PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

De-reverberation using CNN for Non-Reference Reverberant Speech Intelligibility Estimation

Kazushi NAKAZAWA[1]; Kazuhiro KONDO[2]

[1] Graduate School of Science and Engineering, Yamagata University, Japan

[2] Graduate School of Science and Engineering, Yamagata University, Japan

ABSTRACT

Reverberation distorts speech, leading to degradation of speech intelligibility and speech quality. We have been investigating methods to predict the speech intelligibility of reverberant speech. This normally requires clean speech, which is often not available in real-world applications. In a past paper, we attempted to use Deep Neural Networks to reconstruct clean speech from distorted speech. We predicted a frame of clean spectrum from 11 frames of reverberant short-time Fourier transform (STFT) magnitude and evaluated the clean speech that is reconstructed from the predicted clean spectrum. The segmental SNR (SSNR) improved by 3.5 dB in reverberant environments, with a reverberation time from 0.2 to 1.0 second. However, with this method, the features were limited to 11 frames, shown to be inadequate to extract clean features from severe reverberant features. To improve the SSNR further, we need to extract an adequate number of frames of STFT magnitude, and further exploit the temporal characteristics. For this reason, we adapt Convolutional neural networks (CNNs), where we predict 257 frames of clean STFT magnitude from 257 frames of reverberant STFT. However, contrary to expectations, the SSNRs of predicted clean speech exceeds from DNN only in environments with short reverberations.

Keywords: Deep neural network (DNN), Convolutional neural network (CNN), speech intelligibility prediction

1. INTRODUCTION

With the advance of Wireless communication system, people often communicate in various noisy environments. One of the noise types in these environments, reverberation, is known to severely degrade speech intelligibility. Therefore, it is necessary to evaluate the speech intelligibility when constructing and managing a communication system with the necessary levels of intelligibility. Speech intelligibility is normally measured using objective evaluation using human listeners, which takes time and is costly. In order to reduce the cost, speech intelligibility should be predicted, eliminating the need for human listeners. Moreover, the prediction method in a non-reference manner, i.e., without the use of a reference signal, is more desirable than in a full-reference manner, which requires a reference signal. This is because clean (reference) speech is rarely available in practice, though full-reference methods generally perform better than non-reference methods. If we estimate the reference speech signal from reverberant speech, we can mimic full-reference speech intelligibility estimation by using the estimated reference speech signal. Fig. 1 illustrates this process. Thus, we need to estimate reference speech form reverberant speech. However, the reverberant characteristics of a reverberant signal generally depend on the position of the sound source and the receiver in a room because the reflection of the sound waves varies accordingly. In order to resolve this problem, we use Deep Learning for reference speech estimation.

In past research, we adopted fully connected Deep Neural Networks (DNN) which estimate the reference speech spectrograms from reverberant ones. We were able to reduce the reverberation in the tested signals, and segmental SNR (SSNR) improved by around 3 dB [1]. In these experiments, the DNN was able to estimate one temporal frame of the spectrum from 11 temporal frames in many of the experimental conditions. However, we also concluded that 11 frames are inadequate to extract features for some severely reverberant environments.

Accordingly, in this paper, we adopt the Convolutional Neural Networks (CNN), which can potentially process more time frames as a feature value, and may be able to estimate the reverberation more accuracy even with significant reverberation, leading to a more accurate clean speech estimation.

2. THE MODEL

2.1 Feature Values

We use the short-time Fourier transform (STFT) with a Hann window with a length of 512 samples, 50% overlap for feature extraction. Thus, the number of frequency bins is 257. We denote the squared log magnitude in the *k*-th frequency and the *m*-th frame as *X*(*m*,*k*). Therefore, each magnitude frame, denoted as vector $\mathbb{x}(m)$, is:

$$\mathbb{x}(m) = [\, X(m,1), X(m,2), \ldots, X(m,257)\,]^T \tag{1}$$

In order to take into account the transition between frames, we construct the input feature as $\tilde{\mathbb{x}}(m)$:

$$\tilde{\mathbb{x}} = [\mathbb{x}(m), \mathbb{x}(m+1), \ldots, \mathbb{x}(m+256)]^T \tag{2}$$

Therefore, the dimension of the input set is 257×257. The output feature set has the same structure. The feature values are normalized between -1 to 1. Additionally, we shift the *m*-th frame by 50 to make the features redundant.

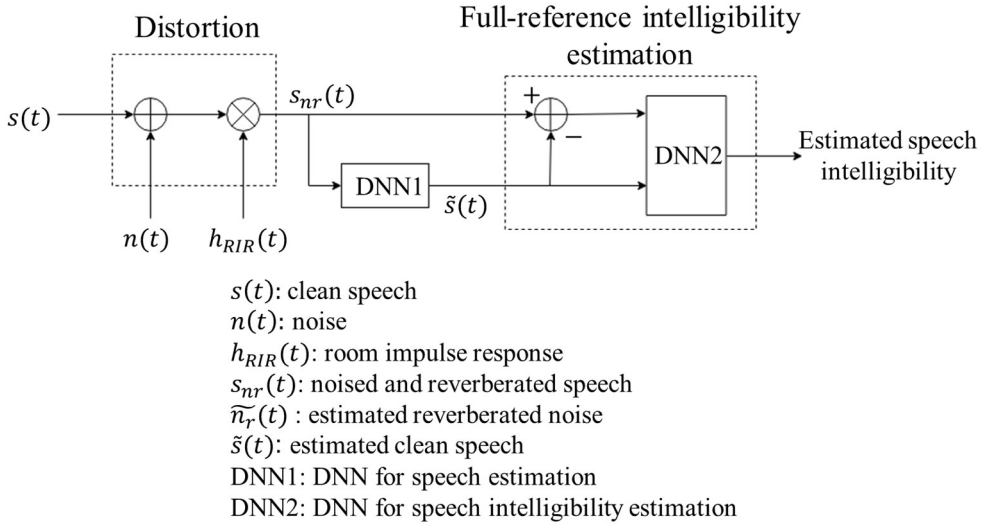

Figure 1. Reverberant speech intelligibility estimation

2.2 CNN Configuration

As described in the previous section, we will use CNN for the estimation of the clean speech signal. The configuration includes an encoder-decoder U-Net [2] that enables CNN to hold positional feature using skip connections. Fig. 2 shows the flow of the feature processing.

In the encoder part, an input feature matrix with a dimension of 257×257 is convoluted with a 2 ×2 dimension filter with a stride of 2 to generate a 128×128 feature matrix. Then, this 128×128 feature matrix is convoluted with a 4×4 size filter with a stride of 2 to generate a 64×64 feature matrix. The same process is repeated until the feature is encoded to a matrix of a single element.

In the decoder part, the encoded feature matrix of size 1 is transposed, and convoluted with a 4× 4 size filter with a stride of 2 to generate a 2×2 size feature matrix. The same process is repeated until the feature matrix is extended to size 128×128, after which this matrix is transposed and convoluted with a 3×3 size filter with a stride of 2 to generate a 257×257 size feature matrix.

The activate function of the CNN is leaky Relu, except for the output layer which uses the hyperbolic tangent.

3. GENERATING TRAINING DATA

We use reverberant sentence speech generated by convoluting dry speech utterances with an artificial Room Impulse Responses (RIR) for training and testing. We generate the RIR in a simulated

room with a size of 6m*6m*3m (length, width, height) with various reverberation times using the RIR generator [3] by changing the wall reflection coefficients. For training, we use 50 utterances read by seven male speakers and generate utterances with eight reverberant times, namely, 0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9, and 1.05 s, respectively. For each reverberation time, we generate two different RIRs with different microphones and speaker positions. For testing, we use 50 utterances read by another three male speakers and generated reverberant utterances with five reverberation times, namely 0.2, 0.4, 0.6, 0.8, and 1.0 s, respectively. All utterances in the training and test set are monaural with 16-bit quantization and 16 kHz sampling frequency.

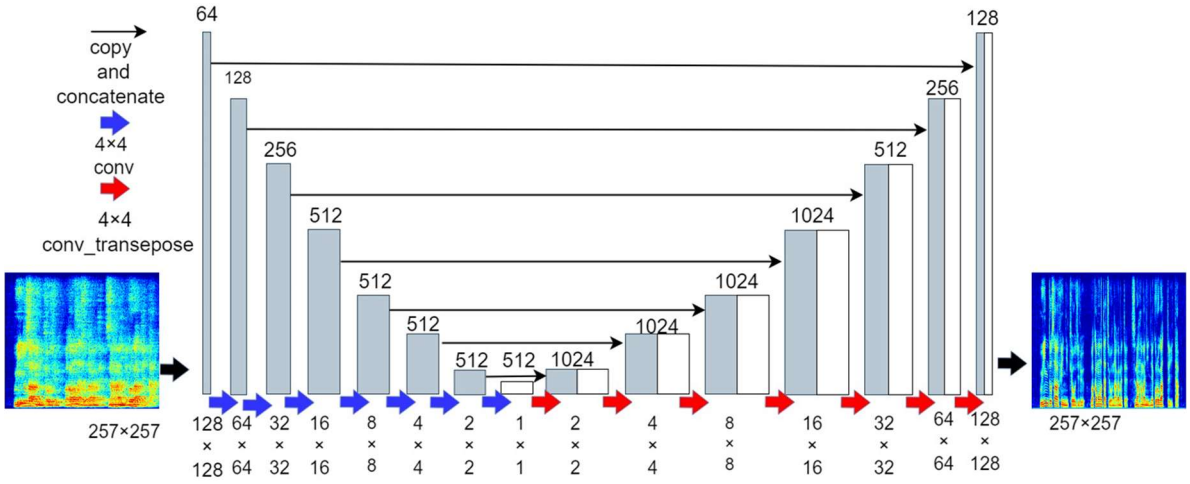

Figure. 2 U-Net configuration of the CNN

4. CLEAN SPEECH RECONSTRUCTION AND EVALUATION

With the trained CNN, we attempted to predict the clean squared log magnitude spectrograms from the reverberant ones using the test set, averaging the redundant parts. We can then reconstruct speech samples from predicted spectrograms using the Inverse STFT (ISTFT). Here we reconstruct speech from the predicted magnitude and the unprocessed (noisy) phase which is calculated from the reverberant speech. We calculated the Segmental SNR (SSNR) between the predicted and the corresponding original speech samples. Segment length for the SSNR calculation was set to 512 samples (32ms).

5. RESULTS AND DISCUSSIONS

Figs. 3 and 4 show spectrograms of clean, reverberated, DNN de-reverberated speech, and CNN de-reverberated speech with reverberation times of 0.2s and 1.0s, respectively. Fig. 5 shows the average SSNRs between the clean speech and the raw reverberant speech (circles), between the clean and the predicted clean signal using DNN (triangles), and between the clean and the predicted clean signal by using CNN (squares), respectively.

The figures show that de-reverberation using DNN and CNN can improve SSNR across all reverberation times tested. However, the estimation quality using CNN can only exceed that of DNN when the reverberation time is 0.2s. With longer reverberation times, the prediction performance was lower than the DNN, contrary to our expectations.

We speculate that this is due to the characteristic of U-Net, which we originally hoped will be able to hold the positional information of spectrogram, such as spectral peak positions. As shown in Fig. 3, estimated spectrogram using CNNs are more accurate, especially the harmonics mapping, than the DNN with a reverberation time of 0.2 s. As expected, the U-Net was able to hold the spectral peak positions with this reverberation time. However, with a longer reverberation time, e.g. 1 s, as shown in Fig. 4, estimated spectrogram using a CNN seems to be smeared compared to the DNN. This probably was caused by the U-Net excessively holding the smeared spectrogram information from the input to the output, resulting in the smeared peaks.

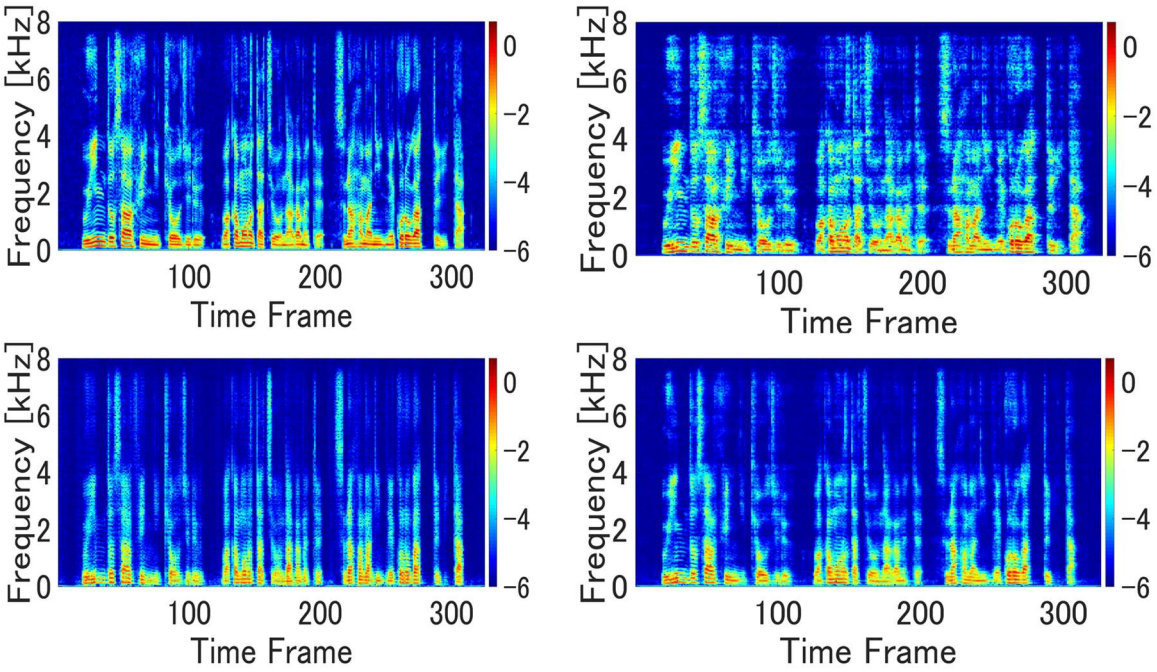

Figure 3. Processed speech spectrograms (reverberation time 0.2 s)

Spectrograms of clean (top left), reverberant (top right),

de-reverberated using DNN (bottom left), de-reverberated using CNN (bottom right) speech

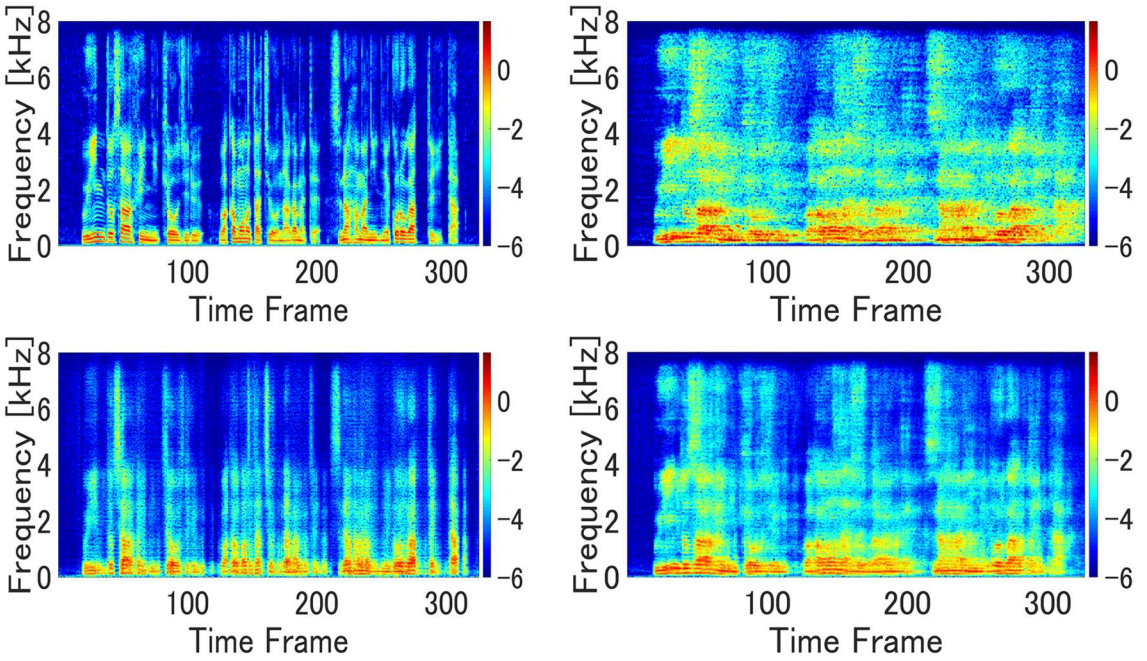

Figure 4. Processed speech spectrograms (reverberation time 1.0 s)

Spectrograms of clean (top left), reverberant (top right),

de-reverberated using DNN (bottom left), de-reverberated using CNN (bottom right) speech

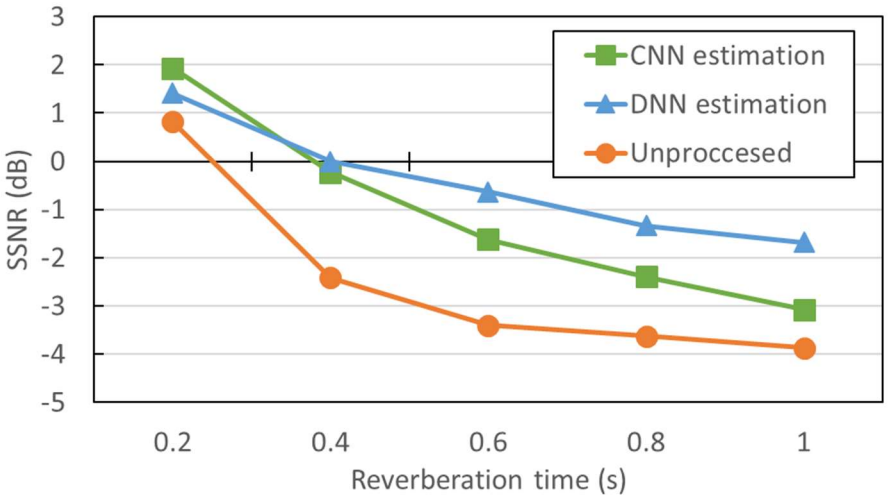

Figure 5. Average SSNR by the reverberation time

6. CONCLUSION

In this paper, we attempted to us a CNN for de-reverberation in order to improve the estimation performance compared to the past results using a densely connected DNN. Accurate estimation of de-reverberated speech is essential in order to estimate the speech intelligibility of reverberant speech accurately. We compared the de-reverberation performance of the DNN, used in previous research, and the newly introduced U-Net using CNN, in terms of SSNR. As a result, we were able to improve the SSNR by up to 2.5 dB using CNN from raw reverberant speech. However, the CNN estimation performance exceeded the DNN performance only with the shortest reverberant environment. The performance was lower with the CNN for all longer reverberation times. We speculate that the result was caused by the characteristics of the U-Net, which was able to hold the positional feature. We found that the U-Net was able to reconstruct harmonics construction accurately only with the short reverberation time. This characteristic of the U-Net seems to be disadvantageous to all longer reverberations, resulting in smearing the estimated spectrograms.

We plan to estimate phase component not only magnitude component in phase-aware processing and use the estimated clean speech to predict the reverberant speech intelligibility without requiring the clean reference signal.

ACKNOWLEDGEMENTS

This work was supported in part by the JSPS KAKENHI 17K00223.

REFERENCES

1. Kazushi N, Kazuhiro K. De-reverberation using DNN for Non-Reference Reverberant Speech Intelligibility Estimation. Proc. GCCE 2018; 9-12 October 2018; Nara, Japan 2018. pp.316-317.
2. Olaf R, Philipp F, Thomas B. U-Net: Convolutional Networks for Biomedical Image Segmentation. Proc. MICCAI 2015; 5-9 October 2015; Munich, Germany 2015. pp. 234-241.
3. AudioLabs. RIRGenerator. https://www.audiolabs-erlan-gen.de/fau/professor/habets/software/rir-generator/