INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Enhancement with Wide Residual Networks in Reverberant Environments

Jorge Llombart, Dayana Ribas, Antonio Miguel, Luis Vicente, Alfonso Ortega, Eduardo Lleida

ViVoLab, Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{jllombg,dribas,amiguel,lvicente,ortega,lleida}@unizar.es

Abstract

This paper proposes a speech enhancement method which exploits the high potential of residual connections in a Wide Residual Network architecture. This is supported on single dimensional convolutions computed alongside the time domain, which is a powerful approach to process contextually correlated representations through the temporal domain, such as speech feature sequences. We find the residual mechanism extremely useful for the enhancement task since the signal always has a linear shortcut and the non-linear path enhances it in several steps by adding or subtracting corrections. The enhancement capability of the proposal is assessed by objective quality metrics evaluated with simulated and real samples of reverberated speech signals. Results show that the proposal outperforms the state-of-the-art method called WPE, which is known to effectively reduce reverberation and greatly enhance the signal. The proposed model, trained with artificial synthesized reverberation data, was able to generalize to real room impulse responses for a variety of conditions (e.g. different room sizes, RT_{60}, near & far field). Furthermore, it achieves accuracy for real speech with reverberation from two different datasets.

Index Terms: speech enhancement, reverberation, deep learning, wide residual neural networks, speech quality measures

1. Introduction

The high capability of the deep learning approaches for discovering underlying relations on the data has been exploited for speech enhancement tasks. Many interesting solutions for modeling the relationship between corrupted and clean data have been recently proposed based on a variety of DNN architectures. Convolutional Neural Network (CNN) based architectures have shown to effectively deal with the corrupted speech signal structure [1, 2]. Also, solutions based on Recurrent Neural Networks (RNN) architectures and the associated Long Short-Term Memory (LSTM) alternative have effectively been able to handle noisy and reverberant corrupted speech [3, 4, 5, 6, 7]. Both, convolutional and recurrent networks, have also appeared combined with residual blocks to further model the dynamic correlations among consecutive frames [8, 9, 10]. Residual connections make use of shortcut connections between neural network layers, allowing to handle deeper and more complicated neural network architectures, with fast convergence and a small gradient vanishing effect [11]. In this way, they are able to provide more detailed representations of the underlying structure of the corrupted signal.

This paper proposes a novel speech enhancement method based on the Wide Residual Neural Networks (WRN) architecture using single dimensional convolutional layers. This approach deals with reverberation in the spectral domain, making a regression from the log magnitude spectrum of reverberant speech to that of clean speech. In this way, it reinforces the importance of the low energy bands of the spectrum in the analysis, which have an impact on the perception of speech. We analyze the method performance through speech quality metrics from two viewpoints, namely the dereverberation level, and the spectral distortion introduced by the enhancement process. We compare the proposal performance with the state-of-the-art method called WPE in an experimental framework inspired by the REVERB Challenge task. This method is based on the LSTM architecture and has reported top performances in this framework [6].

So far, residual connections have been barely exploited for speech enhancement. In [8], the authors proposed an architecture using LSTM, and they briefly studied the performance of residual connections through testing different configurations in the framework of speech recognition. In [10], authors also proposed an architecture based on the recurrent approach but using gated recurrent units. This study also reports results in quality measures to assess the dereverberation. However, they use metrics associated to PESQ, which is actually not recommended as a metric for enhanced or reverberant speech [12]. In this line, our paper contributes to study the role of residual connections in deep speech enhancement solutions, assessing alternative conditions to previous studies.

Section 2 presents the proposal based on the WRN architecture, introducing the characteristics that make it interesting for speech enhancement. Section 3 describes the experimental setup. Section 4 shows results and discussion. Finally section 5 concludes the paper.

2. Proposal

The network architecture proposed (Figure 1) processes input features with a first convolutional layer followed by four Wide Residual Blocks (WRB). The first WRB processes the output of the first convolutional layer and also its input. Following the WRBs there is a Batch Normalization (BN) stage and a nonlinearity (PReLU: Parametric Rectified Linear Unit). The combination of BN and PReLU blocks allows a smoother representation in regression tasks than the combination with ReLU. Finally, there is another convolutional layer with a ReLU, to reduce the number of channels to 1 and obtain the enhanced log-spectrum for each signal. Every WRB increases the number of channels used to get the outputs. The widening operation is done in the first convolution of the first residual block of each WRB.

In order to compute the residual connection as a sum operation, the number of channels in the straight path and in the convolution path has to be the same. Therefore, when the number of channels is increased, a Conv1D with $k = 1$ is added.

http://dx.doi.org/10.21437/Interspeech.2019-1745

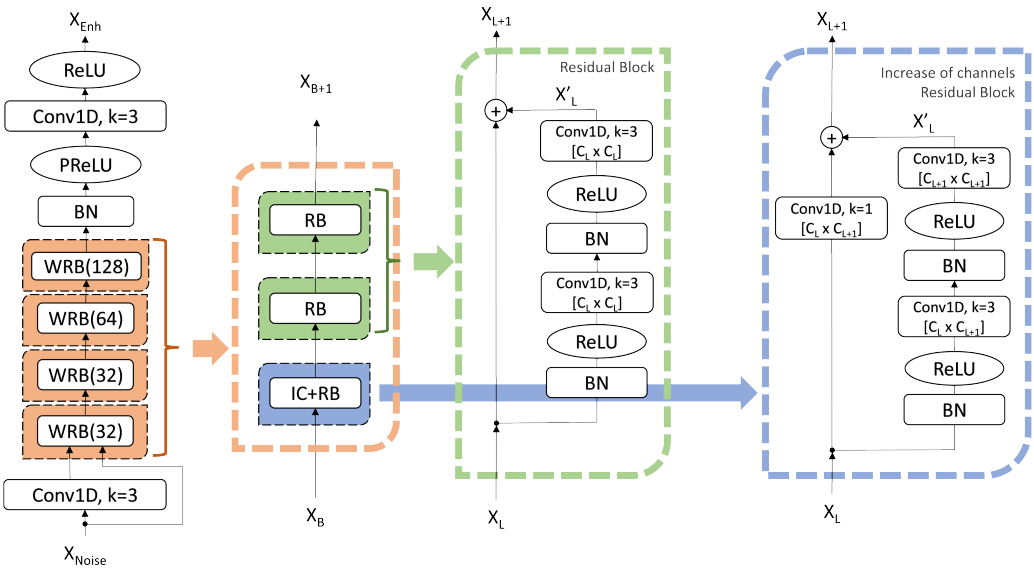

Figure 1: *WRN architecture proposed. From left to right there is the composition of the network blocks, with C_L the number of channels in the layer L.*

This can be interpreted as a position wise fully connected layer to adjust the number of channels from the residual path to the number of channels in the convolutional path in order to add them.

In this work, we want to enhance the logarithmic spectrum of a noisy input signal X_{Noise}. For this objective we use the Mean Square Error (*MSE*) in the training cost function to get an enhanced signal X_{Enh} from X_{Noise} as similar as possible to the clean reference Y. From the experience in our previous work [13], Instead of using a frame by frame enhancement, we process the whole input signal as a sequence. This means that instead of providing for each example the regression error of one frame, we propagate the accumulated error of the regression along the complete sentence. This strategy considerably reduces the computation because instead of generating hundreds of training examples from one input signal, each training example is a complete input sequence. Finally, the cost function is the mean of all input frames *MSE* described in equation (1)

$$J(Y, X_{Enh}) = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{N}\sum_{n=0}^{N-1} MSE(y_{t,n}, x_{Enh,t,n}) \quad (1)$$

where T is the number of frames of the example, N is the feature dimension, $y_{t,n}$ are Y frames and $x_{Enh,t,n}$ are X_{Enh} frames.

3. Experimental setup

The experimental framework developed in this work is inspired by the REVERB Challenge task[1]. We evaluate the performance of speech enhancement methods through speech quality measures, aiming to find a trade-off between the dereverberation and the introduction of spectral distortion with the enhancement.

3.1. Datasets

Approaches were tested on the official Development and Evaluation sets of the REVERB Challenge [14]. The dataset has simulated speech from the convolution of WSJCAM0 Corpus [15] with three measured Room Impulse Responses (RIR) ($RT_{60} = 0.25, 0.5, 0.7s$) at two speaker-microphone distances: far (2 m) and near (0.5 m). It was also added stationary noise recordings from the same rooms (SNR = 20 dB). Besides, it has real recordings, acquired in a reverberant meeting room ($RT_{60} = 0.7s$) at two speaker-microphone distances: far (2.5 m) and near (1 m) from the MC-WSJ-AV corpus [16]. We also used real speech samples from VoiceHome v0.2 [17] and v1.0 [18]. VoiceHome was recorded in a real domestic environment, such that the background noise is that typically found in households e.g. vacuum cleaner, dishwashing or interviews on television.

For training the DNN we used 16 kHz sampled data from the following datasets: Timit [19], Librispeech [20], and Tedlium [21]. This data was augmented by adding artificially generated RIR ($RT_{60} = 0.05 - 0.8s$), stationary and non-stationary noises from Musan dataset [22], SNR = 5-25 dB, including music and speech, and scaling the time axis at the feature level.

3.2. Methods for comparison

We compare the performance of the proposed WRN speech enhancement method with the state-of-the-art dereverberation method called Weighted Prediction Error (WPE), which is known to effectively reduce reverberation in the framework of the REVERB dataset [6]. We used the more recent version of WPE[2] which is also based on DNN [6]. However, WPE uses an architecture based on LSTM, which also provides us the possibility for comparing the speech enhancement solutions from the DNN architecture point of view.

3.3. Performance assessment

The speech quality was measured in terms of the distortion introduced by the enhancement process by means of the Log-

[1] http://reverb2014.dereverberation.com

[2] https://github.com/fgnt/nara_wpe

likelihood ratio[3] (LLR) [23]. This was computed in the active speech segments (determined by a Voice Activity Detection (VAD) algorithm [24]). For this measure, the closer the target to the reference, the lower the spectral distortion, therefore smaller values indicate better speech quality.

On the other hand, we assess the reverberation level of the signal through the Speech-to-reverberation modulation energy ratio (SRMR) [25]. In this case, higher values indicate better speech quality. Note that only SRMR can be used with real data because LLR is computed using the observed/enhanced signal and clean reference.

3.4. Network configuration

The front-end starts segmenting speech signals in 25, 50, and 75 ms Hamming-windowed frames, every 10 ms. We provide this multiple representations of the input in order to maintain as much reverberant impulsive response inside the Hamming window as it is possible, without losing temporal resolution of the acoustic events. For each frame segment, three types of acoustic feature vectors are computed and stacked, to create a single input feature vector for the network: 512-dimensional FFT, 32, 50, 100-dimensional Mel filterbank, and cepstral features (same dimension of the corresponding filterbank). Finally, each feature vector is normalized by variance. Input features were generated and augmented on-the-fly, operating in contiguous vector blocks of 200 samples so that convolutions in the time axis can be performed. The network uses four WRN blocks with a widen factor of 8. AdamW algorithm was used to train the network and PReLUs [26] as parametric non-linearity.

4. Results and discussion

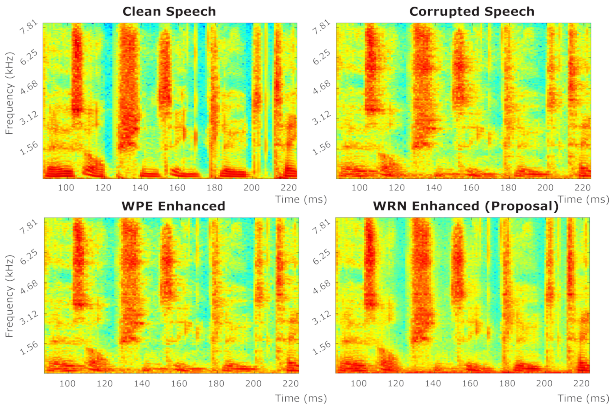

Figure 2: *Qualitative example of the enhancement performance in a single signal of REVERB Dev dataset.*

Figure 2 shows a qualitative example of the enhancement performance in the signal c31c0204.wav of the REVERB Dev dataset (RT_{60} = 0.25s, Distance speaker-mic = 200 cm). Observe the distortion due to reverberation in the corrupted speech spectrogram at the top-right side of the figure. Reverberation provokes a remarkable temporal spread of the power spectrum in active speech segments. Note that the enhanced speech through WPE removes some of this effect, but the WRN method is more accurate in this aim, achieving a better reconstruction of the signal.

[3]Originally known as Itakuta distance

4.1. Speech quality for processing tasks

Table 1 presents speech quality results in terms of distortion with the LLR distance for simulated speech samples. The first row corresponds to the reverberant unprocessed speech, which is compared to the quality achieved by the enhanced signals using WPE and the proposal WRN enhancement method. Both DNN-based methods are able to enhance the corrupted speech data, but the proposal outperforms WPE in terms of spectral distortion.

Table 1: *LLR distance in simulated reverberated speech samples from REVERB Dev & Eval datasets.*

Methods	REV-Dev	REV-Eval
Unprocessed	0.63	0.64
WPE [27]	0.60	0.60
WRN	0.50	0.51

4.2. Speech quality for dereverberation: Simulated vs. Real

Table 2 shows the average of SRMR results over the evaluated conditions for simulated and real speech samples. The first column on the left corresponds to the unprocessed speech data. Shadowed cells highlight the best results for each dataset.

Table 2: *Speech quality through SRMR results for simulated and real reverberated speech samples.*

Datasets	Unprocessed	WPE [27]	WRN
Simulated			
REVERB Dev	3.67	3.90	4.75
REVERB Eval	3.68	3.91	4.63
Real			
REVERB Dev	3.79	4.17	4.79
REVERB Eval	3.18	3.48	4.20
VoiceHome v0.2	3.19	3.28	5.03
VoiceHome v1.0	4.51	4.96	5.92

The proposal outperforms baselines for all datasets evaluated. The consistency in performance through different datasets supports the robustness of the method. This indicates that its parameters are not adjusted to some specific set of speech signals, which is a desirable quality for an enhancement method. These positive results beyond simulations, encourage the use of the method in realistic scenarios. Furthermore, note the WRN model was trained with artificially synthesized reverberation, however, it showed to be effectively dealing with a reverberated speech from real-world scenarios.

4.2.1. Room sizes and reverberation level

Figure 3 shows the evolution of SRMR results with the increase of reverberation level for different room sizes: $Room1 - RT_{60} = 0.25s$, $Room2 - RT_{60} = 0.5s$, and $Room3 - RT_{60} = 0.75s$. The proposed WRN method achieves higher speech quality than the reference for all conditions evaluated. Furthermore, the results of the proposed method have less variability through the RT_{60}, indicating the robustness of the method in different scenarios. See that the speech quality improvement with respect to the reference methods increases with the RT_{60}. However, note that there is less space for improvement in the $Room1 - RT_{60} = 0.25s$ condition, so it is harder to enhance.

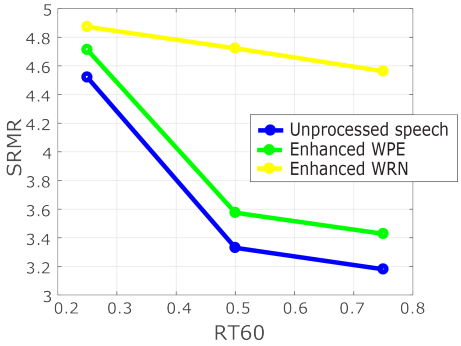

Figure 3: *Speech quality through SRMR measure for different reverberation levels in simulated reverberated speech samples from REVERB Dev & Eval datasets.*

4.2.2. Near and Far field

Figure 4 presents an average of SRMR results for far (250 m) and $near$ (50 m) conditions in the simulated REVERB Dev and Eval datasets. WRN considerably outperformed the WPE baseline for $34,88\%$ in far-field and $8,44\%$ in near-field. Note that the proposal is strongest in the far-field conditions, which is usually the most challenging scenario.

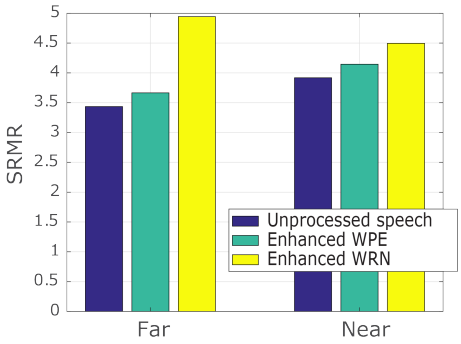

Figure 4: *SRMR results for near- and far-field simulated reverberated speech from REVERB Dev & Eval datasets.*

4.3. Effect of train-test mismatch

As we saw before, the conditions for $Room1, RT_{60} = 0.25$ and $near$ speaker-microphone distance were harder to enhance. These cases correspond to low reverberation, where there is a small margin for improvement. Hence, to increase the focus of the data augmentation for the network training to these conditions could provide a boost of performance. Furthermore, note that due to the lack of exact room size values in the test dataset description, WRN data training included a reasonable estimation of small room size. However, this is probably not small enough for $Room1$. On the other hand, the distance configuration of the training data design considers the speaker/microphone can be randomly situated all-around the room, modeling it with uniform data distribution. This left low probability for the specific test data distances of $near$ (50 cm) and far (250 cm). In order to improve the results in these scenarios, the training data in future approaches should include smaller room sizes, change the function which models the speaker-microphone distances, in order to increase the probability of certain distances. However, caution should be taken with overfitting of the training data to some specific dataset. A better compromise between the network generalization and the test data characteristics will be a more reasonable solution.

4.4. Comparison with previous work

Experimental results evidenced that the proposed WRN architecture outperformed the reference WPE. Despite the potentiality of the RNN-LSTM architecture used in WPE, the combination of CNN with residual connections in the proposal was able to obtain more expressive representations of the reverberant speech. This structure performs the enhancement taking into account the full utterance through convolutions in all temporal domain of the signal, which is higher while deeper is the structure. As WPE is based on RNN-LSTM it only takes into consideration previous context. However, for the enhancement purpose, a representation that considers a context including some future samples may contribute to increasing the overall performance. The proposed WRN architecture implements this idea through the convolutional layers.

With our proposal, the reconstruction of the clean signal achieved improved speech quality more than the reference, with a proper trade-off between the level of dereverberation, and the amount of spectral distortion. These results were also validated through a test in real distorted speech, to show the generalization capability of the model.

5. Conclusions and future work

This paper has introduced a novel speech enhancement method based on a WRN architecture that takes advantage of the powerful representations obtained from a wide topology of CNN with residual connections. Results showed the potentiality of the WRN providing an enhanced speech on top of the state-of-the-art RNN-LSTM-based method called WPE. Best results were obtained for far-field reverberated speech in three different room sizes. The residual mechanism was extremely useful in this case since the signal has always a linear shortcut and the non-linear path enhances it in certain steps by adding or subtracting corrections. In practical applications, this is a valuable property because realistic scenarios could challenge the system with many different conditions [28].

Despite results are encouraging the proposal can be further improved. Future work will focus on fine-tuning the data training configuration with a view to updating the compromise between generalization and accuracy. We also plan to expand the experimental setup to evaluate in speech recognition task with speech data in alternative scenarios from other datasets and comparative baselines. On the other side, the inclusion of perceptual features in the network cost function will be explored in order to improve the performance in the speech reconstruction process.

6. Acknowledgment

Funded by the Government of Aragon (Reference Group T36_17R) and co-financed with Feder 2014-2020 "Building Europe from Aragon". This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. This material is based upon work supported by Google Cloud.

7. References

[1] S.-W. Fu, Y. Tsao, and X. Lu, "SNR-aware convolutional neural network modeling for speech enhancement," in *Interspeech*, 2016, pp. 3768–3772.

[2] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *Interspeech*, 2017, pp. 1993–1997.

[3] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent Neural Networks for Noise Reduction in Robust ASR," in *Interspeech*, 2012.

[4] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. L. Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*, 2015, pp. 91–99.

[5] J. Chen and D. Wang, "Long short-term memory for speaker generalization in supervised speech separation," in *Interspeech*, 2016.

[6] K. Kinoshita, M. Delcroix, H. Kwon, T. Mori, and T. Nakatani, "Neural network-based spectrum estimation for online WPE dereverberation," in *Interspeech*, 2017, pp. 384–388.

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

[8] Z. Chen, Y. Huang, J. Li, and Y. Gong, "Improving mask learning based speech enhancement system with restoration layers and residual connection," in *Interspeech*, 2017, pp. 3632–3637.

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

[10] J. F. Santos and T. H. Falk, "Speech dereverberation with context-aware recurrent neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 7, pp. 1236–1246, 2018.

[11] S. Zagoruyko and N. Komodakis, "Wide residual networks," *CoRR*, vol. abs/1605.07146, 2017. [Online]. Available: https://arxiv.org/abs/1605.07146

[12] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs." *Rec. ITU-T P. 862*, 2001.

[13] J. Llombart, A. Miguel, A. Ortega, and E. Lleida, "Wide residual networks 1d for automatic text punctuation," *IberSPEECH 2018*, pp. 296–300, 2018.

[14] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, E. Habets, R. Haeb-Umbach, V. Leutnant, A. Sehr, W. Kellermann, R. Maas, S. Gannot, and B. Raj, "The REVERB Challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-13)*, 2013.

[15] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJ-CAMO: a British English speech corpus for large vocabulary continuous speech recognition," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 1995, pp. 81–84.

[16] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Proceedings of the 2005 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-05)*, 2005, pp. 357–362.

[17] N. Bertin, E. Camberlein, E. Vincent, R. Lebarbenchon, S. Peillon, ric Lamand, S. Sivasankaran, F. Bimbot, I. Illina, A. Tom, S. Fleury, and ric Jamet, "A french corpus for distant-microphone speech processing in real homes," in *Interspeech*, 2016.

[18] N. Bertin, E. Camberlein, R. Lebarbenchon, E. Vincent, S. Sivasankaran, I. Illina, and F. Bimbot, "VoiceHome-2, an extended corpus for multichannel speech processing in real homes," *Speech Communications*, vol. 106, pp. 68 – 78, 2019.

[19] 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.

[20] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[21] A. Rousseau, P. Deléglise, and Y. Esteve, "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks." in *LREC*, 2014, pp. 3935–3939.

[22] D. Snyder, G. Chen, and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus," 2015, arXiv:1510.08484v1.

[23] P. C. Loizou, *Speech Quality Asssessment. In: Multimedia Analysis, Processing and Communications*. Springer, 2011, pp. 623–654.

[24] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271 – 287, 2004.

[25] T. H. Falk, C. Zheng, and W.-Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Transaction in Audio, Speech and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] L. Drude, J. Heymann, C. Boeddeker, and R. Haeb-Umbach, "NARA-WPE: A Python package for weighted prediction error dereverberation in Numpy and Tensorflow for online and offline processing," in *13. ITG Fachtagung Sprachkommunikation (ITG 2018)*, Oct 2018.

[28] D. Ribas, E. Vincent, and J. R. Calvo, "A study of speech distortion conditions in real scenarios for speech processing applications," in *IEEE Spoken Language Technology Workshop (SLT)*, 2016.