An Autoencoder Baseline for Channel Normalization System

Ge Zhu

Dept. of Electrical and Computer Engineering, University of Rochester

Abstract

Channel normalization system aims to remove the channel effects in speech signals, e.g. quantization noise in coder or speech distortion, such system is able to increase the robustness of automatic speech verification (ASV) systems. However, to the best of our knowledge, previous works were usually dealing with additive noise or convolutional noise. In this project:

- We create a channel corrupted speech dataset
- We applies an autoencoder architecture to serve as a baseline for deep learning based channel normalization.

Introduction

In commercial ASV systems, speech signals are usually recorded through communication devices of various qualities like telephones, cell phones, laptops etc., and different types of devices have different bandwidths and codec standards. To cover a wider users, a lower bitrate codec are offen applied, but at the same time, it will introduce distortions and packet loss. The above mentioned conditions may cause various channel distortions and further affect the performance of ASV.

Neural network based ASV system have been used to minimize channel effects for years, this method is able to achieve great performance as long as training data and test data have similar properties.

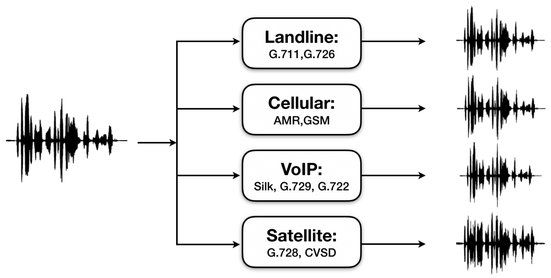

Figure 1: Acoustic Codec Simulator

Dataset Generalization and Dataset for training

We apply acoustic simulator to VCTK corpus to generate different codec corruptions for each clean speech. This simulator comprises mainly four types of codec standards with various parameters. Original clean dataset contains 6626 pieces of speech segments sampling at 48kHz from 8 different people.

Training corruption dataset comprises of cvsd modulation with 128kb/s. Of all the speech segments, 2995 pieces are for training, 1225 for validation and the rest for testing.

Methods: Autoencoder

The aim of autoencoder is to learn robust representations from the data. Autoencoder neural networks first learn to compress input data into code and then uncompress back, reconstruct the original data and simutaneously remove corruption. We first applied the simplest form of an autoencoder: feed-forward, non-recurrent neural network which is basically many single layer perceptrons (MLP).

To store sequential information, we adapt the primary autoencoder into RNN fashion by adding a LSTM layer.

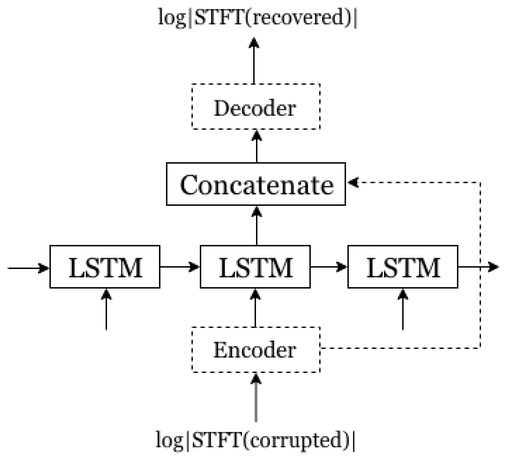

Figure 2: Recurrent Autoencoder

Evaluation

Our results show that this baseline is able to recover clean speech and improve speech quality. From the (perceptual evaluation of speech quality) PESQ curve, we can see the improvement between corrputed speech and recovered speech. The average PESQ improvement in simple autoencoder is 0.46 and 0.49 for autoencoder with LSTM layer.

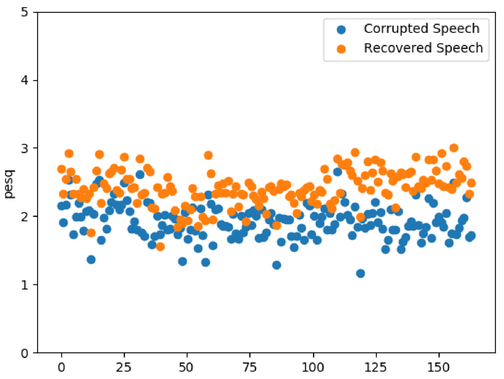

Figure 5: PESQ of Corrupted Speech and Recovered Speech

Results

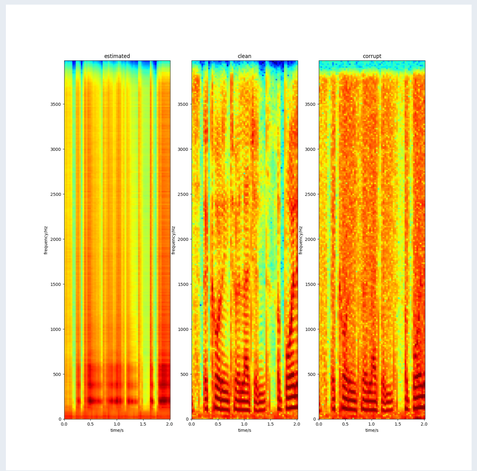

Figure 3: Recovered Spectrogram after 1 Epoch

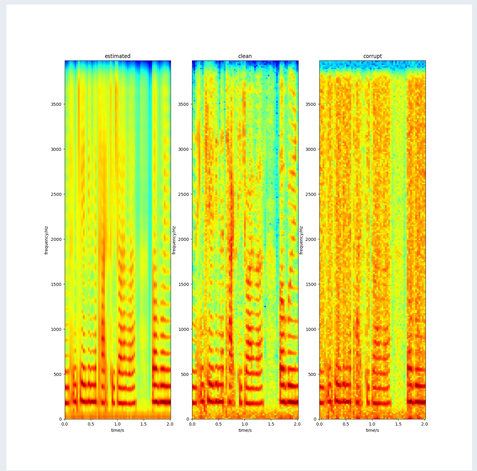

Figure 4: Recovered Spectrogram after 100 Epoch

Conclusions

In this project, we develop an autoencoder baseline to compensate for a certain channel corruption. We trained and evaluated this system on our created dataset. Results show that this model needs further training and parameters tuning. Since this project only develops a baseline for channel compensation, much more work should be done for future work, such as building up a better neural network architecture for a more generalized dataset.

References

[1] A. W. Rix et.al. Perceptual Evaluation of speech quality (PESQ): A new Method for Speech Quality Assessment of Telephone Networks and Codecs. *ITU-T Recommendation.862*, ITU, 2001.

[2] M. Ferras et.al. A large-scale open-source acoustic simulator for speaker recognition. *IEEE Siganl Processing Letters*, **23(4)**,527–531, 2016.

[3] Y. Yan and Z. Duan. HW5: Singing voice separation with neural networks. *U of R Press*, Wilmot, 2018.

Audio Information Research Laboratory