Gender Classification Through Speech Research Proposal

Authors: Hieu Mai, Tai Nguyen, Harpreet Cheema - ECEC 487 - Advisor: Dr. Anup K. Das

I. Introduction

Speech recognition is a complex problem with real-life applications. Hence, this research focuses on exploring different learning techniques: Convolutional Neural Network (CNN - Hieu Mai), Recurrent Neural Network (RNN - Harpreet Cheema), and Spiking Neural Network (SNN - Tai Nguyen), in gender classification and speech data processing. The details of these methods will be shown in separate research paper and compared in a group presentation.

II. Preliminary Data Processing

The dataset consists of 6300 sentences spoken by 630 speakers with genders specified. Below is the initial data processing, in which the frequency distribution shows that male voices have lower pitch while female voices have higher pitch.

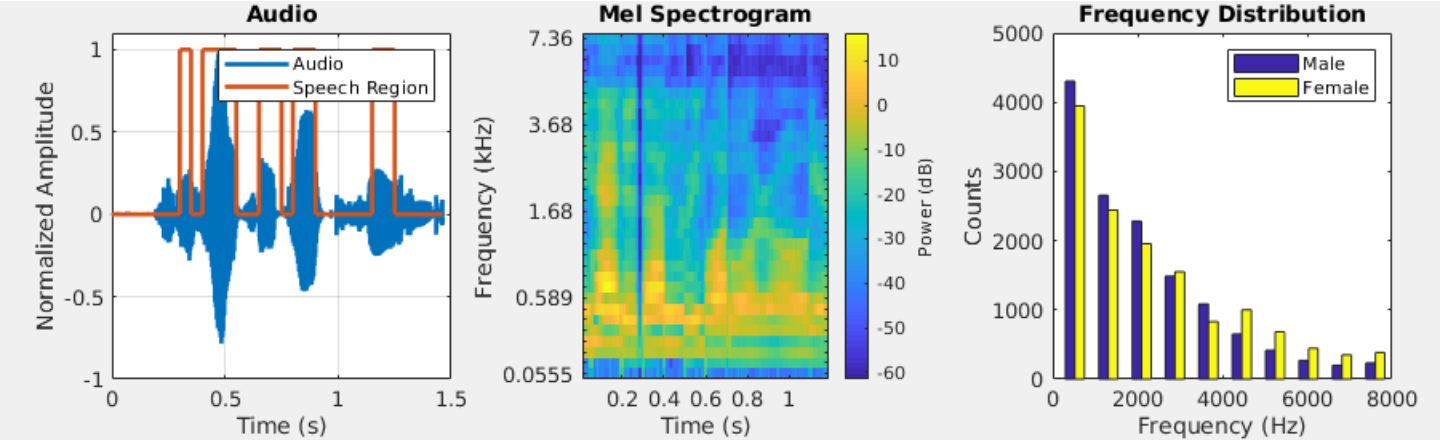

Figure 1 (left): Example of the silence removal algorithm. Figure 2 (center): Mel Spectrogram of a sample audio. Figure 3 (right): Dominant audio frequency distribution after silence removal of the male and female population.

III. Methodologies

The SNN algorithm in this research is a feedforward network using STDP learning rule. The network consists of an input layer, a convolutional layer, and a max-pooling layer, whose output determines the class of the input [1]. To obtain the input signal for the SNN, the Mel-Frequency Cepstral Coefficients (MFCCs) are obtained from the raw data [1], which is encoded with the time-to-first-spike method. This method will be tested against a new input method, in which the input signal is the 10 most dominant frequencies for each recording. The CNN algorithm takes segments of raw signal input with removed silent frames and fed into single or multi convolutional layers of different window sizes [2]. A different implementation takes three MFCC features as inputs, which are analogous to images, then processes them using conventional image kernels. The RNN LSTM algorithm uses input, output, and forget gates in a structure in order to control the cell state and determine the output [4]. These gates are controlled by weights which decide the amount of information fed forward into the cell state. The performance of this algorithm will be evaluated with respect to a simple K-nearest neighbor classifier.

IV. References

[1] Dong M, Huang X, Xu B (2018) Unsupervised speech recognition through spike-timing-dependent plasticity in a convolutional spiking neural network. PLoS ONE 13(11): e0204596.

[2] Kabil, Selen Hande, et al. “On Learning to Identify Genders from Raw Speech Signal Using CNNs.” *Interspeech 2018*, 2018, doi:10.21437/interspeech.2018-1240.

[3] Graves, Alex, et al. “Speech Recognition with Deep Recurrent Neural Networks.” *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, doi:10.1109/icassp.2013.6638947.

[4] Sak, H. & Senior, Andrew & Beaufays, F.. (2014). Long short-term memory recurrent neural network architectures for large scale acoustic modeling. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 338-34