Deep learning-based detection of Dysarthric speech disability

Siddhartha Prakash

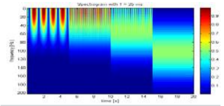
Department of Computer Science Stanford University, sidp@Stanford.edu

Abstract

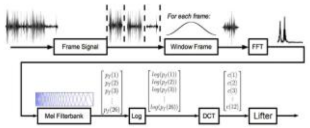
Dysarthria is a form of speech disability affecting 170 per 100 K persons and one third of persons with traumatic brain injury. If we convert audio signal into a two-dimensional representation, then detecting dysarthria using deep neural network is a problem effectively doing image classification. This 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

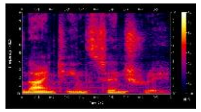
STFT (Short Time Fourier Transform)
Divide long term signal into short segments and compute Fourier transform for each segment.



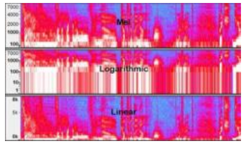
Mel- Filter bank
Decompose input signal STFT/FFT into components, each carrying a single frequency sub-band of original signal. And Mel (similar to log) over it. With 2-D representation.



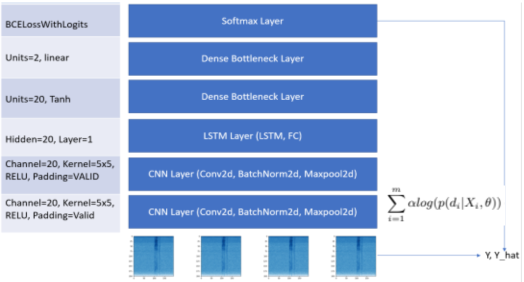
Spectrogram
2-D visual representation of audio signal with time and frequency dimensions



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 Training

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

Test Set	95:2.5:2.5 fold Cross validation
Init	Xavier Initialization
Drop Off	0.3
Training	SDG (lr=0.001, momentum=0.009)
Batch size	4
Data grouping	Single word, multi-word, mixed
Training time	Approx 2 hours
Loss	Binary Cross Entropy + Logits
Training Epoch	41
Audio library	Librosa, TorchAudio
Framework	Azure VM, Python 3.6
Data Source	TORGO Database

Experiment Results

	Words only	Sentence only	Word and sentence Mixed	Reference
Mel-Spectrogram	68%	62%	66%	72 +- 3% [4]
STFT	58%	54%	56%	NA
Spectrogram	63%	58%	61%	NA
Mel-Filterbank	58%	48%	52%	NA

Table1: Result of running 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 cleaner 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

[1] Daniel Korzekwa, Roberto Barra-Chicote, Bozena Kostek, Thomas Drugman, Mateusz Lajszczak, "Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech" in 20th Annual Conference of the International Speech Communication Association, 2019, arXiv:1907.04743v1 (2019)

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097–1105.

[3] Millet, Juliette; Zeghidour, Neil, "Learning to Detect Dysarthria from Raw Speech", in International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019

Video Link

http://urlshortener.at/kotA1