Assessing VOIP Quality – Deep Learning Approach

Allen Ansari

Author Note

I would like to express my very great appreciation to Blake Arensdorf for his valuable and constructive suggestions during the planning and development of this research work.

Abstract

Mean Opinion Score (MOS) is one of the common metrics to measure audio quality. In other words, MOS is the average of call quality ratings given by different users when a VOIP call ended. Most companies like Skype ask users to rate the call but audio signals may be non-stationary in time. In other words, audio quality might be high during calls and suddenly there is a decrease in intelligibility, which made the user rate the final rate lower than expected. In this capstone I am trying to extract voice features using deep learning (COVNET) to predict MOC and compare it with the rates users are given. The data set I am using is the TCD-VoIP dataset which is designed to aid the development and testing of speech quality VoIP systems. It contains a set of five types of VoIP degradations along with their corresponding subjective opinion scores (MOS). To extract voice features I will use the Short-Time Fourier Transform (STFT) to encode the audio signal into four different feature representations. This capstone can be used by VOIP providers to compare their voice quality with the rate received from users.

Keywords: VOIP, Mean Opinion Score, Short-Time Fourier Transform, deep learning

Assessing VOIP Quality – Deep Learning Approach

## Introduction

In voice over IP the voice quality is one of the most important parameters to determine the service quality. There are some technical parameters to measure voice quality like packet loss, jitter and codec protocol but most of VOIP provider like Skype asks user to rate the call based on their experience which is called Mean Opinion Score (MOS). This metric is average of all rating given by statistically significant number of users. Since audio call is not stationary therefore one MOS value may represent different situations. Meaning that the quality of voice may not be the same during the call and it effects user experience based on that. To solve this problem, I am trying to predict MOS based on audio signal features that will be extracted using Short-Time Fourier Transform (STFT) technique to encode the audio signal into four different feature representations:

* The spectral magnitude of the STFT
* The Spectrogram of the STFT
* The Mel-Spectrogram
* The Constant-Q transforms
* The Mel Frequency Cepstral Coefficients (MFCCs)

At the end the constant-Q transforms show the lows validation mean absolute error as 0.39 which is less than 10% of the rating range (1 to 5) to predict MOS.

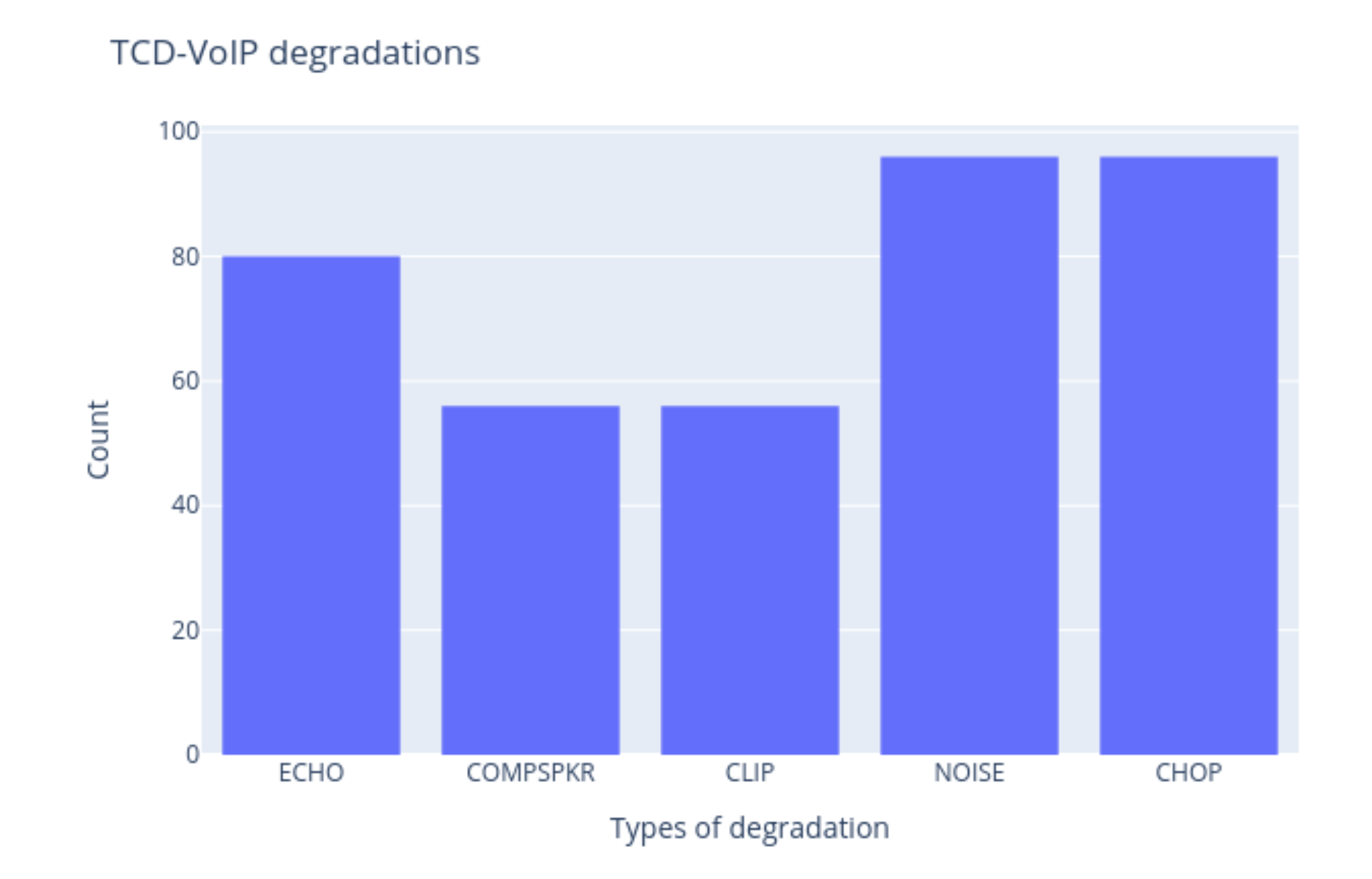
## The Dataset

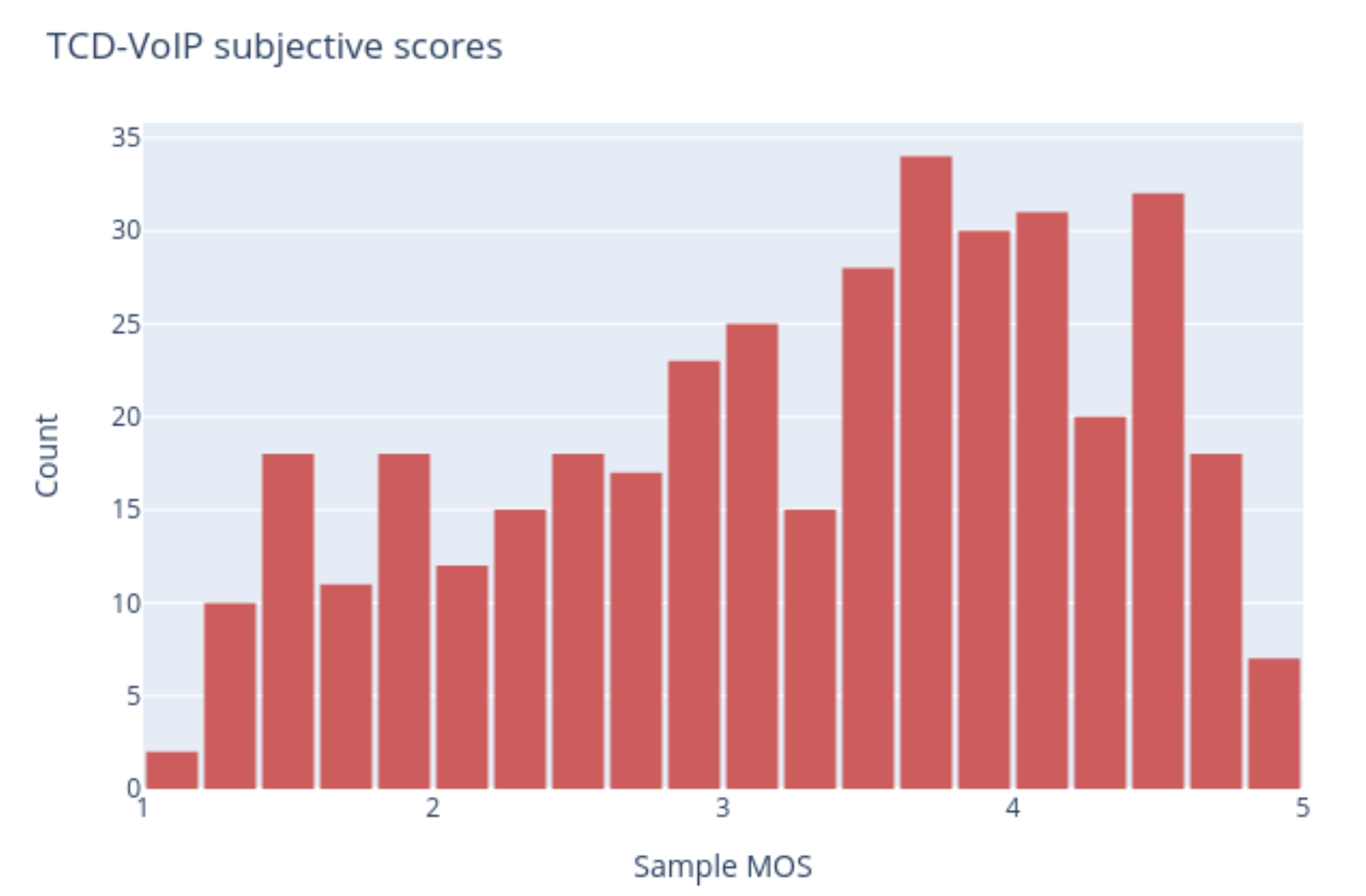
For this capstone I used TCD-VoIP dataset[[1]](#footnote-1) which designed to aid the development and testing of speech quality VoIP systems. It contains a set of five types of VoIP degradations along with their corresponding subjective opinion scores (MOS). The dataset focus on degradation that occurs independently of hardware or network and is freely available.

The TCD-VoIP covers five types of commonly seen degradations in VoIP applications. These are:

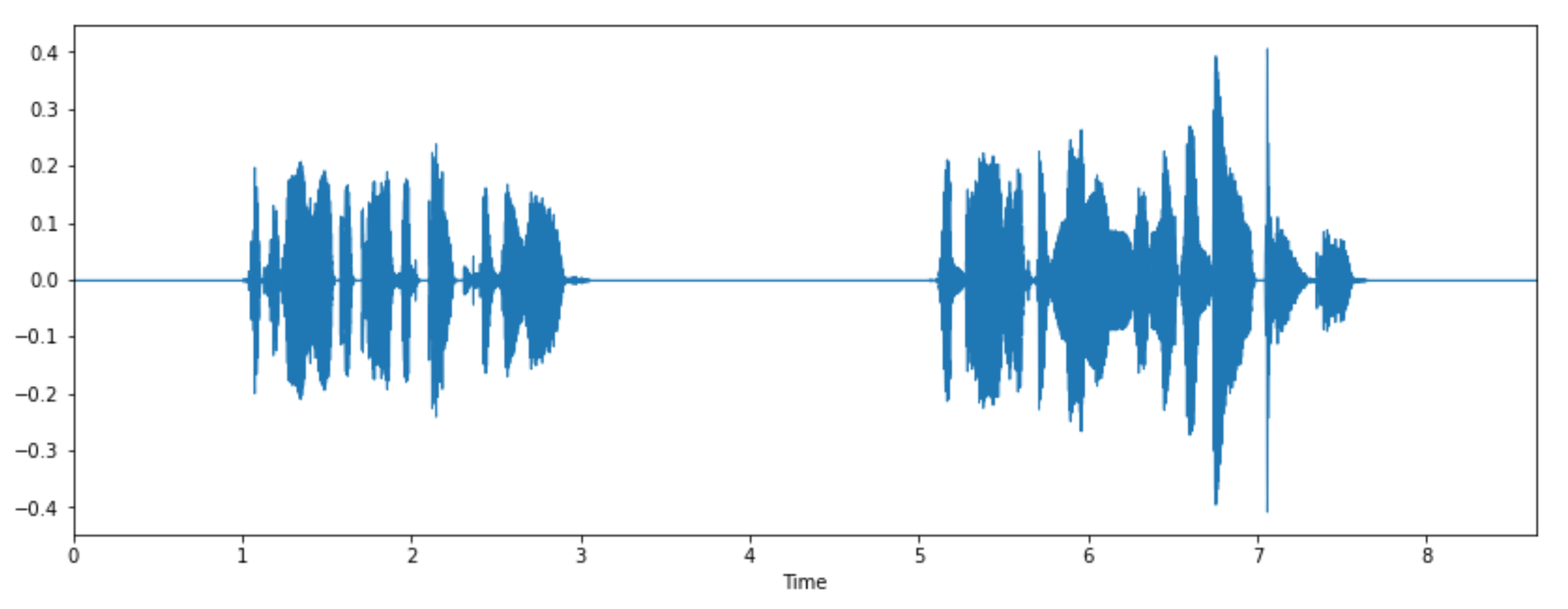
* Background noise
* Intelligible competing speakers
* Echo effects
* Amplitude clipping
* Choppy speech

For each audio sample, there are individual subjective opinion scores from 24 different listeners. Likewise, the final subjective score (MOS) is given as the arithmetic mean across the 24 scores. In total, there are 384 audio files with two male and two female speakers. You can see the distribution of speech-degradations and MOS in the images below.





As an example, the wave form for clip ones would be as below:



## Feature Extraction

The STFT is the most common representation of time-frequency for audio signals. The idea is to compute the Fourier transforms over small portions of the input signal. Since audio samples are highly non-stationary (mainly music signals), the STFT breaks the signal into smaller portions as a way of providing a more robust final representation. For STFT transform, I used a Hamming window that covers 512 sample points of an input audio signal. The window moves with a stride (hop-size) of 64 points which guarantees a 75% overlap. Finally, we take the magnitude of the STFT and use it as the final feature vector. As a side note, to compute the Spectrogram of the STFT, we would just square the magnitude of the STFT. Since the audio samples have different lengths, we pad the STFT using the “wrap” mode so that the feature vectors have the same shape. This way, the STFT has 259 frequency bins and 1241 frames in time.

## The Solution

The final solution consists of a multi-output ConvNet with approximately 58K trainable parameters. The architecture consists of repeated blocks of:

Convolution → Batch Normalization → ReLU → Max Pooling → Spatial Dropout

To balance the contribution of each task to the final loss, we weight each loss inversely proportional to their strengths. Basically, we run a short experiment of 100 epochs and store the individual raw losses per epoch. We proceed by computing the average degree of magnitude of one loss over the other. By doing so, we can measure how larger (on average) one of the losses is compared to the other.

## Results

Despite the small size of the TCD-VoIP dataset, results are reasonably good, especially for MOS estimation, which achieved a mean absolute error of 0.39.

References

Last Name, F. M. (Year). Article Title. *Journal Title*, Pages From - To.

Last Name, F. M. (Year). *Book Title.* City Name: Publisher Name.

1. http://www.mee.tcd.ie/~sigmedia/Resources/TCD-VoIP [↑](#footnote-ref-1)