}} \right\|_2^2 = \min_\theta \sum_{i=1}^{N_{\text{mini\_U}}} \left\| g_\theta(\mathbf{X}^i;\theta) - \mathbf{M}^i_{\text{ref}} \right\|_2^2 \tag{4}$$

where $N_{\text{mini\_U}}$ denotes the number of utterances in each min-batch and $\|\cdot\|_2$ denotes the L2 norm of a matrix. Adam-based back-propagation method [17] is adopted to update the parameters of a neural network in a mini-batch mode.

3. EXPERIMENTAL EVALUATION

3.1. Data Corpus

Now, we present the experimental evaluation of our framework in the CHiME-4 task [19], which was designed to study real-world ASR scenarios where a person is talking to a mobile tablet device equipped with 6 microphones in a variety of adverse environments. Four conditions were selected: café (CAF), street junction (STR), public transport (BUS), and pedestrian area (PED). For each case, two types of noisy speech data were provided: RealData and SimData. RealData was collected from talkers reading the same sentences from the WSJ0 corpus [20] in the four conditions. SimData, on the other hand, was constructed by mixing clean utterances with environmental noise recordings using the techniques described in [21]. CHiME-4 offers three tasks (1-channel, 2-channel, and 6-channel) with different testing scenarios. In this paper, we focus only on the 1-channel case. The readers can refer to [19] for more detailed information regarding CHiME-4.

3.2. Implementation Details

For front-end configurations, speech waveform is sampled at 16 kHz, and the corresponding frame length is set to 512 samples (or 32 msec) with a frame shift of 128 samples. The STFT analysis is used to compute the DFT of each overlapping windowed frame. To train the 2D-RFCNN model, the 2D reference IRM of one utterance with size of $257 \times L$ was used for target. N_{FM1}, N_{FM2} and N_{FM3} are 64, 128 and 190, respectively. PyTorch was used for neural network training [22]. The learning rate for the first 5 epochs was initialized as 0.25 and then decreased by 90% after each epoch, and the number of epochs was 10. And the mini-bath, $N_{\text{mini\_U}}$, is 4. The CHiME-4 challenge [23] training set was used as our training data. Specifically, we used simulated training data from Channel 1, Channel 3 and Channel 5 with 7138 utterances (about 12 hours) for each channel to train the enhancement models. We compare our proposed method with the following three networks:

1) DNN with 3 hidden layers and 2048 nodes for each layer.

2) LSTM with 2 hidden layers and 1024 cells for each layer.

3) EHNET [12] with the best settings of convolution and BLSTM layers in the paper.

The context frame of the above three methods is all set as 11. And the mini-bath, $N_{\text{mini\_F}}$, is 256.

For the back-end configurations, the baseline ASR recognition system is trained on the speech recognition toolkit Kaldi [24]. For TDNN acoustic model training, backstitch optimization method is used. The decoding is based on 3-gram language models with explicit pronunciation and silence probability modeling. The model is re-scored by a 5-gram language model first. Then the Kaldi-RNNLM is used for training the RNN, and n-best re-scoring is used to improve performance. The model is trained according to the scripts downloaded from the official GitHub website[1].

3.3. Experiments on Enhancement

Fig. 3 illustrates a comparison of learning curves of the different regression models using the averaged squared errors normalized by frame on the simulated development set. Clearly, the learning curve

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

Table 1. STOI (%) and PESQ comparisons of conventional IMCRA approach, IRM-based deep learning approach using different neural network architectures for single-channel speech enhancement on the simulation test set.

	STOI(%)					PESQ				
Enhancement	BUS	CAF	PED	STR	AVG	BUS	CAF	PED	STR	AVG
Noisy	84.0	79.0	81.3	80.2	81.1	2.14	1.89	1.92	1.96	1.98
IMCRA	85.4	81.5	82.5	81.6	82.8	2.25	1.98	2.03	2.11	2.09
DNN-IRM	85.2	81.3	82.1	81.2	82.5	2.34	2.06	2.09	2.15	2.16
LSTM-IRM	88.1	83.8	85.8	83.9	85.4	2.46	2.20	2.25	2.28	2.30
EHNET-IRM [12]	89.4	84.7	86.7	85.9	86.7	2.58	2.29	2.36	2.39	2.40
2D-RFCNN-IRM	**89.8**	**86.5**	**88.0**	**86.6**	**87.7**	**2.63**	**2.37**	**2.42**	**2.44**	**2.46**

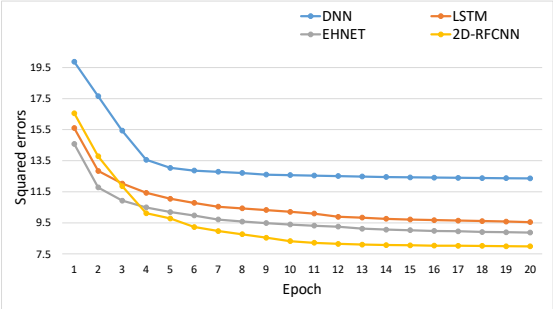

Fig. 3. The learning curve comparison on the development set.

of the proposed 2D-RFCNN model could achieve better convergence than those of the other regression models. More interestingly, the initial point of the learning curve of the 2D-RFCNN model was higher than that of the LSTM and EHNET model, which demonstrated that 2D-to-2D mapping is more difficult to learn than 2D-to-1D mapping initially..

Table 1 shows STOI (%) and PESQ comparisons of conventional IMCRA approach, IRM-based deep learning approach using different neural network architectures for single-channel speech enhancement on the simulation test set. For the first block, "Noisy" denotes the original speech randomly selected from channel 1-6 (except channel 2), namely 1-channel case. "IMCRA" denotes the enhanced speech is obtained by IMCRA-based speech enhancement [25]. For the second block, "LSTM-IRM", "EHNET-IRM" and "2D-RFCNN-IRM" denote the enhanced speech is obtained by the estimated IRM using LSTM, EHNET and 2D-RFCNN regression models. We can find that the proposed method produces consistently better PESQ and STOI performance than the DNN, LSTM and EHNET approaches.

3.4. Experiments on ASR

Table 2 shows WER(%) the comparison of conventional IMCRA approach, IRM-based deep learning approach using different neural network architectures and ISPP-based deep learning approach using different neural network architectures for single-channel speech enhancement on the real test set. The results in Table 2 are obtained by the Kaldi tools without acoustic model retraining.

For the first block, we can find that the IRM estimated by "LSTM-IRM" and "EHNET-IRM" can only improve the ASR performance slightly, comparing to "Noisy". For example, the average WER of "Noisy" is 12.14%, while the average WERs of "LSTM-IRM" and "EnhNet-IRM" are 14.07% and 11.69%, respectively.

Table 2. WER (%) comparison of conventional IMCRA approach, IRM-based deep learning approach using different neural network architectures and ISPP-based deep learning approach using different neural network architectures for single-channel speech enhancement on the real test set.

Enhancement	BUS	CAF	PED	STR	AVG
Noisy	19.05	12.35	9.34	7.81	12.14
IMCRA	24.40	16.62	11.79	8.26	15.26
DNN-IRM	26.19	18.95	12.67	9.12	16.73
LSTM-IRM	22.51	14.76	11.03	7.98	14.07
EHNET-IRM [12]	18.27	12.31	8.67	7.51	11.69
2D-RFCNN-IRM	**16.95**	**11.73**	**7.98**	**7.12**	**10.94**
ISPP-EHNET [12]	14.75	9.21	7.14	5.97	9.27
ISPP-2D-RFCNN	**13.98**	**8.71**	**6.66**	**5.45**	**8.70**

"IMCRA", a kind of classic speech enhancement, is also failed to improve the ASR performance. While our proposed 2D-RFCNN model can significantly improve the ASR performance comparing to "Noisy", with a relative WER reduction of 9.88%.

For the third block, "ISPP" denotes the improved speech presence probability (ISPP)-based method proposed in [26], which combines the classic speech enhancement and IRM-based method. "ISPP-EHNET" and "ISPP-2D-RFCNN" denote the enhanced speech is obtained by the ISPP-based method with the IRM estimated by EHNET and 2D-RFCNN, respectively. For the proposed 2D-RFCNN-based regression model, it still can outperform the EHNET-based regression model using the better ISPP framework. For example, "ISPP-2D-RFCNN" can improve ASR performance with a relative WER reduction of 6.15%, comparing to "ISPP-EHNET".

4. CONCLUSION

In this work we proposed a novel FCNN-based regression model for single-channel speech enhancement with 2-dimensional (2D) noisy lpg-power spectra (LPS) input and 2D time-frequency mask output , denoted as 2D-RFCNN. Because the network input and output are the whole utterance feature map, the deep convolutional layers with a small size of filter can be applied in our architecture for regression task. Experiments on the CHiME-4 challenge task shows that our proposed 2D-RFCNN model not only improves the speech quality (PESQ) and intelligibility (STOI), but also reduces the recognition error rate on real test set comparing to the competing methods. In the future, we will applied our 2D-RFCNN to other regression tasks, for example speech separation.

5. REFERENCES

[1] Li Deng and Xiao Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 5, pp. 1060–1089, 2013.

[2] Philipos C Loizou, *Speech Enhancement : Theory and Practice*, CRC press, second edition, 2013.

[3] Steven F Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[4] Jae S Lim and Alan V Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.

[5] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[6] Israel Cohen and Baruch Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, 2001.

[7] Yong Xu, Jun Du, Lirong Dai, and Chinhui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2015.

[8] Yuxuan Wang, Arun Narayanan, and DeLiang Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[9] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Bjorn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[10] S.-W. Fu, Y. Tsao, and X. Lu, "Snr-aware convolutional neural network modeling for speech enhancement.," in *Proc. Annual Conference of International Speech Communication Association. (INTERSPEECH)*, 2016.

[11] Se Rim Park and Jinwon Lee, "A fully convolutional neural network for speech enhancement," pp. 1993–1997, 2017.

[12] H. Zhao, S. Zarar, I. Tashev, and C. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 2401–2405.

[13] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2263–2276, Dec 2016.

[14] T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun, "Very deep multilingual convolutional neural networks for lvcsr," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4955–4959.

[15] Antony William Rix, John G Beerends, Michael Peter Hollier, and Andries Pieter Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," vol. 2, pp. 749–752, 2001.

[16] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper 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.

[17] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *international conference on learning representations*, 2015.

[18] Tom Sercu, Christian Puhrsch, Brian Kingsbury, and Yann Lecun, "Very deep multilingual convolutional neural networks for lvcsr," in *Proc. IEEE Int'l Conf. Acoust. Speech Signal Process.(ICASSP)*, 2016.

[19] 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," *Computer Speech and Language*, 2016.

[20] J. Garofalo, D. Graff, D. Paul, and D. Pallett, "Csri (wsj0) complete," *Linguistic Data Consortium, Philadelphia*, 2007.

[21] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933—1950, 2007.

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer, "Automatic differentiation in pytorch," 2017.

[23] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third chime speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 504–511.

[24] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," 2011.

[25] Israel Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[26] Yan-Hui Tu, Ivan Tashev, Shuayb Zarar, and Chin-Hui Lee, "A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition," in *international conference on acoustics, speech, and signal processing.(ICASSP)*, 2018, pp. 2531–2535.