Cairo University

Faculty of Engineering

DSP Project 2

Submitted to:

Dr. Mohsen Rashwan

Name	Sec	BN	E-Mail
أحمد رضوان جاد الرب عبداللاه	1	12	ahmed.abdellah991@eng-st.cu.edu.eg
أحمد طه عبد الحليم عبد النبي	1	19	ahmed.abdulanbi98@eng-st.cu.edu.eg
أحمد محمد مكرم ضاهر	1	26	ahmad.daher01@eng-st.cu.edu.eg
محمد إسماعيل عامر عبد الجواد	3	42	mohamed.abelgawad99@eng-st.cu.edu.eg
عبد الرحمن السيد شعبان عبد الرحمن	3	2	abdulrahman.abdulrahman98@eng-st.cu.edu.eg

Speech Recognition System

Ahmed Radwan[1], Ahmed Taha[2], Ahmed Mohammad Makram[3], Abdulrahman Elsayed[4], Mohammad ismael[5]
Faculty of Engineering Cairo University
Electronics & Electrical Communications Engineering Department
[1] ahmed.abdellah991@eng-st.cu.edu.eg
[2] ahmed.abdulanbi98@eng-st.cu.edu.eg
[3] ahmad.daher01@eng-st.cu.edu.eg
[4] abdulrahman.abdulrahman98@eng-st.cu.edu.eg
[5] mohamed.abelgawad99@eng-st.cu.edu.eg

Abstract

Nowadays, Speech recognition is a leading technology as it has many applications in nearly every field from the everyday-use technologies of the ordinary user like mobile phones, computers and internet in general to the most complex technologies of the major institutions like security, organizing and even military-purposes. speech recognition is going through a series of 5 steps which we will talk about in detail later in this paper, but in summary all these five steps goes around analyzing the analog voice that goes into the system as an input to get the features from it then comparing this input through the features we got with a dataset that we established and trained before. in our project we try to take a firm steps in Arabic Speech Recognition which is still below the expected expectations for a language that is spoken by nearly 5 billion people all over the world . our project focuses mainly on catching the spelling mistakes in spelling the letters that is near to each other on the the way it is spoken like the tongue touches the gum in one and not in the other. the methodology goes through using Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time Wrapping (DTW) to compare input words with pre-recorded dataset which is trained before

i. Introduction

We can define the voice recognition process saying that it is a technique that's created to identify, distinguish and authenticate the voice by taking an input word or a speech And extracts the features from it. Then combaring that features with a predefined dataset . This comparison then gives us a result that says is that speech is true with respect to what the system expect and then taking an action based on this like refusing to enter a website of an institute for example or even launch a siren if it was in a security system . speech recognition is going through a series of 5 steps we will discuss it with respect to what we done in this project :

1- firstly we captures a user's utterance by taking an audio with strict specifications like specified sampling rate , mono-channel …etc , 2- this input utterance is an analog signal so we digitize it into digital signal so we can process it easily 3-then we convert them into the basic unit of utterance which is called phonemes 4- we map these phonemes we got –which represents what the system get - to their phonetic representation –which represents what the user likely wanted to say- . the previous two steps can be called feature extraction and for this we used Mel Frequency Cepstral Coefficients (MFCC) 5- finally, According to some presets like grammar, representaions of phonetics and more other. The system compares the input features with the features it has in the dataset using dynamic time warping (DTW) and gives us a list contains the best scores of candidates. In this project we formed a confusion matrix for each category of recordings (Males, females, children) showing a result of all the inputs with its output .

ii. Feature extraction using MFCC

As we said in the introduction, for the step of feature extraction we used Mel Frequency Cepstral Coefficients (MFCC) and preferred it upon other techniques like Linear Prediction Coefficients (LPC), Linear Prediction Cepstral Coefficients (LPCC), (LSF) , (DWT) and others as it seems to be more effective regarding to some papers.

Mel Frequency Cepstral Coefficients:

(MFCC) was firstly provided as an artificially simulation for the hearing system of human beings on the assumptions that it deals with monosyllabic voice that's why all the voices we trained our project on was mono. based on this , mfcc tends to work for low frequencies as the human system is more sensitive to lower than higher frequencies **[4]**

The MFCC technique in summary can be expressed with these steps: divide the signal into windows using overlapping Hamming windows to frame the signal , applying Fast Fourier Transform (FFT) to it to calculate the power spectrum for each frame, multiplying each frame with our Mel-filter bank of 20 filters where we warps frequencies of the signal on, then we get the magnitude and apply a log to it ,and finally applies the inverse Discrete cosine transform (DCT).

Figure 1 shows these steps we talked about :

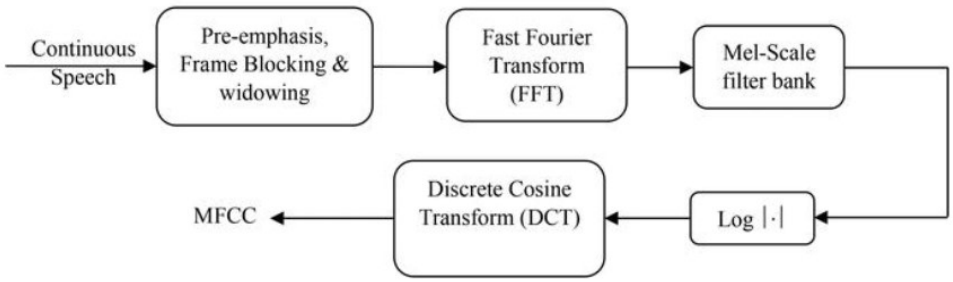


Figure 1

To calculate the mel scale value at a frequency this formula is used :

$$mel(f)=2595*\log10(1+f/700)$$

Figure 2 shows the frequency response (Mel basis functions) using 20 Mel filters, whish shows obviously that they are densed at low frequencies, and sparsed at higher frequencies.

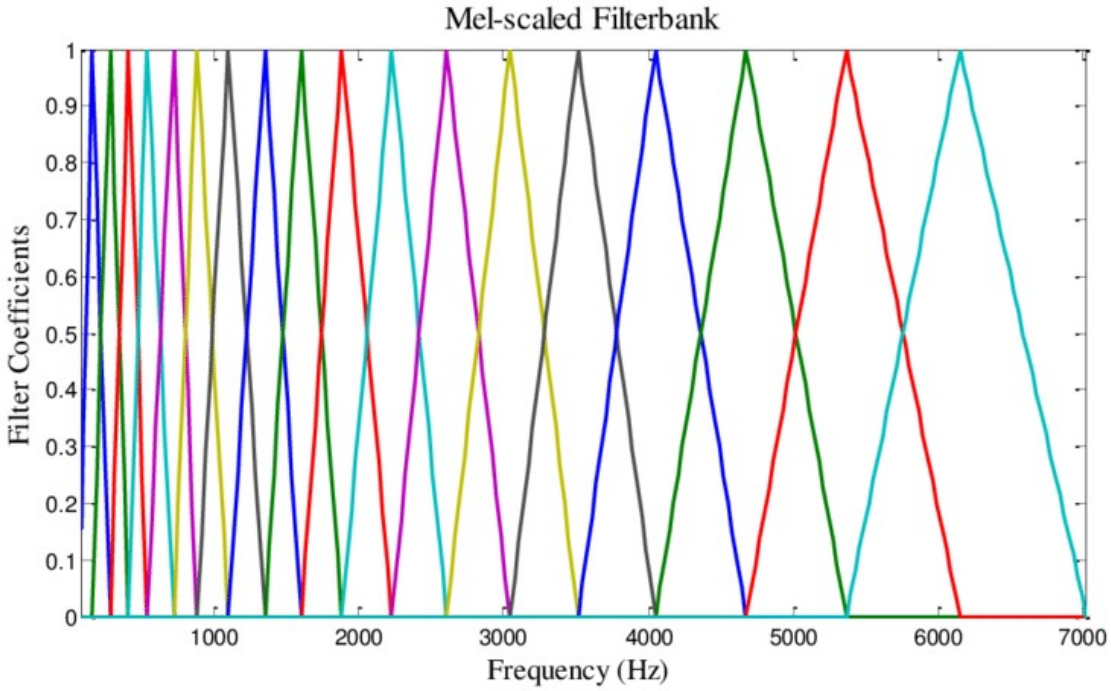


Figure 2

One note to take is that MFCC features is more less accurate if there was a background noise and may make a problems with generalization, that's why in our project we were very careful that all the data we trained the system on was without noise

iii. comparing the similarity:

In this project we used Dynamic Time Warping as a technique to find similarity or calculate the distance between the input signal and what we have in dataset.

To discuss how DTW works regarding to our work we need first to take a brief about what happens before it to know what is the input data it deals with.

The flow of our work is as follows: 1-fistly, we do an error handling script to filter the data and allow only the data with the specifications we specified while training the system like accepting only mono channel signal, rejecting the input with length bigger or smaller than our dataset samples length, Covert the signal from .wav to MFCC so we can process it finally we load the data to an array (test array) and simultaneously form the reference array from our dataset

The flow of DTW is that the system receives a test input signal, then calculates the global distance between the utterance (alhamdullilah) from this input and all of our references from the dataset and based on that it chooses which reference that has the lower DTW as the recognized speech.

After that we split the users to males, females and children and formulate a confusion matrix for each category showing a result of all the inputs with its output (showing how many users spelled the words true and how many of them spelled it wrong

iv. Working with dataset:

in this project we worked hard to accumulate the data needed, the data is 64 pair of words which is spoken by many people, this data is spitted in three main categories: males , females and children. we labeled the data in a manner that make it easy to analyses and process it after that, for example the full label is like G10S5M21WP55W2R where G10 means: Group 10, S5: Student 5, M: Male

Speaker ,21: the age of the speaker: using WhatsApp recording, P55: Pair 55 of words, W2: Word 2, R: Reference Speaker.

v. implementation and results

in this section we will talk in a brief manner about our implementation, how we coded it, what is really does, the challenges encountered us, the results we got and was it as we expected to be look at as accepted results or not.

Firstly, we started with writing a script its major goal is to clean the data and In more detail, the script had to loop over all the input records and check them to insure it meets the primary specification of the project ,checking if it's in mono or stereo and accept mono only and also checking the number of words to be exactly 123 for each user, and when we find that the user recordings does not following these rules, we regretfully delete their files as these files is considered outlined and misleading and if we keep it it will corrupt the whole model and makes it harder to predict a good manner,

The second step is trying to extract the threshold and here is where most of the challenges as it will be known while walking through forthcoming paragraphs. firstly, we started with a random data and making all the effort trying get their threshold but as the records aren't quite good- due to the great diversity in it from the recording background to the way every recording spelling each word-, we had a hard time finding the threshold from the graphs of the recordings , we didn't just try one approach but we also tried many of approaches suggested (like selecting a subset of the features that may give better overlap) but it didn't help a lot and the there was a big overlap as in (Figure 3). so we tried another utterly different approach which finally gave a better result, the idea of this approach is lying around playing with the distance making it large enough to be acceptable as a threshold , it look at the distance between the word spelled by one user and the same word from another user and it finds that it is very small while the distance required between the word and its pair word must be larger , and this exactly what we had in most cases , it stores the distance in 2 different arrays and sort them (ascending for the same word and descending for the word and its pair word),then it takes the first 20 elements in these arrays and extract the threshold from them, and we finally found it to be much easier to extract the thresholds this way and happily getting almost no overlapping as shown in (Figure 4) which means that this step is done with a very good results and now this thresholds is

ready to be used for the program getting utterly new data. In this final step we took an input recording and applied the thresholds to them and saving the results. To be more clear we organized these results in a confusion matrix showing the correct and wrong words and we also separated the males, females and children as for results to be more organized and reflecting the results in a good manner for any further work on them.

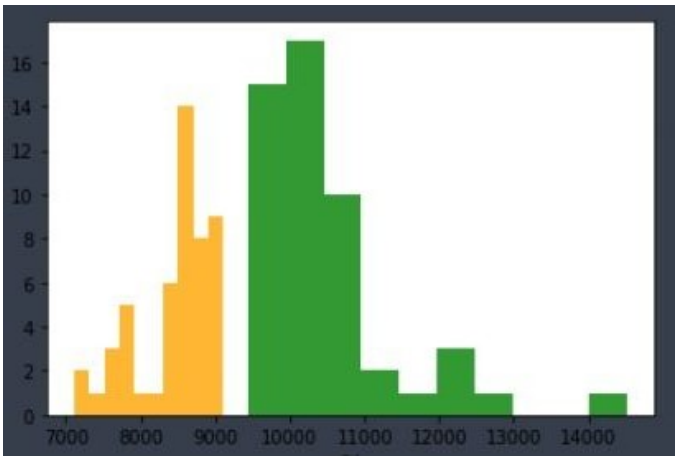


Figure 4

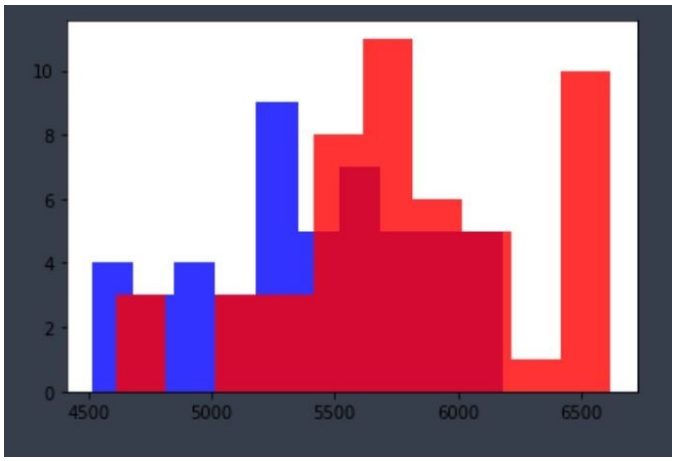


Figure 3

Now talking on results: the results was different between males, females and children, that's because the recordings isn't standardized in a manner that makes it near to each other in the circumstances of recording like background noise level the loudness, the time took in each utterance ...etc so the accuracies we got were as following:

For	Accuracy
Males	75.60 %
Females	62.50 %
Children	69.8 %

as it is obvious, it was a very good results for males while the accuracy for females wasn't that good and the children was between them and we can consider it a good result relating to the issues we talked about previously that is related to the datasets provided and is beyond our control.

We also made a GUI application to do the previous work as a friendly-use app

And (Figure 5) shows the mismatch plot for 5 mismatched words

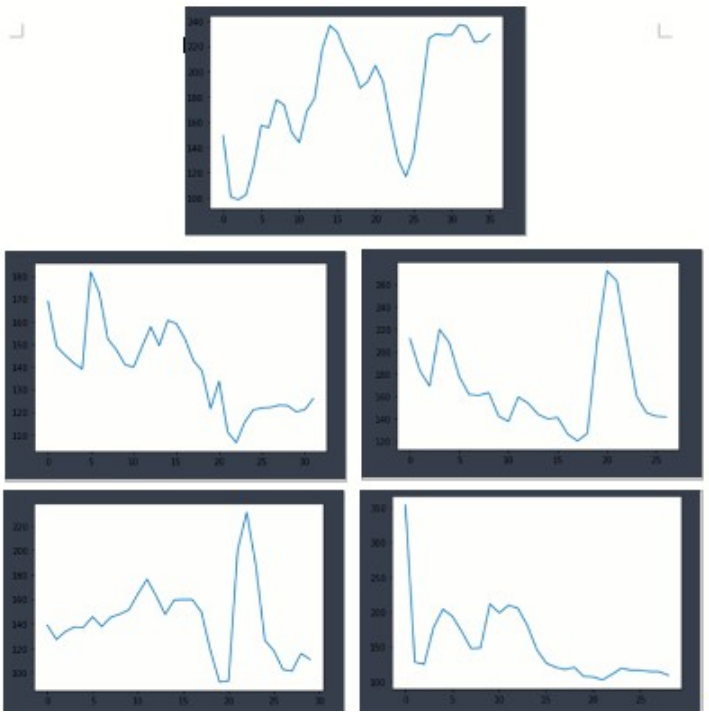
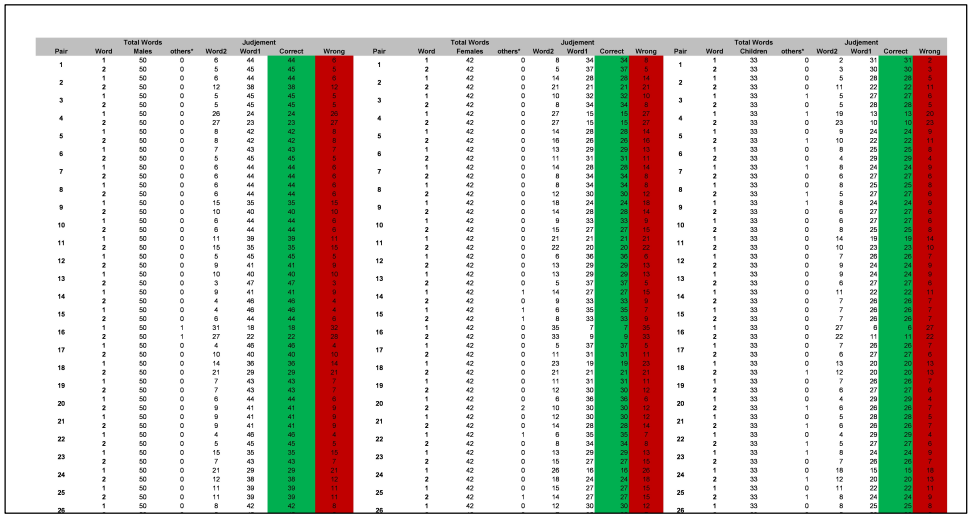


Figure 5

(Figure 6) is the required Confusion matrix for some of the words and the total accuracy for males, females and children, you can find the whole Matrix attached with the paper



	Total Words			Judjement				Total Words			Judjement				Total Words			Judjement		
Pair	Word	Males	others*	Word2	Word1	Correct	Wrong	Pair	Word	Females	others*	Word2	Word1	Correct	Wrong	Pair	Word	Children	others*	Word2	Word1	Correct	Wrong
1	1	50	0	6	44	44	6	1	1	42	0	8	34	34	8	1	1	33	0	2	31	31	2
	2	50	0	5	45	45	5		2	42	0	5	37	37	5		2	33	0	3	30	30	3
2	1	50	0	6	44	44	6	2	1	42	0	14	28	28	14	2	1	33	0	5	28	28	5
	2	50	0	12	38	38	12		2	42	0	21	21	21	21		2	33	0	11	22	22	11
3	1	50	0	5	45	45	5	3	1	42	0	10	32	32	10	3	1	33	1	5	27	27	6
	2	50	0	5	45	45	5		2	42	0	8	34	34	8		2	33	0	5	28	28	5
4	1	50	0	26	24	24	26	4	1	42	0	27	15	15	27	4	1	33	1	19	13	13	20
	2	50	0	27	23	23	27		2	42	0	27	15	15	27		2	33	0	23	10	10	23
5	1	50	0	8	42	42	8	5	1	42	0	14	28	28	14	5	1	33	0	9	24	24	9
	2	50	0	8	42	42	8		2	42	0	16	26	26	16		2	33	1	10	22	22	11
6	1	50	0	7	43	43	7	6	1	42	0	13	29	29	13	6	1	33	0	8	25	25	8
	2	50	0	5	45	45	5		2	42	0	11	31	31	11		2	33	0	4	29	29	4
7	1	50	0	6	44	44	6	7	1	42	0	14	28	28	14	7	1	33	1	8	24	24	9
	2	50	0	6	44	44	6		2	42	0	8	34	34	8		2	33	0	6	27	27	6
8	1	50	0	6	44	44	6	8	1	42	0	8	34	34	8	8	1	33	0	8	25	25	8
	2	50	0	6	44	44	6		2	42	0	12	30	30	12		2	33	1	5	27	27	6
9	1	50	0	15	35	35	15	9	1	42	0	18	24	24	18	9	1	33	1	8	24	24	9
	2	50	0	10	40	40	10		2	42	0	14	28	28	14		2	33	0	6	27	27	6
10	1	50	0	6	44	44	6	10	1	42	0	9	33	33	9	10	1	33	0	6	27	27	6
	2	50	0	6	44	44	6		2	42	0	15	27	27	15		2	33	0	8	25	25	8
11	1	50	0	11	39	39	11	11	1	42	0	21	21	21	21	11	1	33	0	14	19	19	14
	2	50	0	15	35	35	15		2	42	0	22	20	20	22		2	33	0	10	23	23	10
12	1	50	0	5	45	45	5	12	1	42	0	6	36	36	6	12	1	33	0	7	26	26	7
	2	50	0	9	41	41	9		2	42	0	13	29	29	13		2	33	0	9	24	24	9
13	1	50	0	10	40	40	10	13	1	42	0	13	29	29	13	13	1	33	0	9	24	24	9
	2	50	0	3	47	47	3		2	42	0	5	37	37	5		2	33	0	6	27	27	6
14	1	50	0	9	41	41	9	14	1	42	1	14	27	27	15	14	1	33	0	11	22	22	11
	2	50	0	4	46	46	4		2	42	0	9	33	33	9		2	33	0	7	26	26	7
15	1	50	0	4	46	46	4	15	1	42	1	6	35	35	7	15	1	33	0	7	26	26	7
	2	50	0	6	44	44	6		2	42	1	8	33	33	9		2	33	0	7	26	26	7
16	1	50	1	31	18	18	32	16	1	42	0	35	7	7	35	16	1	33	0	27	6	6	27
	2	50	1	27	22	22	28		2	42	0	33	9	9	33		2	33	0	22	11	11	22
17	1	50	0	4	46	46	4	17	1	42	0	5	37	37	5	17	1	33	0	7	26	26	7
	2	50	0	10	40	40	10		2	42	0	11	31	31	11		2	33	0	6	27	27	6
18	1	50	0	14	36	36	14	18	1	42	0	23	19	19	23	18	1	33	0	13	20	20	13
	2	50	0	21	29	29	21		2	42	0	21	21	21	21		2	33	1	12	20	20	13
19	1	50	0	7	43	43	7	19	1	42	0	11	31	31	11	19	1	33	0	7	26	26	7
	2	50	0	7	43	43	7		2	42	0	12	30	30	12		2	33	0	6	27	27	6
20	1	50	0	6	44	44	6	20	1	42	0	6	36	36	6	20	1	33	0	4	29	29	4
	2	50	0	9	41	41	9		2	42	2	10	30	30	12		2	33	1	6	26	26	7
21	1	50	0	9	41	41	9	21	1	42	0	12	30	30	12	21	1	33	0	5	28	28	5
	2	50	0	9	41	41	9		2	42	0	14	28	28	14		2	33	1	6	26	26	7
22	1	50	0	4	46	46	4	22	1	42	1	6	35	35	7	22	1	33	0	4	29	29	4
	2	50	0	5	45	45	5		2	42	0	8	34	34	8		2	33	1	5	27	27	6
23	1	50	0	15	35	35	15	23	1	42	0	13	29	29	13	23	1	33	1	8	24	24	9
	2	50	0	7	43	43	7		2	42	0	15	27	27	15		2	33	0	7	26	26	7
24	1	50	0	21	29	29	21	24	1	42	0	26	16	16	26	24	1	33	0	18	15	15	18
	2	50	0	12	38	38	12		2	42	0	18	24	24	18		2	33	1	12	20	20	13
25	1	50	0	11	39	39	11	25	1	42	0	15	27	27	15	25	1	33	0	11	22	22	11
	2	50	0	11	39	39	11		2	42	1	14	27	27	15		2	33	1	8	24	24	9
26	1	50	0	8	42	42	8	26	1	42	0	12	30	30	12	26	1	33	0	8	25	25	8

Figure 6

vi. Conclusion and discussion

In this paper we talked about how we constructed an Arabic Speech Recognition System , and we showed in the sections above that our system that relies upon MFCC and DTW can do a great job identifying the errors in spelling words and outputting the accuracies and a confusion matrix of all categories .

Refrences :

[1] Chun-Feng Liao Understanding the CMU Sphinx Speech Recognition System National Chengchi University

[2] Yusnita , Paulraj , Sazali Yaacob, Yusuf1 and Shahriman analysis of accent-sensitive words in multi-resolution mel-frequency cepstral coefficients for classification of accents in malaysian english

[3] 1Yurika Permanasari, 2Erwin H. Harahap, 3Erwin Prayoga Al Speech recognition using Dynamic Time Warping (DTW) Departement of Mathematic, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Islam Bandung

[4] Divyesh S. Mistry, Prof. A. V. Kulkarni, 2013, Overview: Speech Recognition Technology, Mel- frequency Cepstral Coefficients (MFCC), Artificial Neural Network (ANN), international journal of engineering research & technology (ijert) volume 02, issue 10 (october 2013),