DNN-Based Linear Prediction Residual Enhancement for Speech Dereverberation

Xinyang Feng, Nuo Li, Zunwen He, Yan Zhang, and Wancheng Zhang

School of Information and Electronics, Beijing Institute of Technology
Beijing, 100081, China
zhangwancheng@bit.edu.cn

***Abstract*—In daily-life scenarios, reverberation inevitably causes a decrease in speech recognizability and speech quality. Exploring methods to eliminate reverberation will benefit both human perception and other speech technology applications such as identity authentication and speech recognition. This paper proposes a speech dereverberation algorithm based on linear prediction (LP) residual processing using deep neural network (DNN). The amplitude spectrum of the LP residual of short-term speech is used as a speech feature to train the DNN, and the mapping relationship between LP residual of the reverberant speech and that of the clean speech is learned. Comparative experiments under different reverberation conditions have verified the effectiveness and robustness of the algorithm.**

***Index Terms*—speech dereverberation, linear prediction residual, deep neural network**

I. Introduction

In daily-life scenarios, sound that reaches the ears usually includes the original source sound and its reflections on various surfaces, especially in some confined spaces, when using microphones or other receiving devices to capture the sound, the original sound is combined with these attenuated, delayed reflections to form a reverberant signal. Reverberation will inevitably cause a decrease in speech recognizability and speech quality [1]. In a reverberant environment, the speech intelligibility for hearing-impaired listeners is significantly reduced [2], and severe reverberation will also have a certain impact on normal-hearing listeners [3]. Reverberation in speech affects human perception and poses a challenge to the robustness of identity authentication systems and speech recognition systems as well [4] [5]. The method to solve the reverberation problem will benefit many speech technology applications.

In order to meet the requirements of human ears for speech intelligibility, people use a variety of methods to study speech dereverberation. One method is to enhance or modify the linear prediction (LP) residuals of the short-term speech segments [6] [7]. In this method, the linear prediction coefficients (LPC) of a short-term reverberant speech segment is assumed to be equal to the ones of the corresponding clean speech segment, or can be estimated to a high accuracy through spatial averaging on reverberant speech segments collected at different positions in the room [8]. Then, the LP residual of the reverberant speech segment is processed through varied algorithms to get it approach that of the clean speech [7] [9] [10], and then the enhanced speech segment is obtained by stimulating the all-pole filter represented by the LP coefficients using the processed LP residual.

In recent years, with the broad application of Deep Neural Network (DNN) in varied signal processing areas, DNN has also been used for speech enhancement [11]. Using DNN to eliminate speech reverberation was firstly proposed in [12] [13]. The DNN was trained to learn the spectral mapping relationship between reverberant speech and clean speech. Subsequently, Wu et al. [14] used the reverberation time (RT60) [15] as known to select the appropriate frame shift time and frame extension length to further improve the DNN-based dereverberation method. A step of denoising was added to the DNN to suppress the noise as well as the reverberation in [16].

In this paper, we propose to use DNN to process the residual of the reverberant speech. Instead of using DNN to model the mapping from the short-term spectrum of the reverberant speech to that of the clean speech as above in [13] [14], the DNN in this method is trained to learn the relationship between the spectrum of the LP residual of the reverberant speech and that of the clean speech, and hereafter this method will be referred to as DNN-based linear prediction residual mapping (DNN-LPRM) method. Since the DNN mapping is nonlinear and it is well-known that nonlinear processing of speech signals can introduce in artifacts [17], in DNN-LPRM we have tried not to apply the direct speech-to-speech mapping, and only apply the mapping to the LP residual to avoid this problem to some extent.

The remainder of this paper is organized as follows. The DNN-LPRM algorithm will be described in Section II. Section III tests the algorithm and shows the simulating experiment results; different performance indicators are employed to verify the effectiveness of our algorithm in speech dereverberation. We summarize the work in Section IV.

II. ALGORITHM DESCRIPTION

The diagram of the DNN-based LP residual mapping dereverberation algorithm model is shown in Fig.1. In the training phase, the amplitude spectrum of LP residual of the short-term

This work was supported by the National Key R&D Program of China under Grant 2019YFE0196400, the National Natural Science Foundation of China under Grant 61871035.

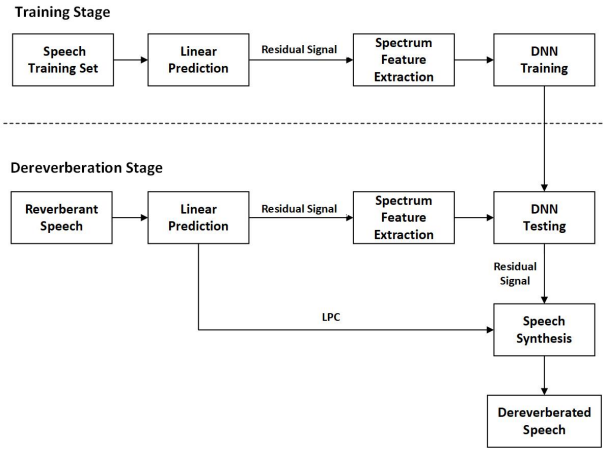

Fig. 1. Block diagram of the DNN-LPRM method.

speech is used as the speech feature. The amplitude spectrum of LP residual of reverberant speech is fed to the input layer of the network.

In the dereverberation stage, the trained DNN is provided with the amplitude spectrum of the LP residual of short-term testing reverberant speech, and an enhanced residual spectrum is obtained after passing through the network. Then, the enhanced amplitude spectrum of the residual is combined with the phase spectrum of the reverberant speech frame. After this the time-domain residual after enhancement can be obtained, and finally speech synthesis is performed to produce the time-domain dereverberated speech signal by using the enhanced residual to stimulate the all-pole filter represented by the LPC.

A. Linear Prediction and Residual Extraction

In a reverberant environment, the speech signal received by the microphone can be expressed as

$$y\left(n\right) = s\left(n\right) * h\left(n\right), \tag{1}$$

where $s(n)$ represents the clean speech signal, $h(n)$ represents the room impulse response (RIR), $*$ represents the convolution operator, and $y(n)$ is the reverberant speech signal.

The principle of the linear prediction is that the value of the current sample of a speech signal can be estimated by a linear combination of a few sample values ahead of it

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k), \tag{2}$$

where $\tilde{s}(n)$ is the predicted value of the nth sample, p is the order of the linear prediction system, and α_k are the LPC.

An error can be formed by subtracting the predicted value $\tilde{s}(n)$ from the true value $s(n)$

$$x\left(n\right) = s\left(n\right) - \tilde{s}\left(n\right). \tag{3}$$

The LPC α_k can be obtained by minimizing the mean-square error [18]. Then, by substituting the resulting α_k back into (2) and using (3), we can obtain the LP residual $x(n)$. The LP residual of the reverberant speech signal, $x_y(n)$, can be obtained in the same way.

Generally speaking, a speech signal is non-stationary, but short-term frames with a duration of, for example, 10-30 ms can be regarded as stable [18]. Therefore, when estimating the LPC and the LP residual, it is necessary to perform framing and windowing. In this study, a frame length of 16 ms and a Hamming window [18] are employed to frame the speech.

Let $X_n(e^{j\omega})$ denote the Short-Time Fourier Transform (STFT) [18] of the residual $x(n)$. The amplitude of $X_n(e^{j\omega})$, $|X_n(e^{j\omega})|$, along with $|X_n^y(e^{j\omega})|$ is used to train the DNN, where $|X_n^y(e^{j\omega})|$ represents the amplitude of the reverberant speech residual spectrum. In the dereverberation stage, let $|X_n^{y_p}(e^{j\omega})|$ denote the amplitude spectrum after mapping, and the phase term of $X_n^y(e^{j\omega})$, $\angle X_n^y(e^{j\omega})$, is used to form the frequency term of the residual after processing

$$|X_n^{y_p}(e^{j\omega})|e^{j\angle X_n^y(e^{j\omega})}. \tag{4}$$

Then (4) can be transform back to time domain using the inverse STFT to produce the enhanced residual.

B. DNN-Based LP Residual Mapping

In previous work which uses DNN to learn the direct mapping relationship between reverberant speech and clean speech, the amplitude spectrum of the short-term speech frame has been used as the speech feature [12] [14]. In this work, the amplitude spectrum of the LP residual is used as the speech feature to train the DNN.

The DNN model based on the Keras and Tensorflow framework is used [19]. The DNN contains 3 hidden layers, and all the layers are fully connected. Except for the output layer, the activation functions of all fully connected layers in the model use the LeakyReLU function [20]. LeakyReLU is a special version of Rectified Linear Unit (ReLU). When it is not activated, LeakyReLU will still have a non-zero output value, so as to obtain a small gradient to avoid the neuron "death" phenomenon that may occur in ReLU. In order to ensure that the final output is non-negative, the activation function of the output layer uses the Sigmoid function to normalize the DNN output of the spectral amplitude to the unit range of [0, 1] [21].

After each fully connected layer and before the activation function, batch normalization is set. The activation value of the previous layer is renormalized on each batch, that is, the average value of its output data is close to 0 and the standard deviation is close to 1. The purpose is to reduce the learning rate and improve the computing speed and performance of DNN training. This is the average variance normalization (MVN) strategy commonly used in speech recognition [22]. In addition, the dropout processing is added at the end of each layer except the output layer [23]. A certain percentage of input neurons are randomly disconnected each time the parameters are updated during the training process to prevent over-fitting. It further improves the DNN training performance.

The objective function is based on the minimum mean-square error (MMSE)

$$E = \frac{1}{MD}\sum_{m=1}^{M}\sum_{d=1}^{D}(\hat{X}_d^m - X_{d,\mathrm{norm}}^m)^2, \tag{5}$$

where E represents the mean-square error, $\hat{X}_d^m$ and $X_{d,\mathrm{norm}}^m$ represent the dth DNN output and normalized target feature at frame index m, D and M represent the size of the feature vector and mini-batch respectively. The DNN output and normalized target features are

$$\hat{X}_d^m = \sum\nolimits_{l=1}^{L} w_{l,d}R_l^m + b_d \tag{6}$$

$$X_{d,\mathrm{norm}}^m = \frac{X_d^m - \mu_d}{\sigma_d} \tag{7}$$

where $w_{l,d}$ and b_d represent the weights and biases between the last hidden layer and output layer respectively, with L denoting the last hidden layer size. R_l^m is the output at the lth neuron of the last hidden layer at frame m. μ_d and σ_d represent the global mean and variance of the target feature at frequency bin d over all target utterances.

III. EXPERIMENTS AND RESULT ANALYSIS

A. Parameters and Evaluation Indicators

The speech data used in the simulating experiments are from the TIMIT database [24]. The reverberant speech segments are obtained by convolving the clean speech with the RIRs generated using the image-source method [25] [26]. The room size is set to 6 m×4 m×3 m (length×width×height), and the speaker and microphone are located at [2, 3, 1] m and [4, 1, 2] m respectively. In the training set, 380 clean speech recordings are convolved with RIRs of reverberation time 0.3 s to 1.0 s with an interval of 0.1 s. In order to test the algorithm, we randomly selected 10 speech recordings from the TIMIT test set to convolve them with RIRs with reverberation times of 0.3s-1.0s, to generate 10×8=80 reverberant speech segments as a test data set.

In the linear prediction stage, the frame length is set to 16 ms, which corresponds to 0.016×16000=256 samples with the sampling rate of 16 kHz. In the windowing processing, the overlap rate between two adjacent frames is 1/2, and a Hamming window function with a length of 256 is used [18]. The order p of linear prediction is set to 12 as usually used in other speech processing techniques [8].

In the training and testing phases, 512-point STFT is used to generate 257-dimension amplitude spectrum features. A frame extension of 5 frames is set at the input layer of the DNN, that is, the input dimension is 257×5=1285. The DNN used has 3 hidden layers, each with 2048 nodes. The maximum number of training rounds is set to 20, the minimum batch size is set to 128, and the input and target features of the DNN are globally normalized to zero mean and unit variance [14]. In the model training process, the mean-square error (MSE) is used as the loss function, and the Adam optimizer is used to train the model [27].

In evaluation, the speech-to-reverberation modulation energy ratio (SRMR) [28], the frequency-weighted Segmented Speech Signal-to-Noise Ratio (fwSegSNR) [29], and the Perceptual Evaluation of Speech Quality (PESQ) [30] are used. These three objective evaluation indicators are the ones used in the REVERB challenge and proved to have high correlation scores with the subjective test results in terms of perceived amount of reverberation and of overall quality [1].

B. Experimental Results

Based on deep neural network, we compare the results of using linear prediction residual processing and not using linear prediction residual processing to verify the feasibility of the algorithm. In the experimental results, the clean speech is labeled with RAW, the reverberant speech is labeled with REV, the dereverberation results using DNN speech amplitude spectrum mapping [14] is labeled with DNN-SM, and the dereverberation result using our algorithm is labeled with DNN-LPRM.

Figure 2 shows an example of the residual signals of a test speech recording with RT60=0.5 s. The speech content

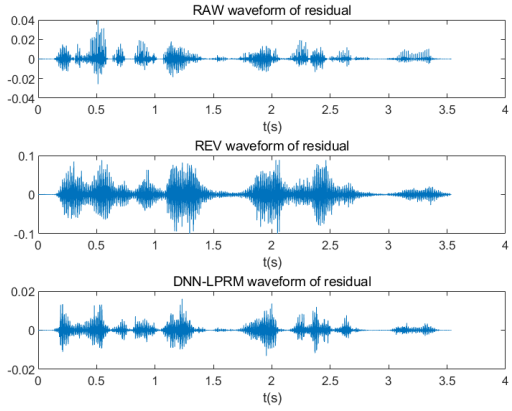

Fig. 2. The residuals of the clean speech, the reverberant speech, and the residual after DNN-LPRM processing.

is: “She had your dark suit in greasy wash water all year.” It can be seen in the figure that the residual after processing is closer to that of the clean signal, and the smearing caused by the strong peaks in the residual of the reverberant speech has been largely eliminated. Figure 3 (a) and Fig. 3 (b) show the spectrogram of RAW speech and REV speech. The spectrum mapping result of DNN-SM is shown in Fig. 3 (c), and that of DNN-LPRM is shown in Fig. 3 (d). In Fig. 3 (c), the tailing energy caused by reverberation is greatly attenuated, which indicates that the spectral mapping result of DNN is a good estimate of the clean spectrogram. Compared with Fig. 3 (c), the spectrogram in Fig. 3 (d) is clearer, and is closer to the spectrogram of clean speech in Fig. 3 (a), indicating that the linear prediction residual enhancement further improves the dereverberation performance.

We take the average of the SRMR evaluation results of 10 speech recordings under RT60=0.3-1.0 s. Figure 4 shows

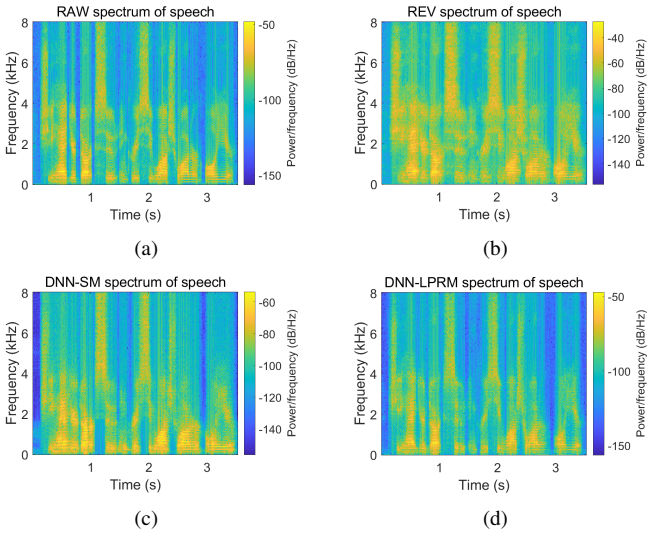

Fig. 3. The frequency spectrum of a testing speech recording with RT60 = 0.5 s: (a) RAW, (b) REV, (c) DNN-SM, (d) DNN-LPRM

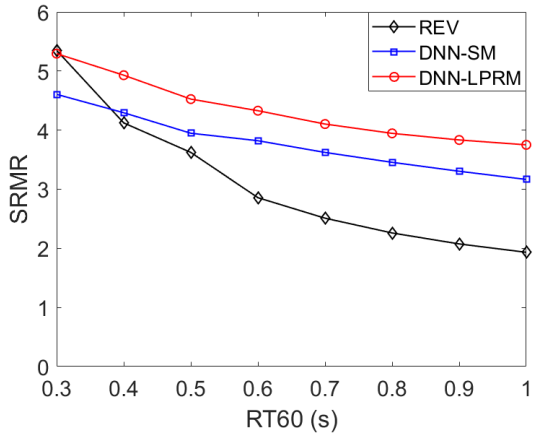

Fig. 4. SRMR results under RT60=0.3-1.0 s.

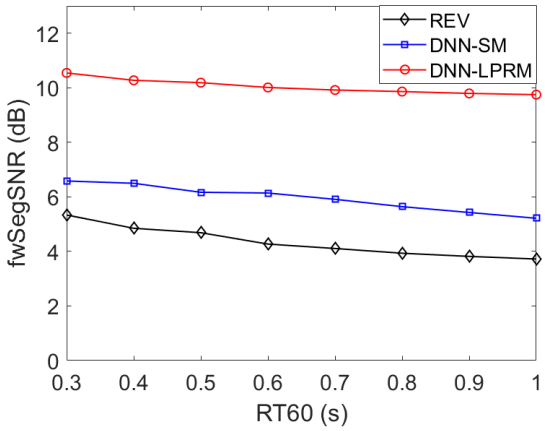

Fig. 5. Evaluation results of fwSegSNR under RT60=0.3-1.0 s.

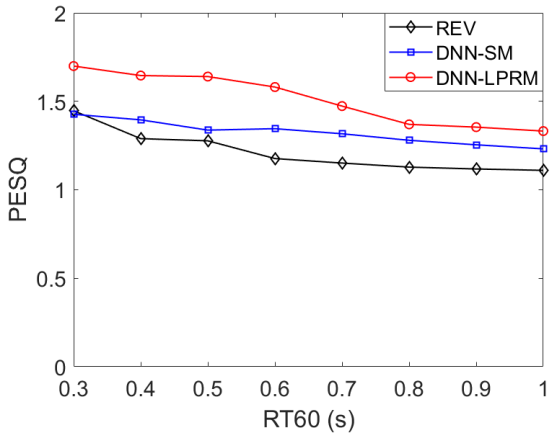

Fig. 6. Evaluation results of PESQ under RT60=0.3-1.0 s.

that the DNN-LPRM method has achieved a big improvement in SRMR compared with DNN-SM. On average, the SRMR value of the speech after DNN-LPRM processing is 1.24 higher than that of the reverberant speech, and 0.56 higher than that of the DNN-SM. Under low reverberation conditions, when RT60=0.3 and 0.4 s, the DNN-SM method does not significantly improve the SRMR, whereas the SRMR value of the DNN-LPRM method can be maintained at a relatively high level. And with the increase of the reverberation time, the SRMR value of the DNN-LPRM method does not drop significantly, and the DNN-LPRM always keeps the gap with the DNN-SM method in SRMR. It also shows that our DNN-LPRM method has high robustness under both low and high reverberation conditions.

The evaluation results of fwSegSNR and PESQ averaged over the same 10 speech recordings under RT60=0.3-1.0 s are shown in Fig. 5 and Fig. 6. Figure 5 shows that the DNN-LPRM method has achieved a significant improvement in the segmental signal-to-noise ratio. The fwSegSNR achieved with the DNN-LPRM is 5.7 dB higher than that of the reverberation speech and is 4.1 dB higher than that obtained with the DNN-SM method on average. With the increase of reverberation time, the value of fwSegSNR drops very little, and maintains at a high level, which shows that our DNN-LPRM method has high robustness in fwSegSNR under different reverberation conditions.

Figure 6 shows that the DNN-LPRM method has also achieved improvements in speech quality. The PESQ value is 0.3 higher than that of the reverberant speech and 0.2 higher than that obtained with the DNN-SM method on average. Please note that PESQ is an indicator of the general speech quality and was not devised specifically for evaluating the severeness of the reverberation in speech. We employed it here as a monitor of the change in speech quality brought about by the DNN-LPRM to ensure that no big decrease in speech quality is introduced by the algorithm, and the results show that small improvement in PESQ has been achieved.

Based on the above results, the performance of our DNN-LPRM method has greatly surpassed the DNN-SM method.

It has effectively removed the reverberation in the speech and improved the speech quality, and has high robustness under different reverberation conditions. We have also tested its robustness to different position of the source-microphone combination in the same room, and the results show that keeping the source-microphone distance unchanged, the different source-microphone position in the room brings very little change in the performance of the algorithm. Robustness of the algorithm to different source-microphone distance in different rooms will be investigated in further research. Comparison of DNN-LPRM with more algorithms using data such as the evaluation set for REVERB Challenge [1] will also be done further.

IV. CONCLUSION

This paper proposes a speech dereverberation algorithm based on linear prediction residual processing using deep neural network. Experimental results show that the algorithm has achieved significant improvements in various objective speech evaluation indicators, which proves the effectiveness of the DNN-LPRM in removing the reverberation in speech. The robustness of the algorithm to different levels of reverberation has also been tested, and it performs consistently well in different reverberant scenarios. The performance of the algorithm has been compared with the method using DNN mapping on the speech-level feature, and the results show that our algorithm surpasses the speech-level feature mapping algorithm by 4.1 dB in fwSegSNR.

REFERENCES

[1] K. Kinoshita, M. Delcroix, S. Gannot, E. Habets, R. Häb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the reverb challenge: state-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP Journal on Advances in Signal Processing*, vol. 2016, pp. 1–19, 2016.

[2] K. Kokkinakis, O. Hazrati, and P. Loizou, "A channel-selection criterion for suppressing reverberation in cochlear implants." *The Journal of the Acoustical Society of America*, vol. 129 5, pp. 3221–32, 2011.

[3] N. Roman and J. F. Woodruff, "Speech intelligibility in reverberation with ideal binary masking: effects of early reflections and signal-to-noise ratio threshold." *The Journal of the Acoustical Society of America*, vol. 133 3, pp. 1707–17, 2013.

[4] S. O. Sadjadi and J. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5448–5451, 2011.

[5] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 836–845, 2014.

[6] B. Yegnanarayana and P. S. Murthy, "Enhancement of reverberant speech using lp residual signal," *IEEE Trans. Speech Audio Process.*, vol. 8, pp. 267–281, 2000.

[7] N. D. Gaubitch, P. A. Naylor, and D. B. Ward, "Multi-microphone speech dereverberation using spatio-temporal averaging," in *2004 12th European Signal Processing Conference*, 2004, pp. 809–812.

[8] N. Gaubitch, P. Naylor, and D. Ward, "On the use of linear prediction for dereverberation of speech," in *Int. Workshop Acoust. Echo Noise Control*, 2003.

[9] B. Yegnanarayana, S. R. M. Prasanna, and K. S. Rao, "Speech enhancement using excitation source information," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2002, pp. I–541–I–544.

[10] C. Zheng, R. Peng, J. Li, and X. Li, "A constrained mmse lp residual estimator for speech dereverberation in noisy environments," *IEEE Signal Processing Letters*, vol. 21, pp. 1462–1466, 2014.

[11] Y. Xu, J. Du, L. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, pp. 65–68, 2014.

[12] K. Han, Y. Wang, and D. Wang, "Learning spectral mapping for speech dereverberation," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4628–4632, 2014.

[13] 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, pp. 982–992, 2015.

[14] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, pp. 102–111, 2017.

[15] H. Kuttruff, *Room acoustics*, 4th ed. Taylor & Frances, 2000.

[16] Y. Zhao, Z. Wang, and D. Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, pp. 53–62, 2019.

[17] 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.

[18] L. Rabiner and R. Schafer, "Introduction to digital speech processing," *Found. Trends Signal Process.*, vol. 1, pp. 1–194, 2007.

[19] C. Anindya, D. Purwanto, and D. I. Ricoida, "Development of indonesian speech recognition with deep neural network for robotic command," *2019 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 434–438, 2019.

[20] J. Xu, Z. Li, B. Du, M. Zhang, and J. Liu, "Reluplex made more practical: Leaky relu," *2020 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–7, 2020.

[21] B. Wu, K. Li, M. Yang, and C.-H. Lee, "A study on target feature activation and normalization and their impacts on the performance of dnn based speech dereverberation systems," *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–4, 2016.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[23] P. Nazreen and A. G. Ramakrishnan, "Dnn based speech enhancement for unseen noises using monte carlo dropout," in *2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, 2018, pp. 1–6.

[24] J. Garofolo, "Getting started with the darpa timit cd-rom: An acoustic phonetic continuous speech database," *national institute of standards & technology gaithersburgh md*, 1988.

[25] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[26] E. A. P. Habets. (2008, May) Room impulse response (RIR) generator. [Online]. Available: http://home.tiscali.nl/ehabets/rirgenerator.html

[27] Z. Zhang, "Improved adam optimizer for deep neural networks," *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pp. 1–2, 2018.

[28] T. Falk, C. Zheng, and W. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1766–1774, 2010.

[29] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions." *The Journal of the Acoustical Society of America*, vol. 125 5, pp. 3387–405, 2009.

[30] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, vol. 2, pp. 749–752 vol.2, 2001.