Discrete Speech Limited Vocab Recognition with Application in Verbal Equation Writing

Researcher: Yulin Zhou Advisor: Prof. Ismail Jouny
Lafayette College

Abstract

In this modern age, speech recognition plays a critical role in hands-free driving, security, home integration systems, and disability assistance. While current speech recognition software has very holistic solutions and is very well-developed, many real usages of speech recognition technology are more situation-specific. This research aims to apply a limited library of vocabulary, derived from mathematic equations, and test the accuracy of different machine learning algorithms such that various performances can be compared. This experiment reveals that limited vocab speech recognition can achieve its best performance using the Convolutional Neural Network(CNN). The CNN takes in the data processed by dynamic time warping, combined with an enlarged dataset by adding white noise to the original audio inputs to obtain the Mel-frequency Cepstral coefficients for further analysis.

Background

- **Speech recognition**
  - Speech in → text out
- **Convolutional Neural Network --- CNN**
  - A class of Artificial Neural Network (ANN)
  - Most applied to analyze image inputs
- **Long Short Tern Memory --- LSTM**
  - An artificial recurrent neural network (RNN) architecture
  - Most applied to analyze sequential inputs
- **Support Vector Machine --- SVM**
  - a non-probabilistic binary linear classifier
  - SVM maps training examples to points in space so as to maximize the width of the gap between the two categories.

Methods

In this research, the below flow chart(Figure. 3) was applied to conduct the experiment.

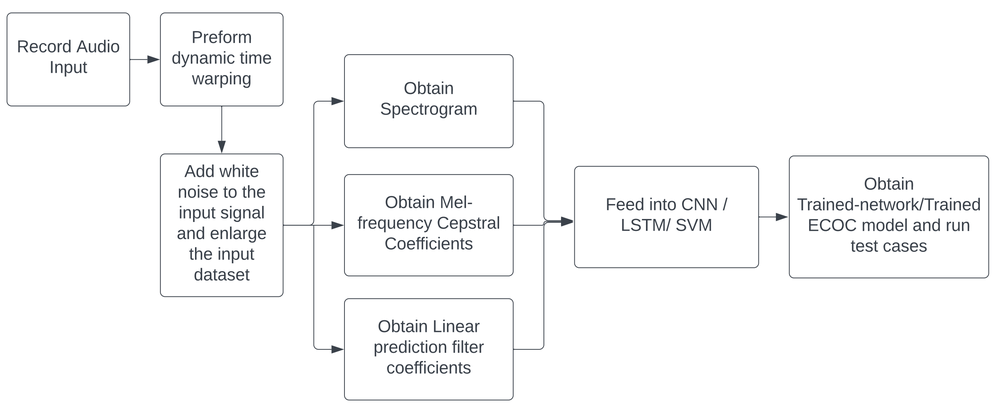

Figure. 3 Flow Chart for Experiment

Dataset ---

The input data is recorded though the record function provided by MATLAB.

72 words were taken, including 28 numbers, 20 signs and operators, 7 trigonometric function, 17 letters and Greek letters. Each words were recorded 10 times for training and 3 times for testing.

The Audio Toolbox in MATLAB was used during signal processing.

Result and Conclusion

The experiment was divided into 3 parts, each use one of the mentioned machine/deep learning algorithm.

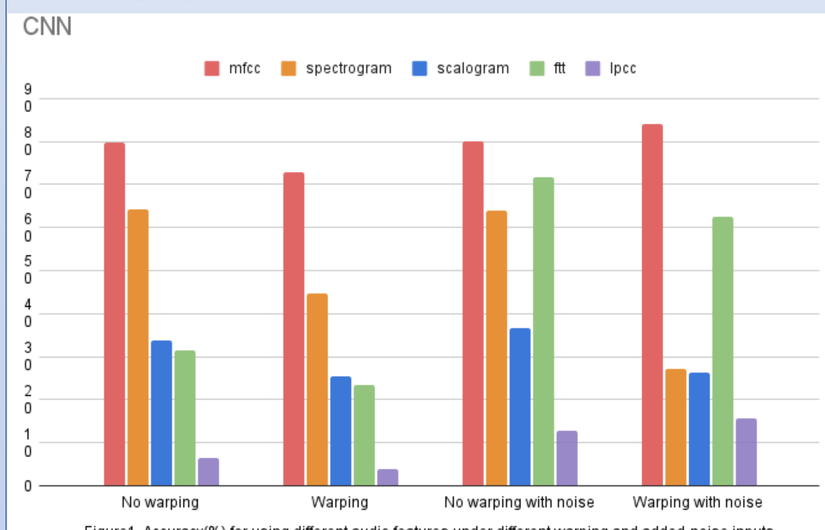

Figure1. Accuracy(%) for using different audio features under different warping and added-noise inputs

Figure 1 shows the results for the major part of the research, CNN. Among the four input scenarios, the models using Mel-frequency Cepstral coefficients(MFCC) performs the best. The models use Linear prediction filter coefficients (LPCC) which results in the worst performance. One reason for that may be that LPCC contains insufficient data points for CNN to train.

Figure 2 shows the comparison of the results from the 3 parts of the research. CNN and LSTM are both have greater performance than SVM, which shows that SVM, a classifier originally designed for binary classification, is limited when attempting to meet multi-class tasks.

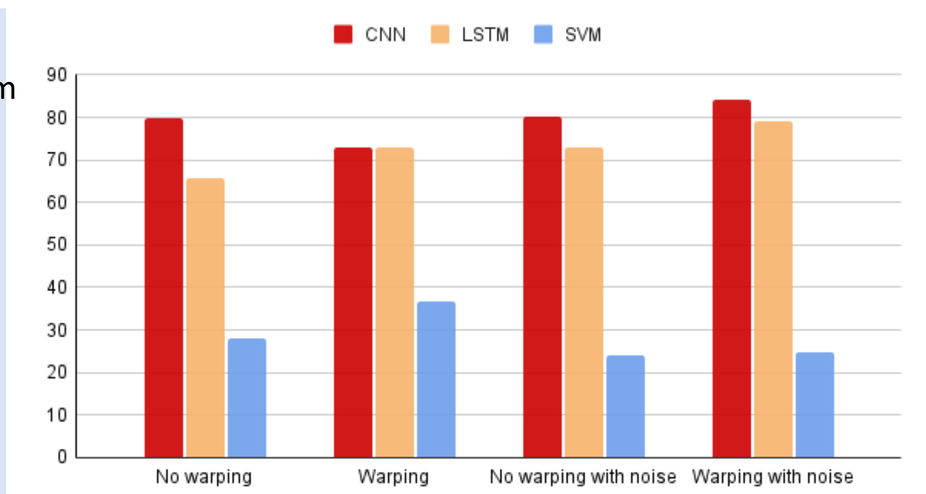

Figure 2. Accuracy(%) under different warping and noise input

Figure 3. shows the detailed results for the model using CNN with MFCC.