LINEAR PREDICTION-BASED PART-DEFINED AUTO-ENCODER USED FOR SPEECH ENHANCEMENT

Zihao Cui, Changchun Bao

Speech and Audio Signal Processing Laboratory, Faculty of Information Technology
Beijing University of Technology, Beijing, 100124, China
cuizihao@emails.bjut.edu.cn, baochch@bjut.edu.cn

ABSTRACT

This paper proposes a linear prediction-based part-defined auto-encoder (PAE) network to enhance speech signal. The PAE is a defined decoder or a defined encoder network, based on efficient learning algorithm or classical model. In this paper, the PAE utilizes AR-Wiener filter as decoder part, and the AR-Wiener filter is modified as a linear prediction (LP) model by incorporating the modified factor from residual signal. The parameters of line spectral frequency (LSF) of speech and noise and the Wiener filtering mask are utilized for training targets. Finally, the proposed the LP-based PAE is compared with the baseline method, namely the Wiener filtering mask-based DNN. The PESQ and STOI results of the LP-based PAE are better than baseline method at lower signal noise ratio (SNR) levels.

***Index Terms*—** Part-defined auto-encoder, speech enhancement, DNN, linear prediction, residual signal

1. INTRODUCTION

Single-channel speech enhancement, aiming at improving speech quality and reduce noise, is an essential part of speech processing and usually used for pre-processing of robust speech recognition [1], hearing aids [2] and robust speech coding [3]. In recent years, with the growing needs of speech processing, single-channel speech enhancement has been payed more and more attention.

Speech enhancement algorithms have been developed over several decades. In 1979, Boll proposed Spectral subtraction [4] algorithm that is used to estimate short-time-frequency spectrum of speech by removing noise spectrum. In 1984, Ephraim and Malah proposed the MMSE-based spectral amplitude estimators [5, 6] to improve the signal noise ratio (SNR) and ensure the quality of speech. However, these unsupervised speech enhancement approaches are not able to reduce non-stationary noise efficiently.

The supervised algorithm is an efficient way to process non-stationary noise. It learns some priori information from the training data and use these priori information to improve intelligibility and quality of noisy speech. As one of the supervised algorithm, Non-negative Matrix Factorization (NMF) [7] assumes that speech and noise are independent. By separating speech and noise components in terms of matrix, the NMF-Wiener filter can be constructed to estimate clean speech spectrum. Deep neural network (DNN) is one of the most progressively supervised approaches for speech enhancement. DNN can predict some features of clean speech, such as log-power spectra (LPS) [8], the Wiener filter parameters [9], and Mel-Frequency Cepstral Coefficient (MFCC) [10]. One of the successful applications of the DNN in speech signal processing is the generative adversarial network (GAN) [11]. It employs a discriminator in network to assimilate true data. Another one is Auto-encoder (AE) [12], which gives a clear coding strategy for feature space, that is, the network is comprised of encoder and decoder. The denoising auto-encoder (DAE) network given in [13] is also an example of the AE aiming to reduce noise. However, the network layer selection of the AE is so general that it is difficult to approach the particular model, such as a speech model.

The NMF-styled reconstruction that uses the NMF to learn basic spectral patterns and build the NMF-layer to drive the Wiener filter is used in DNN [14]. These spectral patterns reflect the physical or perceptual properties of speech. Another classical speech model is the vocal tract model [15]. Auto-regressive (AR) model, as a simplified vocal tract model, was used in well-known codebook-driven speech enhancement [16]. Then, the DNN-based and codebook-driven algorithm [17] was proposed to reduce computation complexity. However, this AR model-based algorithm does not consider the influence of linear residual and use iteration approach to estimate gains of AR parameters. This resulted in some errors for constructing the Wiener filter parameters.

These works showed the combination of the DNN and other valid models as an organic whole might be a more advisable strategy in speech signal processing. Follow this line of thought, a new network is proposed that combines the AE with the AR-Wiener model for speech enhancement. This new network is defined as the part-defined auto-encoder (PAE), in which the encoder or decoder of the AE is set as a fixed or slowly changed network based on classical model or training method. The linear prediction (LP) model-based PAE is used to predict the AR-Wiener filter parameters, in which the line spectral frequency (LSF) parameters, AR gains

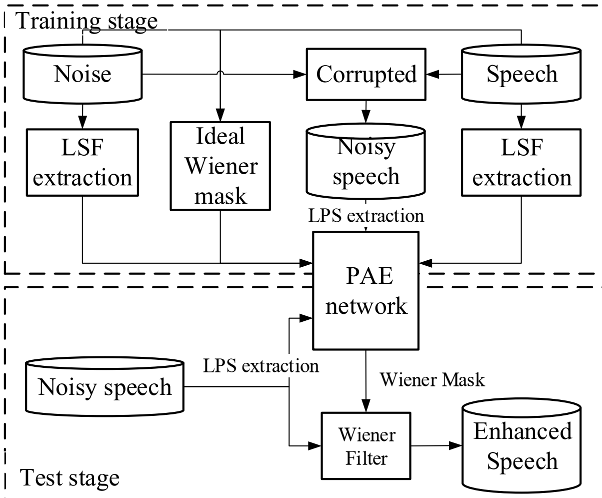

Fig. 1. The block diagram of the proposed PAE

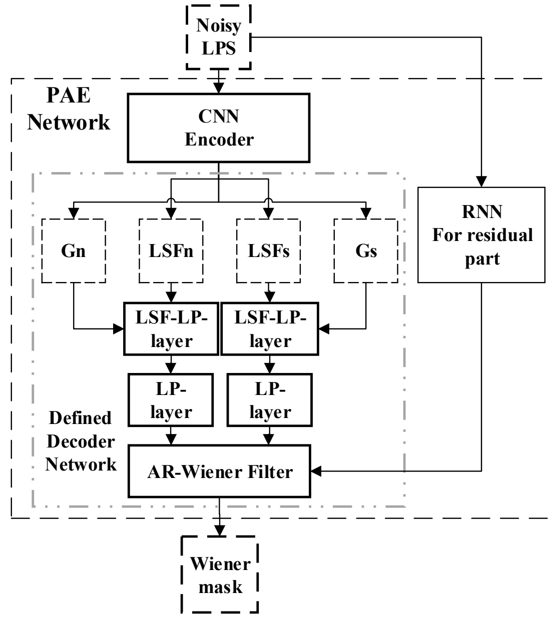

Fig. 2. The proposed PAE (the small block with the dashed line is data space, and the block with solid line is network layer)

are used for middle feature spaces target, and Wiener mask are used for final target. Furthermore, a self-defined the LP layer that can transfer the LP parameters into magnitude spectrum is used to drive the AR-Wiener filter for obtaining magnitude spectrum of clean speech. Based on the feature of spectral envelope, there is a definite range to select activation function and network layers. For modifying the accuracy of the AR-Wiener filter, the modification factor is incorporated by Recurrent Neural Network (RNN) [18].

The paper is organized as follows: The PAE with linear prediction is described in section 2. System evaluation and comparison are given in section 3. The conclusions are summarized in section 4.

2. PAE WITH LINEAR PREDICTION

A block diagram of the proposed speech enhancement framework is illustrated in Figure 1. In the training stage, the ideal Wiener filtering mask, LSF parameters of speech and noise are extracted as the training target for learning the PAE. In test stage, the PAE is used to predict the Wiener filtering mask and drive the Wiener filter. The structure of the PAE is a vital part in this section. The PAE with the AR-Wiener model is described in 2.1. Then, the AR-Wiener model is modified by analyzing vocal tract model in 2.2. Finally, the PAE with the modified AR-Wiener is given in 2.3.

2.1. PAE based on the AR-Wiener filter

The PAE, the vital part in figure 1, is illustrated in figure 2. As the essential part of the proposed model, the decoder network is based on the AR-Wiener model [16], which is:

$$H_{AR-WF}(k)=\frac{\dfrac{g^{(s)}}{\left|A^{(s)}(k)\right|^2}}{\dfrac{g^{(s)}}{\left|A^{(s)}(k)\right|^2}+\dfrac{g^{(n)}}{\left|A^{(n)}(k)\right|^2}} \tag{1}$$

where g is the AR gains. $1/\left|A(k)\right|^2$ is spectral envelope of signal, k is frequency index, and the superscript (s) and (n) respectively are speech and noise. Based on Eq. (1), the AR-Wiener-based PAE is used to decode the AR parameters. For the PAE illustrated in Figure 2 with inverse analysis, from bottom to the top of decoder network, the layer that transfers the gain and LP parameter into magnitude spectrum is defined as the LP-layer. Considering the LSF parameters are more stable and according to the method proposed by Kang et al. [19], the layer that transfers the layer with LSF parameters into the layer with the LP parameters is defined as the LSF-LP-layer. The coding spaces that consist of gain and the LSF of speech and noise are the output of encoder network and the input of decoder network, as shown in the blocks with the thin-dashed line in figure 2.

The loss function E_r of this paper consist of three parts: target error between the estimated value and mask value of the Wiener filtering, LSF error of speech signal and LSF error of noise signal, thus E_r can be expressed as follows:

$$\begin{aligned} E_r=&\frac{1}{K}\left\|\overline{\mathbf{WM}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)-\mathbf{WM}_l\right\|_2^2\\ &+\frac{1}{N}\left\|\overline{\mathbf{LSF}^{(s)}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)-\mathbf{LSF}_l^{(s)}\right\|_2^2\\ &+\frac{1}{N}\left\|\overline{\mathbf{LSF}^{(n)}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)-\mathbf{LSF}_l^{(n)}\right\|_2^2 \end{aligned} \tag{2}$$

where $\overline{\mathbf{WM}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)$ and $\mathbf{WM}_l$ respectively denote the l^{th} estimated and target frequency bins of the Wiener filtering mask. $\overline{\mathbf{LSF}}\left(\mathbf{Y}_{l-t}^{l+t},\mathbf{W},\mathbf{b}\right)$ and $\mathbf{LSF}_l$ respectively denote the l^{th} estimated and target LSF parameters, the superscript (s) and (n) respectively refer to speech and noise. The l is the index of frame and $\mathbf{Y}_{l-t}^{l+t}$ is log-spectral feature vectors of noisy that cover t previous frames, 1 current frame and t future frames. **W** and **b** are weight and bias parameters to be learned. K is the number of sample point, and N is the number of LSF parameters in a frame.

Once the decoder of the PAE is defined, the encoder network layer is selected based on analysis of the coding spaces. Since the LSF parameter is an equivalent expression of mathematics with the LP coefficients used for spectral envelope construction, the convolution neural network (CNN) can be used to learn LSF parameters and AR gains of speech and noise through the LPS of noisy speech. Here, two convolutional layers [20] are used to extract LPS, and then the max-pooling activation function [21] is utilized to find peaks of the LPS. Next, the fully connected (FC) layers are selected to linearly map out LSF parameters of speech and noise from the LPS of noisy. Finally, the Maxout method [22] is employed to find the AR gains of speech and noise.

2.2. Modified AR-Wiener with the residual

The vocal tract model of speech signal is a kind of AR model [15], so the linear prediction residual in time domain can be illustrated by an infinite impulse response (IIR) filter,

$$g \bullet r(n) = x(n) * \left[\delta(n) - \delta(n) * h_{IIR}(p)\right] \tag{3}$$

where $r(n)$ is normalized residual, δ is unit impulse function, $h_{\mathrm{IIR}}(p)$ is the p order IIR filter, g is the gain of the AR model, and n is the index of discrete time.

The expression of frequency domain for Eq. (3) is as follows:

$$X(k) = g\frac{R(k)}{1 - H_{IIR}(k)} \tag{4}$$

where $H_{IIR}(k)$ is the discrete Fourier transform of the IIR filter, $1\text{-}H_{IIR}(k)$ is regarded as a short-term prediction (STP) filter, $R(k)$ is residual spectrum. It is critical to incorporat residual into the AR-Wiener. Substituting Eq.(4) into Eq.(1), the modified AR-Wiener filter is given by

$$H_{M-AR-WF}(k) = \frac{\dfrac{g^{(s)}R^{(s)}(k)}{\left|A^{(s)}(k)\right|^2}}{\dfrac{g^{(s)}R^{(s)}(k)}{\left|A^{(s)}(k)\right|^2} + \dfrac{g^{(n)}R^{(n)}(k)}{\left|A^{(n)}(k)\right|^2}} \tag{5}$$

The Eq. (5) is called the LP-based AR-Wiener filter and can be expressed as follows:

$$H_{M-AR-WF}(k) = H_{AR-WF}(k)H_{r-WF}(k) \tag{6}$$

where the $H_{r\text{-}WF}(k)$ is a modification factor that that will be estimated from residual signal by the RNN. $H_{AR\text{-}WF}(k)$ is defined in Eq. (1). Eq. (6) indicates the Wiener filtering mask influenced by residual signal. The Wiener filtering mask is also influenced by the correlation of spectral envelopes of speech and noise.

Table 1. The detailed structure of the PAE

Summary of the PAE				
Input: 11 frames noisy speech with 129 frequency bins				
Layer index	Type of layer	Number of filter	Output shape (nodes)	Previous output layer
CNN structure				
1	Conv2D(2,2) Maxp2D(2,2)	6	(10,128) (5,64)	input
2	Conv2D(2,2) Maxp2D(2,2)	16	(4,63) (2,31)	1
3	FC		2048	2
4	FC		2048	3
5	FC		12	4 (first 1500)
6	FC		12	4 (last 1500)
7	FC Maxout		20 1	3 (first 100)
8	FC Maxout		20 1	3 (last 100)
Output of CNN: LSFs and gains of speech and noise				
RNN structure				
9	RU(2FC)		1419	input
10	FC		2048	9
11	FC		129	9
Output of RNN: Modified factors in 129 frequency bins				

3. SYSTEM EVALUATION AND COMPARISON

3.1. Data set and PAE structure

All experiments were conducted on the TIMIT corpus [23]. Speech signal was down-sampled to 8 kHz. The Babble, F16, Factory, Office and Street noises were chosen from NOISEX-92 noise database [24], in which the Babble, F16 and Factory noises were used both in training and test part, and Factory and Office noises were used for generalization test. 40 minutes utterances were added with the aforementioned three training noises at -5 dB, 0 dB, 5 dB and 10 dB SNR levels. Another 201 randomly selected utterances from the TIMIT test set were used for five noises at four same SNR levels. The Wiener filtering mask from the DNN-based algorithm [9] was used as reference mask to evaluate performance of the proposed method. For comparison, this reference is named as the W-DNN. The proposed AR-Wiener filter-based PAE algorithm is named as the Pro. A, and the modified AR-Wiener filter-based algorithm is named as the Pro. B.

As for signal analysis, the corresponding frame length was set to 256 samples with an overlap of 128 samples. The 129-dimension LPS was used as input in the training and test stages for three methods. The LSF parameters of speech and noise and the Wiener filtering mask were extracted from parallel speech corpus and noise database.

The details of the PAE were illustrated in table 1. The input shape of the CNN was shown in the 2D spaces in terms of time and frequency. The input shape of the RNN was

Table 2. PESQ results

SNR(dB)	Noise Type	Noisy	W-DNN	Pro A	Pro B
-5	Babble	1.719	1.941	1.793	**1.986**
	F16	1.595	2.125	2.037	**2.139**
	Factory	1.761	2.298	2.207	**2.333**
	Office	1.860	2.067	1.975	**2.078**
	Street	1.990	2.570	2.502	**2.614**
	Average	1.785	2.200	2.103	**2.230**
0	Babble	1.931	2.338	2.180	**2.377**
	F16	1.832	2.520	2.401	**2.523**
	Factory	2.073	2.699	2.577	**2.734**
	Office	2.175	2.489	2.403	**2.511**
	Street	2.320	2.946	2.868	**2.983**
	Average	2.066	2.598	2.486	**2.625**
5	Babble	2.221	2.733	2.567	**2.749**
	F16	2.136	2.898	2.771	**2.908**
	Factory	2.391	3.064	2.932	**3.086**
	Office	2.492	2.889	2.811	**2.906**
	Street	2.650	3.287	3.198	**3.297**
	Average	2.378	2.974	2.856	**2.989**
10	Babble	2.518	**3.098**	2.930	**3.098**
	F16	2.444	3.225	3.105	**3.239**
	Factory	2.705	**3.374**	3.244	3.37
	Office	2.803	**3.256**	3.171	3.253
	Street	2.973	**3.577**	3.477	3.565
	Average	2.689	**3.306**	3.185	3.305

shown in 1D spaces in terms of requency. The Conv2d(2,2) denotes a 2D convolutional layer with the filter size of 2×2, and the Maxp2D(2,2) denotes the down-sampling by 2 both in time and frequency. The previous layer 4 (first 1500) means that the layer is connected to the first 1500 nodes in the output of the 4^{rd} layer. The resdual unit (RU) consisted of two FC layers without bias. The rectified linear unit (Relu) [25] was used for activation function of the 1^{st}, 2^{nd}, 3^{rd}, 4^{th} and 10^{th} layers, and the sigmoid function with weight π was used for activation function of the 5^{th} and 6^{th} layers. The activation function was not used for the 7^{th}, 8^{th} and 11^{th} layers. The dropout layers with 0.2 rate were used behind the 3^{rd}, 4^{th} and 10^{th} layer. The W-DNN has three hidden layers with 2048 points and the Relu activation layer with 0.2 rate dropout.

Three speech enhancement systems were implemented by PyTorch, all networks were trained by Adam optimizer [26], and the initial learning rate was set to 0.0001. Our evaluation measures for speech enhancement are the Perceptual Evaluation of Speech Quality (PESQ) [27] and the short-time objective intelligibility (STOI) [28].

3.2. Results

Table 2 shows the PESQ results among Noisy, W-DNN, Pro. A and Pro. B at different SNR levels for five noise types. From Table 2, we can see that the PESQ results of all methods are better than the Noisy. However, the PESQ result of Pro. A is poorer than the W-DNN. For this case, the reason may be that the residual signal is not considered in Pro. A so that

Table 3. The average results of the PESQ, STOI

	Noisy		W-DNN		Pro. B	
SNR	PESQ	STOI%	PESQ	STOI %	PESQ	STOI %
-5	1.785	57.14	2.200	68.39	**2.230**	**69.04**
0	2.066	68.24	2.598	78.67	**2.625**	**78.80**
5	2.378	78.01	2.974	**85.77**	**2.989**	85.56
10	2.689	85.70	**3.306**	**90.45**	3.305	90.06
Avg.	2.230	72.27	2.770	80.82	**2.787**	**80.87**

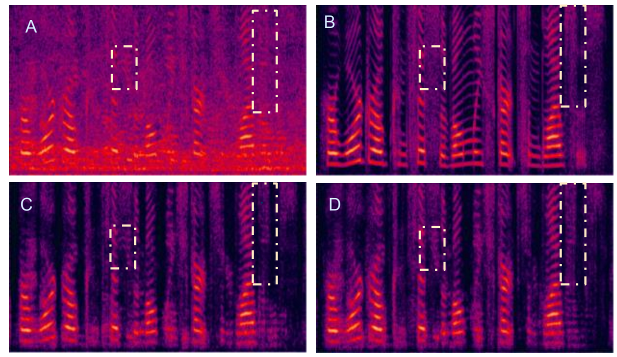

Fig. 3. Spectrogram comparison. A. Speech corrupted by Babble at 0 dB; B. Clean speech; C. Enhanced speech by the W-DNN; D. Enhanced speech by Pro. B

lots of structures of voiced speech is broken. Since the modified factor of the AR-Wiener from residual is incorporated, the PESQ results of Pro. B are better than W-DNN in most cases, especially in lower input SNR.

Since the PESQ results of the Pro. A are not good, the performance of the Pro. A is not shown in table 3 and Fig. 3. In table 3, the average PESQ and STOI results are showed among five abovementioned noise types. The Pro. B is better than W-DNN in many cases, and particularly at 0dB and -5dB. The spectrograms of an utterance are displayed in Fig. 3. The labeled part shows that the unvoiced speech signal is estimated more accurate based on the Pro. B. This phenomenon may be derived from the results of incorporating the modified AR-Wiener factor and may lead to better PESQ and STOI results at lower SNR levels.

4. CONCLUSIONS

In this paper, the AR-Wiener filter to obtain an AR-Wiener based PAE replaced the decoder in the AE. Then, the modified factor of the AR-Wiener filter was incorporated so that speech enhancement method derived from the LP-based PAE was given. The test results showed that the proposed method got a satisfactory result compared with the reference, especially at the low SNR levels. In the future work, the enhanced model of residual can be considered, and some classical algorithms can be combined with the PAE.

5. ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61831019 and No. 61471014). The authors are also grateful to the thorough reviewers.

6. REFERENCES

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

[2] C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi, "An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device," *IEEE signal processing letters,* vol. 24, no. 11, pp. 1601-1605, 2017.

[3] Z. Li, and M. W. Hoffman, "Evaluation of microphone arrays for enhancing noisy and reverberant speech for coding," *IEEE Transactions on Speech and Audio Processing,* vol. 7, no. 1, pp. 91-95, Jan, 1999.

[4] S. 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.

[5] Y. Ephraim, and D. 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] Y. Ephraim, and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing,* vol. 33, no. 2, pp. 443-445, 1985.

[7] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," in Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on, 2011, pp. 45-48.

[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] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 22, no. 12, pp. 1849-1858, 2014.

[10] 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 (TASLP),* vol. 23, no. 1, pp. 7-19, 2015.

[11] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[12] C.-Y. Liou, W.-C. Cheng, J.-W. Liou, and D.-R. Liou, "Autoencoder for words," *Neurocomputing,* vol. 139, pp. 84-96, 2014.

[13] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in Interspeech, 2013, pp. 436-440.

[14] S. Nie, S. Liang, W. Liu, X. Zhang, and J. Tao, "Deep Learning Based Speech Separation via NMF-Style Reconstructions" *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 11, pp. 2043-2055, 2018.

[15] P. C. Loizou, *Speech enhancement: theory and practice*: CRC press, 2013.

[16] M. Kuropatwinski, and W. B. Kleijn, "Estimation of the excitation variances of speech and noise AR-models for enhanced speech coding," in icassp, 2001, pp. 669-672.

[17] Y. Yang, and C. Bao, "Dnn-Based Ar-Wiener Filtering for Speech Enhancement," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 2901-2905.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition." pp. 770-778.

[19] G. Kang, and L. Fransen, "Application of line-spectrum pairs to low-bit-rate speech encoders," in Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85., 1985, pp. 244-247.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[21] J. J. Weng, N. Ahuja, and T. S. Huang, "Learning recognition and segmentation of 3-d objects from 2-d images," in Computer Vision, 1993. Proceedings., Fourth International Conference on, 1993, pp. 121-128.

[22] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n,* vol. 93, 1993.

[24] 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.

[25] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in Proceedings of the fourteenth international conference on artificial intelligence and statistics, 2011, pp. 315-323.

[26] D. P. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] 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 Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on, 2001, pp. 749-752.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, J. J. I. T. o. A. Jensen, Speech,, and L. Processing, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," vol. 19, no. 7, pp. 2125-2136, 2011.