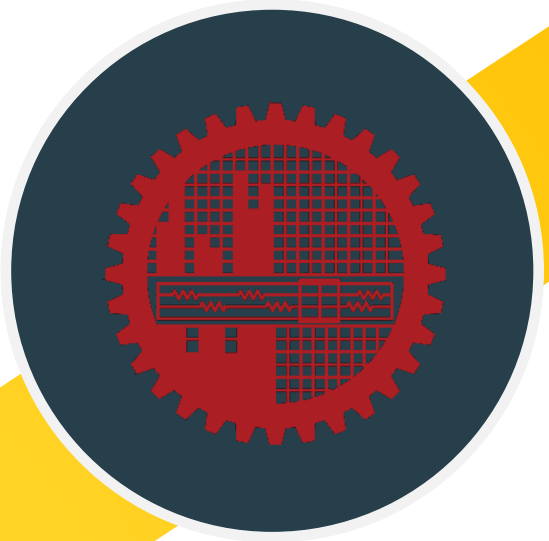
Isolated Word Recognition

PROJECT REPORT

Group No. 5

ID: 1906082 1906083
1906084 1906085



EEE 312

Digital Signal Processing I Laboratory

Submitted To,

Shahed Ahmed
Lecturer
Department of EEE, BUET

Barproda Halder
Lecturer (PT)
Department of EEE, BUET

Submitted By,

Ramim Hassan Shawn
ID: 1906082

Mohammad Toki Bin Alam
ID: 1906083

Mushfiquzzaman Abid
ID: 1906084

Eftekhar Sadik
ID: 1906085

Table of Contents

Isolated Word Recognition

Introduction

Speech recognition technology has gained significant popularity in recent years due to its ability to precisely detect and convert human speech into a machine-readable format, enabling its use in various applications. This technology has revolutionized the way people interact with electronic devices, allowing hands-free control of various devices, such as smartphones, smart home systems, and virtual assistants. It is a valuable tool in many applications, enabling seamless human-machine interaction, increasing efficiency, and providing an accessible means of communication for individuals with disabilities. With continued advancements in speech recognition technology, we can expect its usage to expand further, offering even more benefits in the future. In this project, we tried to build a small model of user independent isolated keyword speech recognition system.

Problem Statement

Isolated keyword detection is the task of identifying specific keywords from an audio input. Six words were selected as keywords for this purpose. Our selected words are given below:

Apple	Banana	Coconut	Jackfruit	Mango	Oranges

The total workflow of problem is subdivided into three parts:

a) Collecting voice data of keywords for different users and storing them in a database
b) Taking voice data of unknown keyword from new user as input and performing feature extraction techniques
c) Comparing the features of previously stored voice data and new data to find which keyword was spoken.

Audio processing techniques can be used to extract features from the audio signal and identify the presence of the keyword.

There are many feature extraction techniques like Mel-Frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), Perceptual Linear Prediction (PLP), Mel-Spectrogram, Discrete Wavelet Transform (DWT), Wavelet Packet Transform (WPT), Probabilistic Linear Discriminate Analysis (PLDA) etc. In this project, Mel-Spectrogram and Mel-Frequency Cepstral Coefficient (MFCC) are used.

After extracting the features, a classifier is used to measure the similarity and dissimilarity between the coefficients and therefore the speech signals. For classification, various techniques are available like Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Network (ANN), Dynamic Time Warping (DTW), K-Nearest Neighbor (KNN) etc. We used Dynamic Time Warping (DTW) method in this project.

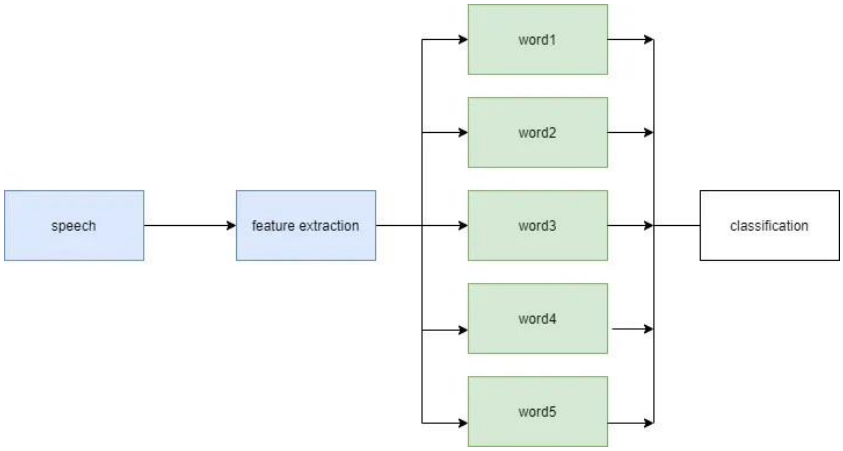


Figure 1 Project Workflow

Proposed Methodology

One common approach to isolated keyword detection is to use a technique called "keyword spotting". Keyword spotting involves training a model to recognize the specific keyword or phrase by analyzing its acoustic characteristics.

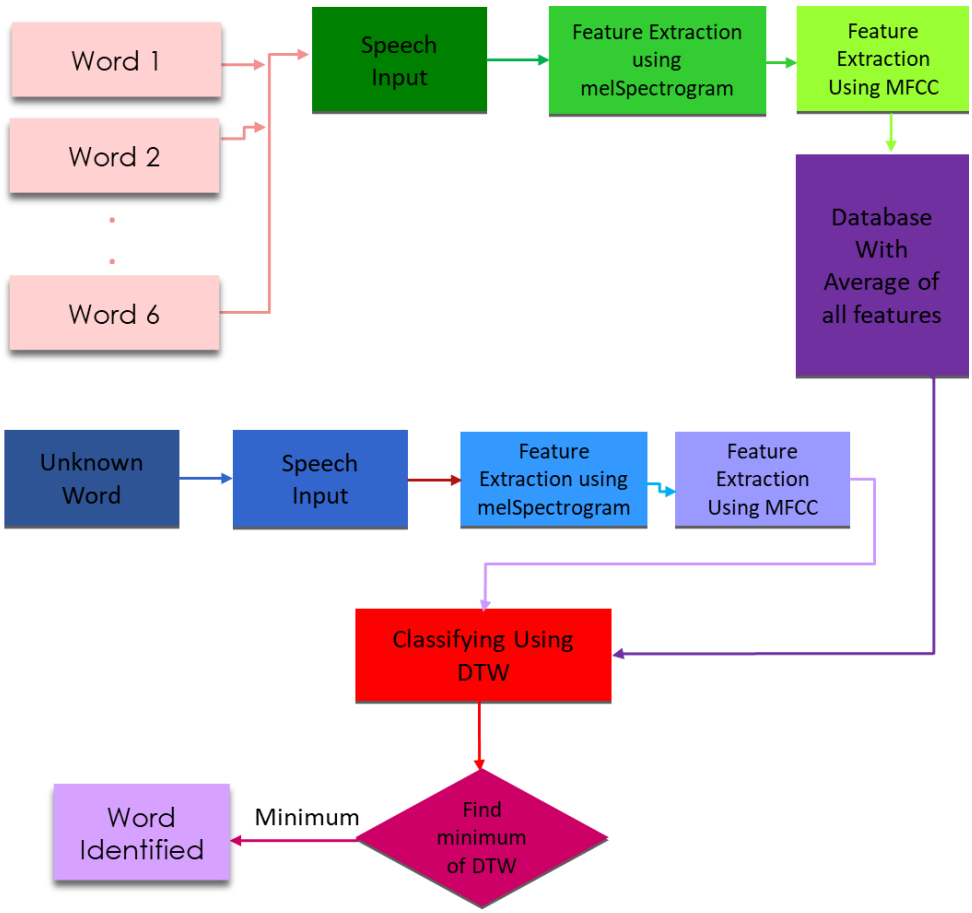


Figure 2 Algorithm Flowchart

We collected 80 sets of training data for each of the keywords. Roughly 3 samples for each word per person was collected. Unfortunately, many sample data was noisy. But we could not discard them, as our data set was not large enough.

We used the following key algorithms in our project. All of them are MATLAB built in function. These algorithms are shortly discussed below:

Mel-Spectrogram

– Used this algorithm to extract features from audio

MelSpectrograms are a type of spectrogram commonly used in audio signal processing that utilizes a filterbank to separate audio signals into different mel-frequency bands. They provide a more compact and informative representation of an audio signal's frequency content, with improved perceptual relevance.

Mel-frequency cepstral coefficients (MFCCs)

– Used this algorithm to extract features from audio

MFCCs are a commonly used feature in speech recognition. They are derived by taking the logarithm of the power spectrum of the audio signal, then applying a filterbank to extract frequency bands that are more closely related to human perception.

The MFCC uses the MEL scale to divide the frequency band to sub-bands and then extracts the Cepstral Coefficents using Discrete Cosine Transform (DCT). MEL scale is based on the way humans distinguish between frequencies which makes it very convenient to process sounds.

Dynamic Time Warping (DTW)

- Feature Matching is done by DTW

The principle of DTW is to compare two dynamic patterns or time sequences and measure its similarity by calculating a minimum distance between them. DTW matches one sequences with the others and obtain the optimum matching.

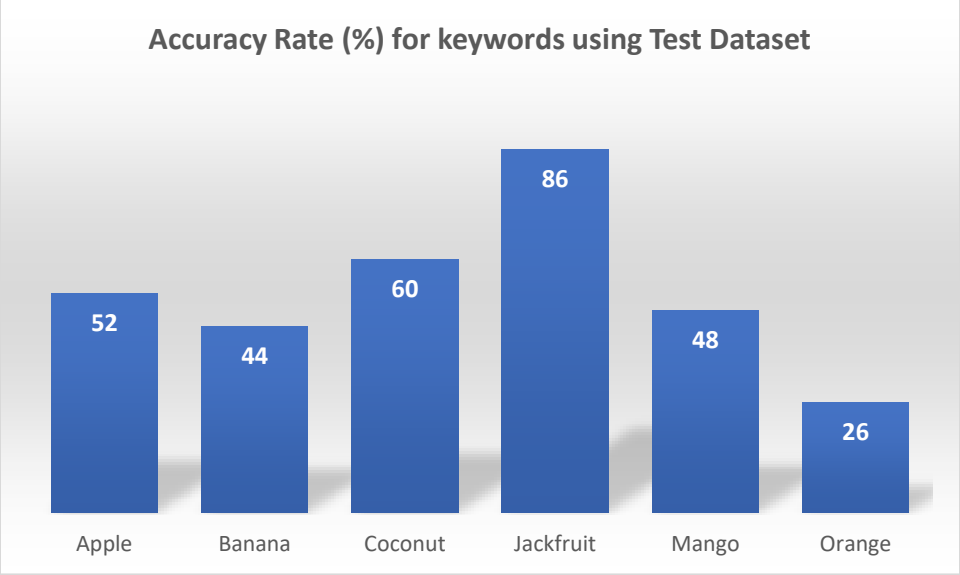
DTW finds the best alignment between two sequences, even if they have different lengths or vary in time. The alignment is found by warping the time axis of one sequence so that it matches the time axis of the other sequence. DTW has many applications in speech recognition, where it can be used to match unknown speech patterns to a known set of speech patterns. If similarity between two sequences increases, DTW returns smaller value.

During the training phase, we extracted features from each train data. We used both Mel-Spectrogram and MFCCs to extract features. Then an average is taken of the feature vectors found from each training sample for each word. The averaged feature vectors are then stored in a data file. There are total 12 sets of feature vectors for 6 words. For each word, there are 2 vectors, one for MelSpectrogram and other for MFCC. These are later used for detecting any unknown word spoken.

For any unknown word spoken, we similarly calculated both Mel-Spectrogram and MFCCs. The feature vectors for unknown word are then compared using DTW algorithm with each of the average feature vectors previously stored. The DTW result found from MelSpectrogram and MFCC are multiplied for each word. The results of multiplication are stored in an array. For 6 words there are 6 results of multiplication. We selected the lowest value of the array as detected word. The word for which DTW of both features returns minimum value is the detected word.

Results

We used a set of 50 data for each word to test the program. The test dataset was collected from the train dataset. We got an accuracy rate of 63.2% overall. The accuracy for different words are given below:



In our live testing, the program performed slightly worse. We got an accuracy rate of about 50%. The accuracy rate for each word also varied. Some words were easily detected, and some were hard to detect. The accuracy ranking during live testing is given below:

Jackfruit > Banana > Mango > Apple > Orange > Coconut

There can be several reasons behind these inadequate results. We could have used a better algorithm or further optimizations could have been done. We tried to use Hidden Markov Model (HMM) at first but was not successful. Deep learning models were not feasible because of small training data.

The main reason for the moderate result is the poor quality of train data. As stated before, the train dataset had many noisy data. Noise in training data makes it harder to detect. Moreover, pronunciation of different people is different. It also contributed to incorrect detection. The live environment was speaker independent detection, so the accuracy rate decreased than that we got with the test dataset.

Future Works

Further improvement on the project can be done to enhance accuracy and robustness of the system, by taking the following steps:

- Increasing the number of training datasets with quality data
- Obtaining large number of datasets (more than 1000) with diversified utterances from the users.
- Using deep learning-based models
- Further optimizing the existing model
- Adding other feature extraction models. With augmentation of more speech features like voice pitch, intensity, dynamics, speech recognition reliability will increase hugely.

Discussion

Isolated keyword detection has many practical applications, such as in voice assistants, where the system needs to recognize specific wake words or commands to activate. It can also be used in security systems or call center applications to monitor for specific keywords that may indicate a problem or concern. The idea of this project was to build a program which would be able to detect words from voice signals for some selected words. The accuracy level achieved using the current algorithm was moderate. Accuracy during live testing becomes slightly worse because of the minimal modification of the user's voice from the stored sample speeches and due to surrounding noise. Using more advanced classifier algorithms, such as deep neural networks, may improve accuracy further. Unfortunately, due to limited datasets, these algorithms were not utilized, and we had to accept the results obtained. Although results are still modest, it can be a foundation for future developments if necessary.

References

[1] Bhadragiri Jagan Mohan, Ramesh Babu. N, "Speech Recognition Using MFCC and DTW"

[2] H.Mansour, Abdelmajid & Zen Alabdeen Salh, Gafar & Mohammed, Khalid. (2015). Voice Recognition using Dynamic Time Warping and Mel-Frequency Cepstral Coefficients Algorithms. International Journal of Computer Applications. 116. 34-41. 10.5120/20312-2362.

[3] L. Muda, M. Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, No. 3, March 2010, pp. 138-143

[4] Linlin Pan," Research and simulation on speech recognition by Matlab",Dec 2013

[5] Risto Hinno, "Single word speech recognition"