BANGLADESH UNIVERSITY OF ENGINEERING AND TECHNOLOGY

Department of Electrical and Electronic Engineering

Course No.: EEE 312

Course Title: Digital Signal Processing Laboratory

Project Tittle : Keyword Spotting

Date of Submission: 07.03.2023

<u>Submitted To:</u>

Shahed Ahmed	**Barproda Halder**
Lecturer, EEE, BUET	**Part-Time Lecturer, EEE, BUET**

<u>Submitted By:</u>

1906074- Shakir Ahmed
1906075- Mehedi Hasan
1906076- Md. Abu Sayed Chowdhury
1906077- Md. Sharif Uddin

Problem Statement:

In this project, we were focused on detecting specific keywords. The project's necessity is becoming ubiquitous with the increase of technologies like voice-based vending machines, home automation systems, and voice assistance.

Previously machine learning and deep learning techniques were performed for this purpose. But it has a crucial drawback. It requires a vast data set for training. But, for a few keywords, relying on such a method is time- and resource-costly. Simplely, feature extraction and statistical processing gives a satisfactory result.

MFCC features are extracted, and the comparison is conducted using DTW to detect the uttered keyword. Detection of keywords is done in three modes: isolated keywords, foreign words, and keywords in a sentence.

Methodology:

Part 1: Data collection

For each word, total 83 voice samples was collected from 29 individuals, among whom 27 individuals gave 3 voice samples and 2 individuals gave a single voice sample. The voice clips (.wav) were 2 seconds long and was collected using 8000Hz sampling frequency. So, total 498 voice sample was present in our data set.

Part 2: Parameter adjust for highest accuracy

2.1 MFCC extraction

Mel frequency cepstral coefficients for 498 voice samples were extracted using built-in matlab function mfcc: coeffs = mfcc(AudioIn,fs,Name,Value). For name-value pair, the considered names were" LogEnergy" "NumCoeffs", "OverlapLength" and "Window". Periodic hamming window was constantly used. Number of coefficients, overlap length and window length were varied for different accuracy. Overlap length and window length combinedly varies the number of frames.

2.2 DTW and comparison

Dynamic time warping was done between MFCC of each sample against all 498 MFCCs. This resulted in 498 distances from which 6 final distances for each key word was found by averaging 83 distances of each word. The minimum of the 6 distances is spotted as the input key word. So, for 498 samples, 498 output is found which is compared with the inputs to find %accuracy.

$$Total\ Accuracy = \frac{Matched\ keywords\ \times\ 100\%}{498}$$

$$Individual\ Accuracy = \frac{Matched\ keywords\ \times\ 100\%}{83}$$

2.3 Varying parameters and repeating part 2 & 3

In part 2.1, different "name-value" s are used based on the given approaches and in part 2.2, accuracy for each approach each found.

For the best outcome of each approach, the best parameters are determined; and is passed and kept constant in the next approach.

Approach 1: From default 14 output coefficients (log energy + 13) of MFCC function, 4 combinations of log energy and 1st coefficient were tested :

Kept coefficients.
Y = yes
N = No
	Log energy	First

	Y	Y
	Y	N
	N	Y
	N	N

Approach 2: Varying the number of coefficients (default 13) through varying the value of "NumCoeffs" from **6 to 26**.

Approach 3: Varying overlap length and window length.

Overlap length (ms)	Window Length (ms)
0.015	0.030
0.020	
0.025	
0.020	0.025
0.010	0.020
0.015	
0.010	0.015

Part 3: Handling 3 problems

3.1 Spotting the correct keyword

I. Input voice from user is taken through microphone in a 2 second and fs = 8000 Hz voice clip
II. MFCC with the best parameters is extracted.
III. 6 dtw distances are found as in part 2.2.
IV. Minimum distance is the keyword.

3.2 bonus 1: spotting not-found-in-the-given keywords

I. Same steps as 3.1 (I -III).
II. The difference between maximum and minimum distance calculated.
III. Threshold is set upon this difference

3.3 bonus 2: spotting keyword in a sentence

I. Input voice from user is taken through microphone in a 4 second and fs = 8000 Hz voice clip.
II. The total clip was framed with using window of length WL and overlap of length OL.
III. Kaiser window β was applied in each frame.

IV. The frames where 'mean of absolute amplitude > k*maximum of absolute amplitude of the frame' were considered for the next steps.
V. Each frame was assigned to a 2 second audio clip.
VI. For each clip, step 3.2 (I - II) were applied.
VII. Maximum difference indicated the frame with the keyword.
VIII. WL, OL, β, K were varied for the best outcome.

Results:

Best found parameters in part 2:

Approaches	Parameters	Values
1	Log energy – 1st coefficient	N -N
2	Number of coefficients	10
3	Window length	0.020ms
3	Overlap length	0.015

Accuracy:

	%
Forest	90.3614
Jungle	100
Lake	85.5422
Moutain	81.9277
Ocean	84.3373
River	81.9277
Overall	**87.3794**

Bonus 1: Unreliable

Bonus 2:

Parameters	Value
WL	0.75s
OL	0.5s
β	5
K	0.07

Results in the following outcomes:

I. Highly accurate detection (85%)
II. Run time error in code

Discussion:

1.Main problem:

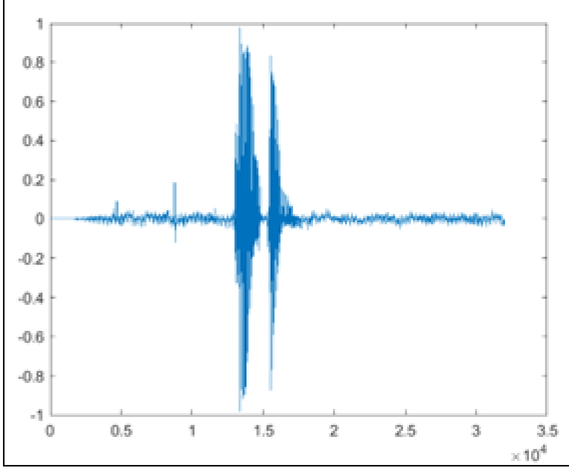
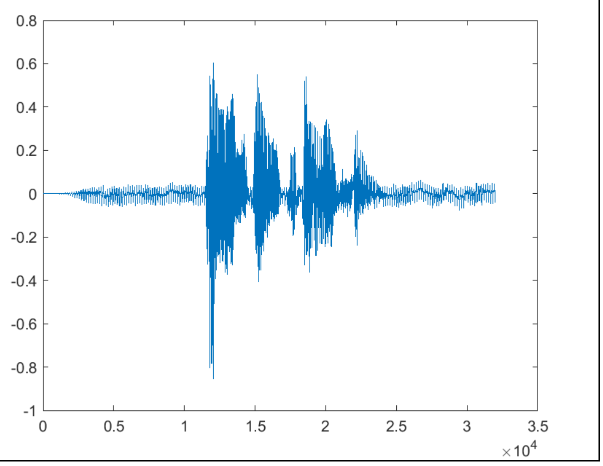
In the main problem part, the final accuracy was over 87% which we received through trialling different combinations of the number of coefficients, number and size of the frames, and size of frame-shifts. 100% accuracy could not be achieved as our dataset had some faulty samples. Moreover, Words with close pronunciation show wrong results, which would have been solved if we had worked with the features of the phonemes rather than whole words.

2. Bonus Problem 1:

In the bonus 1 problem, proper result was not found. The main reason is we could not find any general threshold value. Threshold was changing with loudness, device, environment etc. Though it is not practical, a solution can be calibrating the threshold whenever used. Loudnessness normalization could be a method too.

3. Bonus Problem 2:

For the condition "mean of absolute amplitude > 0.07*maximum of absolute amplitude of the frame", when input voice is not properly given and becomes impulse like (really high peak but localized in small place), no frame proceeds to the next steps and all the matrixes remains empty; hence the runtime error.

Shows runtime error	No error + correct detection
	

This can be easily solved by adding a new step after 3.3-IV:

- If no frames met the condition, declared "voice was not properly inputted" and retake the input voice from the user.

List of References:

- Hasan, M.R., Hasan, M.M.,Hossain, M.Z., How many Mel-frequency cepstral coefficients to be utilized in speech recognition? A study with the Bengali language. J. Eng. 2021, 817–827, 2021.

https://doi.org/10.1049/tje2.12082

- S. Dhingra, G. Nijhawan and P. Pandit, Isolated Speech Recognition using MFCC and DTW, International journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering,8(2), 2013 **https://www.ijareeie.com/upload/2013/august/20P_ISOLATED.pdf**
- P. P. Shingh, P. Rani, An Approach to Extract Feature using MFCC, IOSR Journal of Engineering, 8(4) , 2014
- H. Shaikh1, L. C. Mesquita2, S. D. C. S. Araujo, Recognition of Isolated Spoken Words and Numeric using MFCC and DTW, IJSEC, 4(7), 2017