Deep learning-based detection of Dysarthric speech disability

Siddharta Pukash
Department of Computer Science Stanford University, sidp@Stanford.edu

Abstract

Dysarthria is a form of speech disability affecting 170 per 100 K persons and one-third of persons with traumatic brain injury. If we convert audio signal into a two-dimensional representation, then detecting dysarthria using deep neural network is a problem effectively doing image classification. This project explores building a deep learning model to detect Dysarthria using various 2-D representations of audio signal namely STFT (Short Time Fourier Transform), Mel-Filterbank, Spectrogram, Mel-Spectrogram.

Audio Signal Processing

STFT (Short Time Fourier Transform)
Divide long term signal into short segments and compute Fourier transform for each segment.

Mel-Filterbank
Decompose input signal STFT/FFT into components, each carrying a single frequency sub-band of original signal. And Mel (similar to log) over it. With 2-D representation.

Spectrogram
2-D visual representation of audio signal with time and frequency dimensions.

Mel-Spectrogram
Log of Spectrogram values, represented still in 2-D space. Adjoining image show linear, log and Mel representation of audio signal.

Model Design

Model design is inspired from [1] and training was done on all four audio processing as input. The dataset from TÖROG [2] database was divided into single word, multi word and mixed groups. Label was detected from folder nomenclature. E.g. “K” folder contains control group of female non-dysarthric female audio. Input was zero-center normalized and padded.

Model Training

Model training is divided into three phases:
1. Data preparation.
2. Model building.
3. Model evaluation.

Experiment Results

<table>
<thead>
<tr>
<th></th>
<th>Words only</th>
<th>Sentence only</th>
<th>Word and sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>STFT</td>
<td>69%</td>
<td>62%</td>
<td>66%</td>
</tr>
<tr>
<td>Mel-Spectrogram</td>
<td>71%</td>
<td>54%</td>
<td>56%</td>
</tr>
<tr>
<td>Mel-Filterbank</td>
<td>71%</td>
<td>48%</td>
<td>52%</td>
</tr>
</tbody>
</table>

Table 1: Result of training the model with different audio processing.

Analysis

We achieved 68% accuracy while baseline target was 72% with 3% variability [3]. Direct acoustic time-frequency signal needs very deep (up to 24 layers) neural network to give same performance of feature identification based on another study. Mel-Spectrogram representation of input audio signal gave best results, though this could vary depending on hyper-parameter tuning or design of deep learning model. Representation method of acoustic audio affects the result of dysarthria detection using deep-learning model.

Detection of dysarthria is clearer using only single word audio input.

Future work

Explore multi-dimensional representation of acoustic signal before passing it into deep learning network.
Feed all the four-audio signal input to the input layer and let the model learn which representation to use during its training process.
Train the model along with reconstruction of the normal audio from dysarthric audio. If we do that, then we can compare the reconstructed audio with the original control/non-dysarthric audio and try to reduce the distance between the two during learning process.

References

Video Link

http://urlshortener.at/kotA1