INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Real-time single-channel deep neural network-based speech enhancement on edge devices

Nikhil Shankar, Gautam Shreedhar Bhat, and Issa M.S Panahi

Department of Electrical and Computer Engineering, The University of Texas at Dallas, Richardson, TX-75080, USA
Nikhil.Shankar@utdallas.edu

Abstract

In this paper, we present a deep neural network architecture comprising of both convolutional neural network (CNN) and recurrent neural network (RNN) layers for real-time single-channel speech enhancement (SE). The proposed neural network model focuses on enhancing the noisy speech magnitude spectrum on a frame-by-frame process. The developed model is implemented on the smartphone (edge device), to demonstrate the real-time usability of the proposed method. Perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI) test results are used to compare the proposed algorithm to previously published conventional and deep learning-based SE methods. Subjective ratings show the performance improvement of the proposed model over the other baseline SE methods.

Index Terms: speech enhancement, neural networks, real-time, smartphone.

1. Introduction

The presence of background noise in a speech signal degrades the quality and intelligibility of the speech. Speech enhancement (SE) focusses on eliminating or suppressing the unwanted noise from the desired signal. SE plays a crucial role in many applications such as speech recognition, speech communication systems, and hearing aids. Several conventional and neural network-based SE algorithms have been proposed in the past decade.

The Boll [1] approach for spectral subtraction focuses on the subtraction of the noise magnitude spectrum from the noisy speech magnitude spectrum. At the high signal to noise ratio (SNR) levels, statistical model-based approaches developed by Ephraim and Malah [2, 3] were effective in eliminating background noise. In [4, 5], Maximum A Posteriori (MAP) estimation based computationally efficient SE algorithms are proposed. However, such conventional SE methods are based on some premises and cannot be successful for non-stationary forms of background noise. Some of the statistical-based single-channel SE methods also introduce speech distortion in the form of musical noise, especially at low SNR.

Based on the recent progression in deep neural networks (DNN) for different signal processing tasks, several deep learning methods for single-channel SE have been developed. The supervised SE methods are divided into masking and mapping-based techniques depending on the description of the clean speech targets for training [6]. In [7], the ideal binary mask (IBM) from noisy input speech is estimated by a feed-forward neural network. Compared to mask-based techniques, signal-based approximations reduce the difference between predicted and target gain [8]. DNN-based SE framework proposed in [9] predicts the clean speech log-power spectra (LPS) from noisy speech input LPS features. Recent innovations in the convolutional neural network (CNN) make them beneficial for SE to train the model using spectrogram features [10]. In [11], a fully convolutional neural network (FCN)-based SE is proposed with input raw audio data. Recurrent neural network (RNN) layers and long short-term memory (LSTM) layers are implemented to perform SE [12 - 14]. A mixture of convolutional and LSTM networks [15] outperforms other neural networks for SE at lower SNRs. In general, RNN layers are much more complex than CNN layers as they do not have weight sharing. However, RNNs are most suitable for time series data, as they can be used for processing random input data sequences with their internal memory.

In this paper, we propose a novel framework for real-time single-channel SE on edge devices, where a convolutional recurrent neural network (CRNN) model is trained to predict the clean speech magnitude spectrum. Also, the CRNN is computationally efficient and can be used for real-time processing [16]. A smartphone with an inbuilt microphone is used as an edge device example to capture the noisy speech data and perform complex computations using the proposed SE algorithm. The enhanced speech signal from the developed model implemented on the smartphone can be transmitted through wired or wireless earphone connection to the user [17, 18]. The paper provides a detailed description of the real-time implementation on the smartphone. The proposed algorithm can run on any stand-alone platform such as a smartphone and will serve as a critical element in the signal processing or communication pipeline. To prove the real-time implementation and application of the proposed method, we evaluate the developed model using a variety of noise types (both stationary and non-stationary), a wide range of speakers, and SNRs. The objective evaluations and subjective test results conducted for the proposed CRNN based SE method demonstrate the operational potential of the methodology developed.

2. Proposed algorithm description

This section describes the proposed SE algorithm. Figure 1 shows the block diagram of the proposed SE pipeline.

2.1. Problem formulation

We consider noisy speech $y(n)$ to be an additive mixture model of clean speech $s(n)$ and noise $d(n)$.

$$y(n) = s(n) + d(n) \quad (1)$$

http://dx.doi.org/10.21437/Interspeech.2020-1901

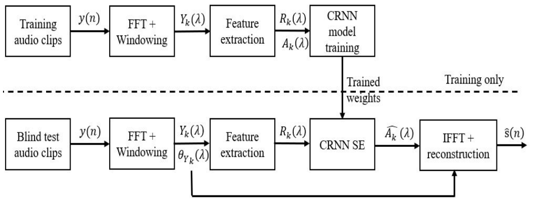


Figure 1: *Proposed SE pipeline.*

The input noisy speech signal is transformed into the frequency domain by taking short-time Fourier transform (STFT).

$$Y_k(\lambda) = S_k(\lambda) + D_k(\lambda) \tag{2}$$

$Y_k(\lambda)$, $S_k(\lambda)$, and $D_k(\lambda)$ represent the STFT of $y(n)$, $s(n)$ and $d(n)$ respectively for the frame λ and frequency bin k. In polar coordinates, (2) can be written as,

$$R_k(\lambda)e^{j\theta_{Y_k}(\lambda)} = A_k(\lambda)e^{j\theta_{S_k}(\lambda)} + B_k(\lambda)e^{j\theta_{D_k}(\lambda)} \tag{3}$$

Where $R_k(\lambda)$, $A_k(\lambda)$, $B_k(\lambda)$ are the magnitude spectrum of noisy speech, clean speech, and noise respectively. $\theta_{Y_k}(\lambda)$, $\theta_{S_k}(\lambda)$, $\theta_{D_k}(\lambda)$ represents the phase of noisy speech, clean speech, and noise respectively.

2.2. Feature selection

For effective neural network training, it is necessary to select suitable features. We consider magnitude spectrum as the input feature. The proposed CRNN system is trained with the noisy speech magnitude spectrum $R_k(\lambda)$ as input and the clean speech spectrum $A_k(\lambda)$ as the output label. Hence, the proposed model focuses on estimating a speech spectrum $\widehat{A_k}(\lambda)$. We consider the noisy phase for reconstruction. Finally, the estimate of clean speech for reconstruction is,

$$\widehat{S_k}(\lambda) = \widehat{A_k}(\lambda)\, e^{j\theta_{Y_k}(\lambda)} \tag{4}$$

The time domain signal output is obtained by taking Inverse Fast Fourier Transform (IFFT) of $\widehat{S_k}(\lambda)$.

2.3. Proposed CRNN architecture

CNNs process the input image or matrix by performing convolution and pooling functions. In CNNs, a small image region can be compacted by a series of weighted learning filters (kernels) to form a convolutional layer. The kernel generates a feature map for every forward pass of input. Maxpooling layers follow the convolution layers to reduce the size or dimension of the feature maps. Compared to CNN, RNNs permit us to model sequential data since they have feedback connections. The RNN cell has a dynamic behavior to make use of its internal state memory for processing. Thus, making it very reliable for speech analysis.

The proposed model is a combination of both CNN and RNN layers [19]. The model takes in one frame of noisy speech magnitude spectrum and outputs one frame of enhanced/clean speech magnitude spectrum. The input noisy magnitude spectrum is reshaped to form an image input, due to the presence of convolutional layers at the start. This is then fed into a neural network twice as shown in Figure 2. Figure 2 shows the block diagram representation of the proposed CRNN architecture. Different hidden layers such as convolutional layers, maxpool layers, long short-term memory (LSTM) layers, and fully connected (FC) layers are used to design the proposed model. There are 4 convolutional layers with a maxpool layer in between them. The first, second, third, and fourth convolutional layer uses 257, 129, 65, and 33 feature maps respectively. The feature maps gradually decrease in order to reduce the computational complexity and number of parameters, making the developed model suitable for real-time applications. The kernel and bias for all the convolution layers is given in Table 1. Followed by the convolutional layers, there are two LSTM layers consisting of 33 neurons each. The output of the LSTM layer is flattened out and the respective outputs from both the paths are added together before sending them to the FC layer. The FC hidden layer has 257 neurons and is followed by a linear output layer to predict the speech spectrum. The CRNN architecture proposed is given in Table 1. The specific numbers for designing the CRNN model was fixed after several experiments and training.

Table 1: *Architecture of the proposed CRNN model.*

Layers	Kernel	Bias
Conv 2D_1	5 × 5 × 1 × 257	257
Conv 2D_2	5 × 5 × 257 × 129	129
Conv 2D_3	5 × 5 × 129 × 65	65
Conv 2D_4	5 × 5 × 65 × 33	33
LSTM	33 × 132	264
Dense_1	1089 × 257	257
Dense_2	257 × 257	257

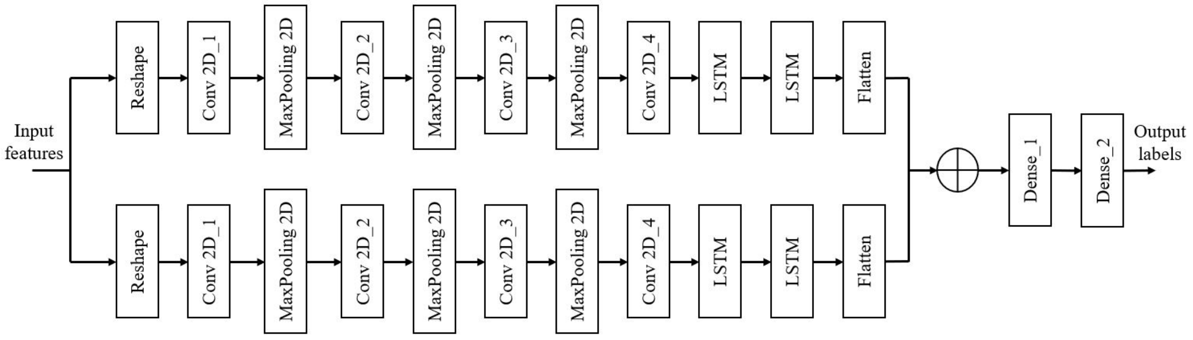


Figure 2: *Block diagram of the proposed CRNN architecture.*

Adam optimization algorithm [20] was used with a mean absolute error loss function to train the model.

Activation functions are used in each hidden layer to allow the network to learn complex and non-linear functional mapping between the input and output labels. We selected rectified linear unit (ReLU) as activation function because it has been successful in solving the vanishing gradient problem [21].

3. Experimental evaluation and results

3.1. Dataset and experimental setup

For the training and evaluation of the proposed CRNN model, a clean speech dataset is built from the Librivox dataset [22] of public audiobooks. Librivox has individual recordings in several languages, most of them are in English, that is read over 10,000 audio public domain books. Overall, there are 11,350 speakers present in the dataset. A portion of this dataset is considered to generate the noisy speech input features and clean speech labels for training the model. The noise dataset from Audioset and Freesound is considered [23]. Audioset is a series of approximately two million ten seconds sound clips made of YouTube videos, belonging to 600 audio classes. Finally, 150 audio classes, 60000 noise clips from Audioset, and 10000 noise clips from Freesound are mixed with the clean speech dataset considered. The resulting noisy speech audio clips are sampled to 16 kHz before feature extraction. A total of 100 hours of clean speech and noisy speech constitutes the training set. The clean speech files are normalized, and each noise clip is scaled up to have one of the five SNRs (0, 10, 20, 30, 40 dB). We randomly pick a clip of clean speech and noise, before combining them together to create a noisy speech clip. Due to the real-time application of the proposed method, reverberation is added to a portion of clean speech (30 hours) [24]. The reverberation time (T60) is randomly drawn from 0.2 s to 0.8 s with a step of 0.2 s. The proposed model is trained using the entire training dataset and we evaluate the model once the training is complete using a blind validation test set made available in [23]. The blind test set consists of real noisy speech recordings with and without reverberation. Challenging non-stationary noise cases were included in the blind set such as Multi-talker babble, keyboard typing, a person eating chips, etc. The blind test set comprises of 150 noisy speech clips.

The audio clips are sampled at 16 kHz with a frame size of 32ms with a 50% overlap. A 512-point STFT is computed to determine the input magnitude spectrum features. The first 257 magnitude spectrum values are taken into consideration due to the complex conjugate property in STFT and reshaped to form an image of 257 × 1 × 1. The final output layer predicts the clean speech signal magnitude spectrum. The model is trained for a total of 50 epochs.

3.2. Objective and subjective test results

The blind test set explained in the previous section is considered for speech quality and intelligibility evaluations. Perceptual evaluation of speech quality (PESQ) [25] and short-time objective intelligibility (STOI) [26] scores are included in the paper. PESQ ranges between -0.5 and 4.5, 4.5 is for high-quality speech. An increase in the STOI score increases the intelligibility. The test scores are the average of 150 audio clips present in the blind test set (validation data – not seen by the model). The proposed CRNN-based SE is compared with noisy speech and two baseline SE methods. Classical single microphone SE method i.e. log-MMSE [3] and RNN-based SE [27] are implemented and tested for our comparison. We also compare the number of parameters present in both models. Table 2 presents the objective test results and the proposed method outperforms the noisy speech and the other two SE approaches considered.

In addition to the objective evaluation, we conducted the mean opinion score (MOS) tests on 10 subjects. 10 random audio clips from the objective evaluation test set are considered for subjective testing. The subjects were instructed to score the noisy speech, proposed enhanced speech, log-MMSE output, and RNN SE output speech. The scoring is in the range 1 to 5 and the instructions are based on the following criteria: 5 for excellent speech quality and an imperceptible degree of distortion. 4 for decent speech quality with minimum distortion. 3 for providing enough degree of distortion for equivalent quality of speech. 2 for the low quality of speech with plenty of residual noises and distortions. 1 with the lowest speech content and an unacceptable level of distortion. Table 3 displays the average subjective test results and the findings indicate the usefulness of the proposed SE method. A comprehensive description of the scoring procedure is explained in [28].

Table 2: *Comparison of PESQ and STOI scores.*

Method	**Number of Parameters**	**PESQ**	**STOI (%)**
Noisy	-	2.18	89.1
logMMSE [3]	-	2.29	85.4
RNN [27]	61.2 K	2.42	89.5
Proposed	2.58 M	**2.57**	**91.3**

Table 3: *Subjective MOS test results.*

Method	**MOS**
Noisy	2.45
logMMSE [3]	2.93
RNN [27]	3.15
Proposed	**3.87**

Table 4: *CRNN vs CNN.*

Method	**Number of Parameters**	**PESQ**	**STOI (%)**
CNN	2.54 M	2.4	90
Proposed	2.58 M	**2.57**	**91.3**

3.3. Comparison of CRNN with CNN

In this experiment, we compare the performance of CRNNs with CNNs. Two different models are evaluated and compared using the magnitude spectrum input features. We propose a baseline CNN model, wherein the entire architecture is the same as that of the proposed CRNN model without the presence of the LSTM layers. The training data explained in the previous subsection remained the same for training the CNN model. Table 4 shows the PESQ and STOI scores of both the CRNN and CNN models. The same blind test set is used for evaluating and comparing both models. The presence of additional gates and the capability to use time-series information in the LSTM layer results in better noise suppression.

4. Real-time implementation

The proposed CRNN based SE algorithm is implemented on an iPhone. However, due to the real-time usability of the proposed application, it can be implemented on any processing platform. The microphone on the smartphone captures the input noisy speech at a 48 kHz sampling rate and then we downsample it to 16 kHz with the help of a low-pass filter and a decimation factor of 3. The input frame size is set to be 32ms. Figure 3 shows the screenshot of the proposed method implemented on the iPhone. By pressing the SE button present in the figure, the implemented model is initialized. The application simply replays the audio on the smartphone without processing when the ON/OFF switch is in off mode. By clicking on the ON/OFF switch button, the CRNN based SE module will process the input audio stream and suppress the background noise. A slider is provided to the smartphone user to control the amount of output volume.

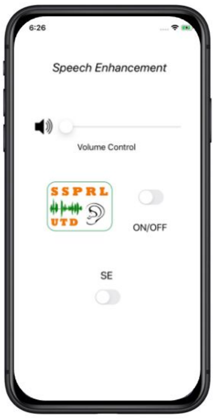
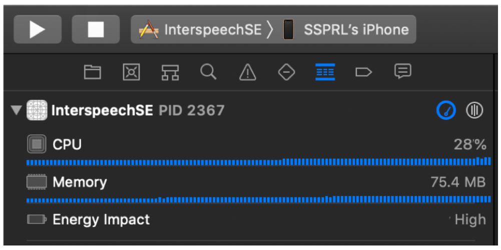


Figure 3: *Screenshot of the developed CRNN application.*

To run deep learning models on the smartphone, TensorFlow Lite offers a C/C++ API [29]. The proposed model is compressed and deployed on the smartphone using libraries such as the TensorFlow Lite converter and interpreter. The trained weights are frozen, thus eliminating backpropagation, training, and regularization layers. The final frozen model with the weights is saved into a file that includes a .pb extension. To test the computational complexity of the proposed application, an iPhone 11 smartphone is considered. For these appliances, the audio latency for the iPhone 11 was 12-14ms. The processing time for the input frame of 32ms is 0.705ms. Since the processing time is lower than the length of the input frame, the proposed SE application works smoothly at low audio latency on the smartphone. Based on our measurements, the application runs on a fully charged iPhone 11 with a 3046 mAh battery for approximately 5 hours.

Figure 4 shows the CPU, memory, and battery usage of the proposed SE application running on the iOS smartphone used. The CPU usage of the app is 28% and the maximum memory consumption after the processing is turned on is 75.4 MB. The obtained frozen model with the trained weights is of size 11.5 MB, meaning the actual memory consumption of the SE application is around 65 MB. The smartphones present in the market usually have 12-16 GB memory; thus, the proposed application uses only 0.5 % of the entire smartphone memory.


InterspeechSE 〉 SSPRL's iPhone

InterspeechSE PID 2367

CPU 28%

Memory 75.4 MB

Energy Impact High

Figure 4: *CPU and memory consumption of the proposed CRNN-based SE application.*

5. Conclusions

A single channel CRNN-based SE application is proposed. The proposed application operates in real-time on an edge device. The developed algorithm is computationally efficient and implemented on an iPhone with minimal audio latency. The objective and subjective test results presented in the paper show that the proposed CRNN-based SE method outperforms conventional and neural network-based single-channel SE algorithms in terms of speech quality and intelligibility.

6. Acknowledgements

This work was supported by The National Institute of the Deafness and Other Communications Disorders (NIDCD) of the National Institutes of Health (NIH) under award number 1R01DC015430-04. The authors would like to thank the National Institute of the Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH) for their support. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

7. References

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Trans. Acoustic, Speech and Signal Process*, vol. 27, pp. 113-120, Apr 1979.

[2] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[4] P. J. Wolfe and S. J. Godsill, “Efficient alternatives to the Ephraim and Malah suppression rule for audio signal enhancement,” *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 10, pp. 1043–1051, 2003, special issue: Digital Audio for Multimedia Communications.

[5] Lotter, P. Vary, “Speech Enhancement by MAP Spectral Amplitude Estimation using a super-gaussian speech model,” *EURASIP Journal on Applied Sig. Process*, pp. 1110-1126, 2005.

[6] X. Zhang and D. Wang, "A Deep Ensemble Learning Method for Monaural Speech Separation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967-977, May 2016.

[7] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, “A regression approach to speech enhancement based on deep neural networks,” IEEE

Transactions on Acoustics, Speech and Signal Processing, vol. 23,no. 1, pp. 7–19, 2015

[8] Y. Wang and D. Wang, "Towards Scaling Up Classification-Based Speech Separation," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 7, pp. 1381-1390, July 2013.

[9] Y. Xu, J. Du, L. Dai and C. Lee, "An Experimental Study on Speech Enhancement Based on Deep Neural Networks," in IEEE Signal Processing Letters, vol. 21, no. 1, pp. 65-68, Jan. 2014.

[10] G. S. Bhat, N. Shankar, C. K. A. Reddy and I. M. S. Panahi, "A Real-Time Convolutional Neural Network Based Speech Enhancement for Hearing Impaired Listeners Using Smartphone," in IEEE Access, vol. 7, pp. 78421-78433, 2019.

[11] S. Fu, T. Wang, Y. Tsao, X. Lu and H. Kawai, "End-to-End Waveform Utterance Enhancement for Direct Evaluation Metrics Optimization by Fully Convolutional Neural Networks," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 9, pp. 1570-1584, Sept. 2018.

[12] Shankar, Nikhil, et al. "Efficient two-microphone speech enhancement using basic recurrent neural network cell for hearing and hearing aids." *The Journal of the Acoustical Society of America* 148.1 (2020): 389-400.

[13] Weninger, Felix, et al. "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR." *International Conference on Latent Variable Analysis and Signal Separation*. Springer, Cham, 2015.

[14] Goehring, Tobias, et al. "Using recurrent neural networks to improve the perception of speech in non-stationary noise by people with cochlear implants." *The Journal of the Acoustical Society of America* 146.1 (2019): 705-718.

[15] Healy, Eric W., et al. "A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation." *The Journal of the Acoustical Society of America* 145.3 (2019): 1378-1388.

[16] K. Tan, X. Zhang and D. Wang, "Real-time Speech Enhancement Using an Efficient Convolutional Recurrent Network for Dual-microphone Mobile Phones in Close-talk Scenarios," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 5751-5755.

[17] C. Karadagur Ananda Reddy, N. Shankar, G. Shreedhar Bhat, R. Charan and I. Panahi, "An Individualized Super-Gaussian Single Microphone Speech Enhancement for Hearing Aid Users With Smartphone as an Assistive Device," in IEEE Signal Processing Letters, vol. 24, no. 11, pp. 1601-1605, Nov. 2017.

[18] N. Shankar, A. Küçük, C. K. A. Reddy, G. S. Bhat and I. M. S. Panahi, "Influence of MVDR beamformer on a Speech Enhancement based Smartphone application for Hearing Aids," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Honolulu, HI, 2018, pp. 417-420.

[19] T. N. Sainath, O. Vinyals, A. Senior and H. Sak, "Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 4580-4584, doi: 10.1109/ICASSP.2015.7178838.

[20] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).

[21] Maas, Andrew L., Awni Y. Hannun, and Andrew Y. Ng. "Rectifier nonlinearities improve neural network acoustic models." *Proc. icml*. Vol. 30. No. 1. 2013.

[22] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, 2015, pp. 5206-5210.

[23] Microsoft DNS Challenge [Online] Available: https://github.com/microsoft/DNS-Challenge

[24] E. A. Lehmann, A. M. Johansson, and S. Nordholm, “Reverberation- time prediction method for room impulse responses simulated with the image-source model,” in *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct 2007, pp. 159–162.

[25] 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," 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221), Salt Lake City, UT, USA, 2001, pp. 749-752 vol.2.

[26] C. H. Taal, R. C. Hendriks, R. Heusdens and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, 2010, pp. 4214-4217.

[27] Reddy, Chandan KA, et al. "A scalable noisy speech dataset and online subjective test framework." *arXiv preprint arXiv:1909.08050* (2019).

[28] ITU-T Rec. P.830, “Subjective performance assessment of telephone- band and wideband digital codecs,” 1996.

[29] Google TensorFlow. [Online] Available: https://www.tensorflow.org/lite/