SPEECH DEREVERBERATION WITH U-NET ARCHITECTURES

Zachary Neveu

Technical University of Denmark

ABSTRACT

This work compares 2 deep neural architectures for the speech dereverberation.

***Index Terms*—** One, two, three, four, five

1. INTRODUCTION

In real world environments, speech is heard as a mix of direct signal with reflections caused by the surrounding environment. Late reflections decrease speech intelligibility, causing problems for listeners, as well as tasks such as automatic speech recognition (ASR). This work explores the potential of convolutional networks to remove problematic reverberation in speech signals.

2. MODEL ARCHITECTURE

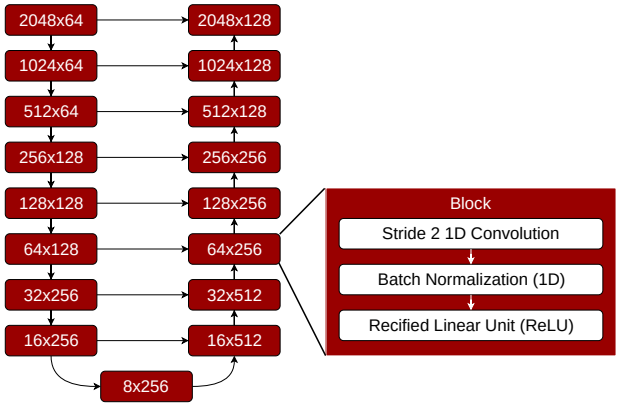

Fig. 1: Model Architecture

The two networks used are both based on the U-Net architecture originally proposed for speech enhancement/denoising in [1]. The specific layer sizes and the architecture of each block is detailed in figure 1. The time-domain network replaces the final rectified linear unit activation with a hyperbolic tangent activation to produce values between -1 and 1. The frequency domain network is one block in a larger traditional signal processing system shown in figure 2.

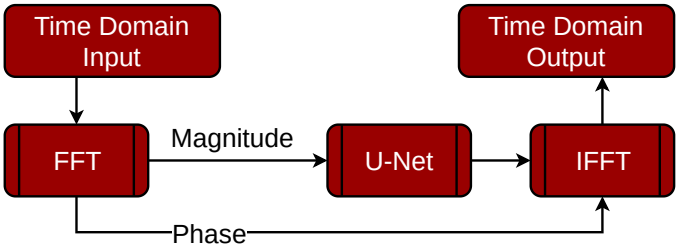

Fig. 2: System Architecture of Frequency domain Network

3. DATA AND TRAINING

The training data consists of a subset of the LibriSpeech dataset [2] which is processed according to the pipeline shown in figure 3. Both models are trained on this data for 15 epochs using the Adam optimizer. Training is stopped at this point because losses have largely stopped decreasing and training is slow because of large file sizes of inputs.

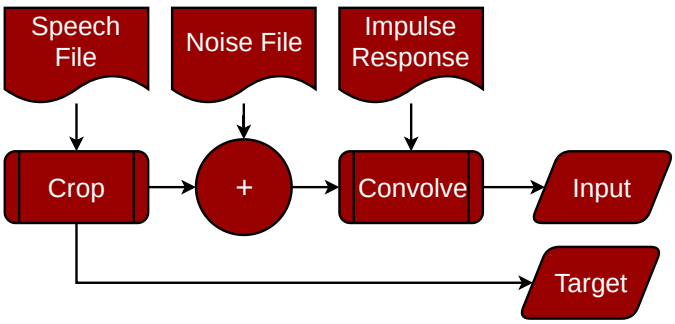

Fig. 3: Data pipeline

4. RESULTS

To evaluate the results, two common speech metrics are used: STOI and PESQ. STOI measures speech intelligibility, whereas PESQ measures speech quality. Figure 4 shows the scaling for each score.Table 1 shows the mean change in the speech metric scores for 10 random reconstructed examples from the validation set. Both the time-domain and frequency-domain networks improve the STOI while making the PESQ

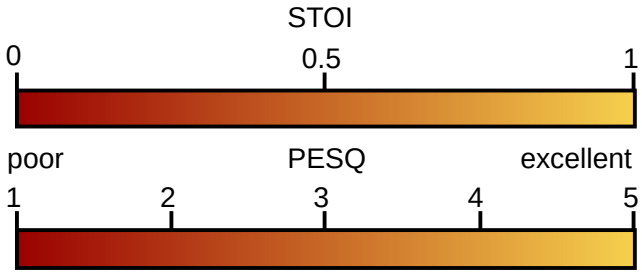

Fig. 4: PESQ and STOI metric scales

Table 1: mean change STOI and PESQ scores for a subset of the validation set

Method	avg. ΔSTOI	avg. ΔPESQ
Frequency-domain	0.070	-0.383
Time-domain	0.023	-0.159

worse. These networks may be useful for tasks such as ASR where intelligibility is valued over quality. Notably, the frequency-domain network achieved both a larger boost in STOI and a larger detriment to the PESQ than the time-domain network. Figure 5 shows spectrograms for the input, target, and both models for an example speech sequence.

5. REFERENCES

[1] Ashutosh Pandey and DeLiang Wang, "A New Framework for CNN-Based Speech Enhancement in the Time Domain," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1179–1188, July 2019.

[2] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210, IEEE.

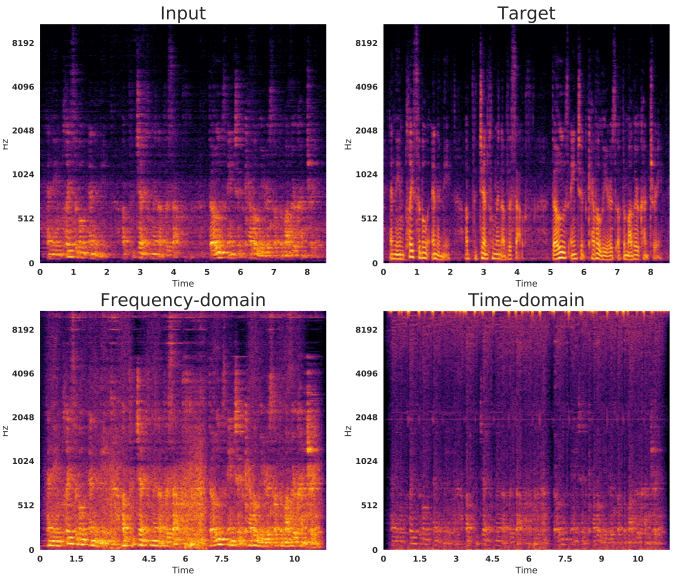

Fig. 5: Spectrograms of Reconstructed Example