SPEECH DEREVERBERATION WITH U-NET ARCHITECTURES

Zachary Neveu

Technical University of Denmark

github.com/zacharyneveu/DeReverb

ABSTRACT

This work compares the performance of time and frequency domain approaches to speech dereverberation using u-net architectures. Towards this end, a neural architecture is constructed consisting of a u-net with a recurrent latent layer. Two models of this architecture are trained, one using time-domain inputs, one using frequency domain magnitudes. Results are compared using the Short Time Objective Intelligibility Measure (STOI) and Perceptual Evaluation of Speech Quality (PESQ) metrics. Additionally, a simple frequency-domain perceptually weighted loss function is presented.

***Index Terms*—** Speech, Enhancement, Reverb, Noise

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). Traditionally, the best approach for speech dereverberation is to use a multi-channel approach. The basic principle of these approaches is that direct path sound will have high coherence between multiple microphones, while reverberant sound will have low coherence

add citation here

. Based on coherence, a gain function can be applied in the time-frequency domain to remove reverberant sound. These approaches have the downside that they require extra hardware. Additionally, traditional coherence based approaches do not take advantage of the structure of speech, only exploiting spatial differences between sounds. A method is then desired which can remove reverberation from a recording using only a single microphone which can exploit the structure of speech. The first of these problems has been approached using traditional signal processing methods to estimate the impulse response of the reverberation and perform de-convolution to obtain the clean signal. The problem with these approaches is that noise is often mixed with the signal, and this is often amplified greatly by small errors in the impulse response estimation. Based on their success when used in problems such as ASR, Neural networks are able to learn the characteristics of speech. This should allow a neural network to remove reverberation that is mixed with non-stationary noise.

2. MODEL ARCHITECTURE

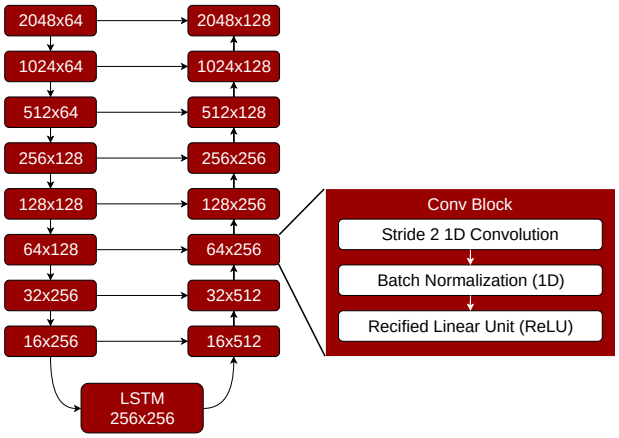

Fig. 1: Model Architecture

Two networks are trained which are both based on the U-Net architecture originally proposed for speech enhancement/denoising in [1]. The reasoning behind using a U-Net architecture is that speech is known to be sparse, and can be represented well with a small amount of data. The information bottleneck of a U-Net architecture should allow the network to distill salient information about the speech, while the pass-through connections allow for the recreation of utterance-specific details. Additionally, the 1-dimensional convolution layers used are known to be useful for frequency analysis, enabling the time-domain network to learn frequency-domain features. The specific layer sizes and the architecture of each block is detailed in figure 1. An long short term memory (LSTM) is used to filter the latent representation of the speech stored in the bottleneck of the u-net. This architecture modification is used in

Cite this from demucs

for source separation and is based on the intuition that each frame of sound processed is highly dependent on past frames. 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.

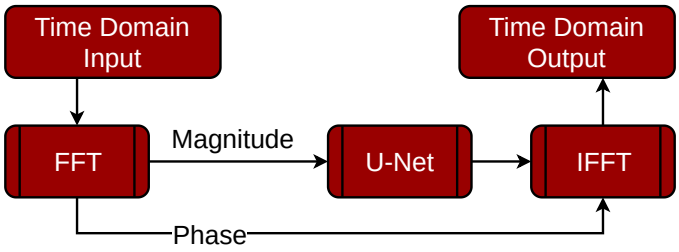

Fig. 2: System Architecture of Frequency domain Network

3. DATA

The training data consists of a subset of the LibriSpeech dataset [2] which is processed according to the pipeline shown in figure 3. In total, the dataset consists of 2255 speech utterances, 36 impulse responses, and 20 noise recordings. In order to avoid overfitting, each time a speech file is loaded, it is randomly cropped, then paired with a segment of a randomly chosen noise file and a randomly chosen impulse response. In the data pipeline presented, noise is added before the convolution takes place. This is done in an effort to model real rooms where ambient noises are also convolved with the reverberation of the room. This ordering is important to note, as most approaches use the reverse ordering. It is also important to note that the result of the convolution operation is truncated to retain the same length as the target signal. This is necessary so that the loss function can be easily be computed between the network output and the target.

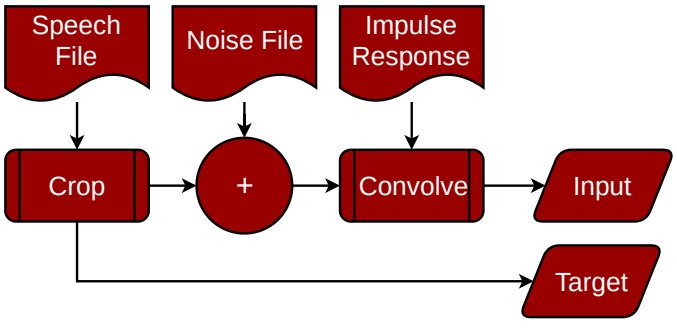

Fig. 3: Data pipeline

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

4. TRAINING

The time and frequency domain models are each trained for 30 epochs using the Adam optimizer

cite this

. Training is stopped at this point because losses have largely stopped decreasing. Both models show converging trends, however it is important to note that because of the random transforms used in the data pipeline, large changes in loss can occur between epochs, and thus losses are not strictly decreasing for the training or validation set. For the time-domain model, a simple mean squared error (MSE) loss is used. It should be noted that while this loss is frequently used, it has several theoretical shortcomings when used in this context. The first of these is that a white signal (equal power at each frequency) will have larger amplitudes at lower frequencies. Errors at low frequencies will then take higher priority than errors at high frequencies. The second shortcoming of MSE loss is that it is not scale invariant. Relative losses in quiet frames will be optimized far less than relative losses in loud frames. One final shortcoming is that MSE in the time domain is highly sensitive to phase, unlike the human ear. For the frequency-domain model, a naive MSE loss is also tried, in addition to a custom loss function. In the frequency domain when working with only amplitude, MSE makes more sense than in the time domain, because it is phase-invariant. The other two shortcomings still apply, however. In order to counteract the over-weighting of low-frequency content, a simple custom technique is developed.

The idea behind this custom loss function is to multiply each frequency bin of the model output and target by the perceptual relevance of that frequency before computing the loss function. This technique ensures that the largest penalties are applied to differences at the most perceptually relevant frequencies. To compute the perceptual weights, an equal loudness contour is used as detailed in ISO 226

cite iso 226

. In order to convert this curve to perceptual weights, the curve is inverted, then normalized to have a maximum of 1.

5. RESULTS

To evaluate the results, two common speech metrics are used: STOI and PESQ. STOI measures speech intelligi-

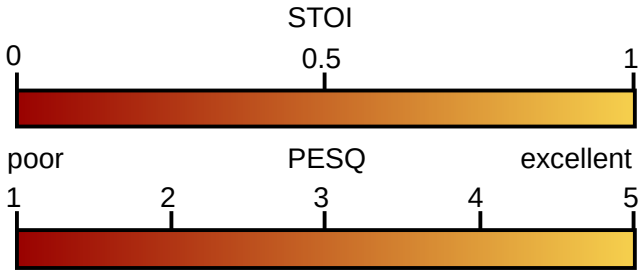

Fig. 4: PESQ and STOI metric scales

bility, whereas PESQ measures speech quality. Figure 4 shows the scaling for each score. Both PESQ and STOI are intended to be used on segments of speech that are many frames long. In order to measure results in this manner, a long segment of speech is reconstructed using the windowed overlap-add (WOLA) method

cite if i can find this...

. This method was chosen over concatenating adjacent frames because it lessens the severity of discontinuities at frame boundaries. The WOLA method consists of reconstructing frames that overlap by a constant number of samples. Reconstructed frames are multiplied by a Hann window. All frames are then added together with respective offsets to create the final reconstructed signal. Table 1 shows the mean change in the speech metric scores for 10 reconstructed random examples from the validation set. Both the time-domain and frequency-domain networks improve the STOI while making the PESQ 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.

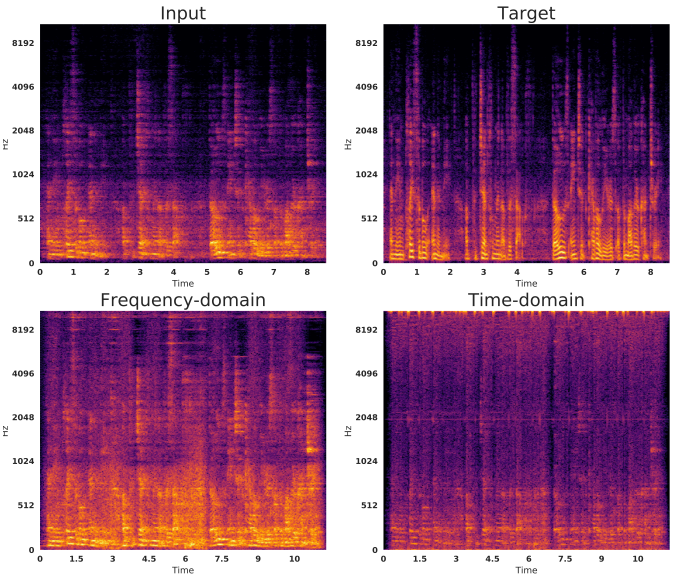

Fig. 5: Spectrograms of Reconstructed Example

6. 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.