REHABILITATION OF SPEECH IMPAIRED-PERSONS FOR ARABIC LANGUAGE SPEAKERS

By

Amira Yehia Anwar
Dalia Lotfy Abdelhay
Radwa Saeed Elmohamedy
Tarek Mohamed Rashad
Mayar Tarek Hasan

A Thesis Submitted to the
Faculty of Engineering at Cairo University
in Partial Fulfillment of the
Requirements for the Degree of
BACHELOR OF SCIENCE
in
Systems and Biomedical Engineering

Under the Supervision of

Prof. Dr. Ahmed Hisham Kandil
Associate Professor
Systems and Biomedical Engineering
Faculty of Engineering, Cairo University

FACULTY OF ENGINEERING, CAIRO UNIVERSITY
GIZA, EGYPT
2022

Acknowledgments

Throughout the process of working on this project, and the writing of this book we received a great deal of support and assistance.

We would like to thank our supervisor, Professor Ahmed Hisham, for his support and encouragement, his feedback pushed us to sharpen our thinking and brought our work to a higher level.

Awards and participation

We would like to announce that we had qualified to the final stage of International Science and Engineering Innovations Competition (ISEIC'2022) at March 23 - 24, 2022, Air Defense College (ADC), Alexandria, Egypt. And our project had been stated as one of the winners' projects.

Table of Contents

List of Tables

List of Figures

Abstract

Rehabilitation services are needed for people who have lost the ability to speak normally, often because of an injury, a stroke, an infection, a tumor, surgery, or a progressive disorder. The goal of rehabilitation is to establish the most effective means of communication. People having problems in pronouncing some letters are facing many challenges as their speech wouldn't be totally understandable and that may cause psychological issues. So, they will not be able to deal with society in effective way.

Our application for speech-impaired persons consists of two modes: The first mode is coach mode (frequently repeating words to the person). The patient has to select, listen, and re-pronounce selected word to practice on follow up the progress of patients and the second mode is assistant mode (speaking words and presenting objects that can be seen) help people regain some ability to use language. This part is done by selecting a category from different categories (names, objects, food, ..., etc.) which had been stored in database and playing the audio associated with the category selected in order to help stroke Alzheimer people to remember things, family and different objects and contribute with society in an effective way.

The performance of Speech Recognition System is mainly depending on the quality of Signal Preprocessing Stage. The Preprocessing quality is giving the biggest impact on the Speech Classification performance. Signal Pre-processing consist of removing silence and normalize signals. Processing stage consist of extracting Mel-Frequency Cepstral Coefficients (MFCC). MFCC is more preferred in Feature Extraction technique as it generates the training vectors by transforming speech signal into frequency domain, and therefore it is less affected by noise. Then, we apply Dynamic Time Warping (DTW), an algorithm for achieving similarity between to temporal sequences. An Improvement in any individual part can improve the overall system performance. For effective working of Back-End there should be more efforts in Front-End processing. So, we are trying to improve the accuracy to get the better results.

Chapter 1 : Introduction

Speech is the semantic element of human communication through which humans convey their message to each other. The Arabic language is the second most spoken language in terms of the number of speakers (more than a billion people worldwide) [1]. Unfortunately, about 30.8% of the children in Egypt suffer speech disorders [2]. People having problems in pronouncing some letters are facing many challenges as their speech wouldn't be totally understandable and that may cause psychological issues as they will lose self-confidence and will not be able to deal with society anymore. This problem is obvious in children, too.

1.1. Some of speech impairment cases:

- Aphasia: Aphasia patients may have trouble speaking, writing, reading, and interpreting language. When language-processing regions of your brain are harmed by a stroke or other trauma, the disorder may emerge.
- Apraxia: People with apraxia are aware of what they want to express but struggle to put it into words. Reading, writing, swallowing, and other motor skills may be difficult for them.
- Articulation disorders: Certain word sounds cannot be produced by kids with articulation issues. For example, they may substitute one sound for another — like saying "وامى" instead of "رامى" or "ثناء" instead of "سناء,". With articulation difficulties, early intervention speech therapy can be helpful.
- Cognitive-communication disorders: Communication problems can occur when the part of the brain that regulates thinking is disrupted. Cognitive-communication impairments can affect a person's ability to listen, speak, remember information, and solve problems.
- Dysarthria: Some neurological conditions, such as a stroke, MS, ALS, or multiple sclerosis, can cause the muscles that govern your speech to weaken. Dysarthria patients may speak slowly or erratically.
- Expressive disorders: People with expressive disorders may have difficulty getting words out or conveying their thoughts. Expressive disorders are linked to stroke or other neurological events, developmental delays or hearing loss.
- Fluency disorders: Fluency disorders disrupt the speed, flow and rhythm of speech. Stuttering (speech that’s interrupted or blocked) is a fluency disorder. So is cluttering (speech that’s merged together and fast).

So, trying to get suitable treatment for children at an early age will be more effective. But unfortunately, most parents don’t have enough awareness of this problem in their children. Even those who are aware of taking their children to a speech therapist, their children show great grumbling before each session, which puts pressure on parents to convince children about the importance of going to these therapists, which often leads to failure.

1.2. Computer-Aided Language Learning (CALL)

Therefore, in the last decade, Computer-Aided Language Learning (CALL) has received considerable attention due to its adaptability, which enables people at their own place to improve their language abilities. Despite the diversity of factors posing pronunciation disorders (vocal pathologies, stroke, Alzheimer, psychological state, age, etc.), no work has been extended to identify these factors and to assist speakers with pronunciation defects specially in Arabic language [3].

1.3. Application

Our work in this paper is divided into two main apps:

1.3.1. Speech coach app:

The main goal for speech coach App is based on speech similarity. This is done by allowing the user to record his voice and the app will repeat the word to him in a correct way and measure user's progress along period of time. This will help persons to participate in society in more efficient way.

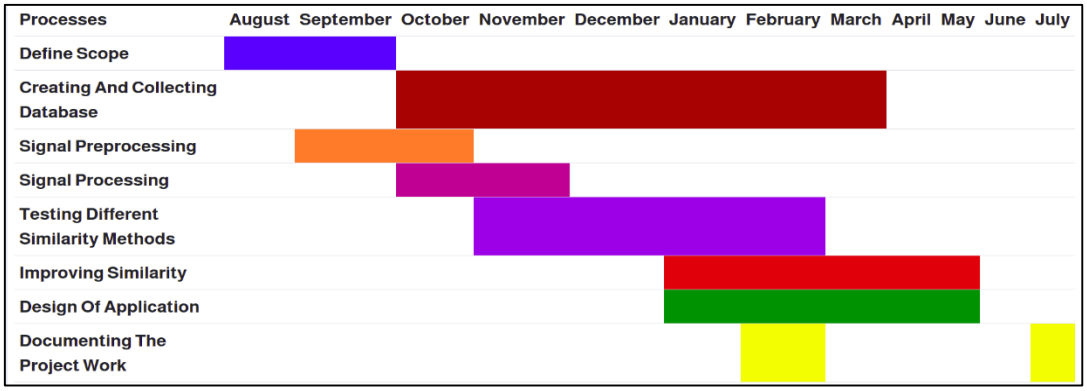
1.3.2. Speech Assistant app:

The second objective (Assistant speech app) targets people who suffering from Alzheimer's. Alzheimer's is a general term for memory loss and other cognitive abilities serious enough to interfere with daily life. Alzheimer's disease accounts for 60-80% of dementia cases [4]. People suffering from this disease don't have the ability to form complete sentence in a correct way. This app will display many categories for the user to choose from them according to what he wants to say, and the app will help him and say the complete sentence for the user in order to convey the speech that cannot be delivered correctly to others.

1.4. Roadmap of project

The following figure illustrates the application roadmap to visualize our application strategy and provide a high-level overview of the software development process.

Figure 1.4 Roadmap of project



Chapter 2 : Market Research

2.1. Primary research:

For our primary research, we have gathered data from our friends, family, and relatives about problems facing their children who have speech disorders in addition to visiting more than one clinic/hospital like Al-Kasr Al Aini. They all agreed to the need of such application for speech therapy.

Not only that but also most of the elderly in our families faced speech disorders like Alzheimer, dementia. From all these deductions, we noticed the need for having an application that tries to solve the problems mentioned in the previous chapter and save time for both patient and therapist.

2.2. Secondary research:

The global voice and speech recognition market size was valued at USD 14.42 billion in 2021 and is anticipated to grow at a compound annual growth rate (CAGR) of 15.3% from 2022 to 2030. The growing use of voice biometrics is among the major factors driving the market growth and the use of mobile applications in speech therapy is also growing as an avenue to bring treatment into the home.

So, the target customers for our application are mainly patients who suffer from speech impairment like (Aphasia, Apraxia, Articulation disorders, Alzheimer, Cognitive-communication disorders, Dysarthria, etc.), speech therapists, hospitals and any healthcare facility providing speech therapy sector.

Speech and voice recognition technologies are mostly used in the healthcare sector to report health checkups, data entry, and when the doctor or the attendant/nurse is unavailable. Such software solutions enable healthcare professionals to enter notes into the electric health record (EHR) system or their computers without taking time out from patient care and remain productive throughout the day. Easy to use and hands-free features of an automated speech recognition system in medical applications enable doctors to get their work done efficiently, driving the speech and voice recognition market growth. Thus, increased productivity leads to increased cash flow.

All language disorder applications in the market now don't support Arabic language so we can say that there are no competitors for us since we target the Egyptian market. But if we will talk about the global market of speech therapy and language disorder applications so here are list of these applications:

- **Articulation Station:** it provides multiple levels and categories that help users evolve their speaking skills. (Owned by small app development company called little bee speech)
- **Bitsboard:** it provides flashcards that can help users communicate with others.
- **WORD VAULT:** it provides fun games and interactive tools that help users practice their speaking skills. (Owned by HomeSpeechHome)
- **ArtikPix:** it provides fun card matching game so users can have fun while practicing their speaking skills. (Owned by Expressive Solutions)
- **Speech Blubs:** it provides videos of other children speaking that help other children imitate what they see.

2.3. SWOT analysis: **Table 2.3 SWOT analysis**

	Pros	Cons
	Strengths	**Weaknesses**
Internal	• Easy to use. • Enable therapists to get their work done efficiently. • The first application to Support Arabic language as basic language in speech recognition field. • Increase the number of patients the speech therapist can receive a day. • Decrease the number of offline speech therapy sessions for every patient. • Save a lot of time for both of the therapists and patients. • Increase productivity and obtain more profitability for the therapist or hospital. • It can help children who are shy and do not like to talk to people they don't know • Cost effectiveness.	• The application accuracy still low and need to be improved. • Limited database of the application. • Lack of specialization and testing of systems. • Competitors can offer similar products quickly. • Limited market • Lack of domain experts to evaluate its feasibility • Some training time is required initially. • Failure in case of background noise. • Less reliable if voice changes in time of distress. • Lack of activities and fun for kids may make it less attractive for them to use.
	Opportunities	**Threats**
External	• Applying this technology will drive the speech and voice recognition market growth in Egypt. • Using more effective techniques for creating the reference dataset will increase the accuracy of the application. • Adding an emotional human element to the design • Development of the application to support the Arabic language for other Arab countries • Integrating with google assistant to provide a more recognizing system for people with speech impairment • Development of the application to provide group therapy services. • Applying more advanced machine learning algorithms to improve accuracy of the application. • Updating the application continuously and increasing features to meet customer needs.	• Child protection issues. • Lack of trust • Being less effective in speech therapy than traditional offline used techniques. • Potential copycats

Chapter 3 : Table 3 Literature Review

Author/ Date	Year	Data Source	Methodology	Analysis & Results	Comments
Shaikh Naziya S.1, R.R. Deshmukh2	2016	Not mentioned	speech signal is transformed into sequence of feature vectors by different speech processing techniques. It converts feature vector to phoneme lattice by applying algorithms. A recognition module transforms the phoneme lattice into a word lattice by lexicon and then grammar is applied to word lattice to recognize the specific words or text.	In this study, total seven different approaches which are widely used for SRS have been discussed and after comparative study of these approaches it is concluded that Hidden markov method(HMM) is best suitable approach for a SRS because it efficient, robust, and reduces time and complexity.	
Mohamed S.abdo A. H. Kandil Ahmed Mohammed El-Bialy S. A. Fawzy	December 2010	The well-recognized recitation rules of the Holy Quran. A database of correct readings for the chosen uttered letters was recorded from six famous referenced readers to be the control set of our system	An algorithm was developed for automatic segmentation for the phonetic segmentation of the phonetic unit from uttered speech to be verified. "Using *Delta-MFCCs based segmentation"*	the overall accuracy of automatic segmentation reached to 73%.	More experiments are needed to validate the system

Mohamed S.abdo Ahmed H kandil	November 2016	The recordings from the reader “Mahmoud Khaleel El-Hosary”. this choice is based on the well-known of his good realization for the rules of recitations of Quran verses.	The developed method employs the vector of local maxima picked from peaks of the delta function of first Mel Frequency Cepstrum Coefficient as cutting tools that predict possible location of syllables boundaries inside the continuous speech	The results have shown that the system was able to break up a set of 276 Arabic utterances into its syllables with up to 91.5 % accuracy	This is a great step in speech recognition processes
FARIA NAZIR , MUHAMMAD NADEEM MAJEED , MUSTANSAR ALI GHAZANFAR , AND MUAZZAM MAQSOOD	March 22, 2019	Not mentioned	In their paper, they developed three models . In HandCrafted_Features model, we take an audio dataset and extracted handcrafted features from audio files and pass them to classifiers like KNN, SVM, and NN to detect mispronunciation	The experimental results show that handcrafted_features method, CNN_features, and transfer learning-based method achieve an accuracy of 82%, 91.7%, and 92.2%, respectively. The performance analysis shows that transfer learning-based method outperforms handcrafted_features and transfer CNN_features-based methods and achieve an accuracy of 92.2%. The proposed transfer learning-based method also outperforms the state-of-art techniques in term of accuracy.	

Divya Gupta, Poonam Bansal and Kavita Choudhary	January 2018	Not mentioned	The well known techniques of feature extractions and d advantages and disadvantages of it like LPC, MFCC, RASTA, PCA, LDA, PLP.	They attempt to provide a comprehensive survey of six feature extraction techniques which help to researchers in the field of automatic speech recognition area. We have also summarized the performance comparison of various ASR systems	Mfcc technique is the best way to feature extraction it,s accuracy is 94.35%
Nitin Washani Sandeep Sharma, Ph.D.	No. 18, April 2015	Not mentioned	This paper presents the advances made as well as highlights the pressing problems for a speech recognition system. The paper also classifies the system into Front End and Back End for better understanding and representation of speech recognition system in each part.	In this System End Point of speech utterance is detected by concept of Energy & ZCR and MFCC is used as feature extraction technique. Hence GMM gives a poor Recognition Rate (70%) as compared to other classifiers, but in case of Speaker Recognition it gives efficient results	The result can be improved by increasing the training data size.

Chapter 4 : Materials and Methods

4.1. Data collection

Data collection is a vital stage in any research. Our application is simply dependent on voice verification by comparing it to a reference database. So, searching for the reference database was the main step in our project. Because working with Arabic language is very rare, we didn't find any available audio database. So, we had to create an audio database by ourselves. The key steps in the data collection processes are:

- Identify the research issue that needs to be addressed and set goals for the project.
- Gather data requirements to deliver the research information.
- Identify the data sets that can provide the desired information.
- Set a plan for collecting the data, including the collection methods that will be used.
- Collect the available data and begin working to prepare it for analysis.

So, the first trial was to search for an API to help us. There were several APIs available to convert text to speech in Python. One of such APIs is the Google Text to Speech API commonly known as the gTTS API. gTTS is a very easy-to-use tool which converts the text entered, into audio which can be saved as an mp3 file. This API was supporting many languages except for Arabic which led to a different accent from ours as Arabic native speakers. So, this database didn't give us the desired output. As a result, we created the database with our voices.

As we mentioned before that our application has two modes of operation, so we needed different types and categories of datasets. The first one was the assistant mode which helps in creating a complete sentence. We created about 400 recorded audios from different categories (names, objects, food, etc.) which are mostly used by people in their daily activities. Another option is the availability for adding new words to this dataset as family members.

The second one was the coach mode which helps users to practice pronouncing words for speech therapy (speech-therapist's alternative). The dataset for this mode was gathered from number of volunteers with different genders and ages to assure the coverage of a wider range of frequencies. The number of volunteers was 182 and their ages were ranging from seven to sixty-five years old. They were categorized to sixty-six for males, sixty for females and fifty-six for children. The most frequent age in males and females was 22 years old while the most frequent age for children was 12 years old. The dataset is composed of 123 words divided into two groups which differ from each other by one letter. These words are selected according to the similarity between them to train the user how to differentiate between pronunciation of the different letters of the two words.

4.2. Signal preprocessing

4.2.1. Removing silence from audio file

We can use librosa.effects.trim() function from librosa library to remove silence from the signal. It will trim leading and trailing silence from an audio signal. After reading the audio file and converting it into frames, it checks voice activity detector VAD to each set of frames using Sliding Window Technique. The Frames having voices are collected in separate list and non-voices(silences) are removed. Hence, all frames which contain voices is in the list are converted into "Audio file".

4.2.2. Signal normalization:

Sometimes you'll have audio files where the speech is loud in some portions and quiet in others. Having this variance in volume can hinder transcription. We need to normalize a signal when we want to compare it with respect to other signals. Normalization means scaling the signals in identical level. It is a rescaling of the data from the original range so that all values are within the range of 0 and 1. If you normalize the signals in power level, that means all the signals have same power now.

Normalizing the amplitude of a signal is to change the amplitude to meet a particular criterion. One type of normalization is to change the amplitude such that the signal's peak magnitude equals a specified level.

Luckily, PyDub's effects module has a function called normalize() which finds the maximum volume of an AudioSegment, then adjusts the rest of the AudioSegment to be in proportion. This means the quiet parts will get a volume boost.

4.3. Signal Processing

4.3.1. Recording Audio File & storing it

At first, we need to record the audio of the user and save it to apply our functions on this record. Python can be used to perform a variety of tasks. One of them is creating a voice recorder. We can use python's sound device module to record and play audio. This module along with the wavio or the scipy module provides the way to save recorded audio.

you can use the scipy.io.wavfile module to store NumPy arrays as WAV files. The wavio module similarly lets you convert between WAV files and NumPy arrays. If you want to store your audio in a different file format, pydub and soundfile come in handy, as they allow you to read and write a range of popular file formats (such as MP3, FLAC, WMA and FLV).

Now, before starting the recorder, Parameters should be considered for each function:

The second is recording duration

The first one is the sampling frequency of the audio.

We used audio sample rate 44.1kHz. The most common audio sample rate you'll see is 44.1 kHz, or 44,100 samples per second. This is the standard for most consumer audio, used for formats like CDs.

This is not an arbitrary number. Humans can hear frequencies between 20 Hz and 20 kHz. Most people lose their ability to hear upper frequencies over the course of their lives and can only hear frequencies up to 15 kHz–18 kHz. However, this "20-to-20" rule is still accepted as the standard range for everything we could hear.

The computer should be able to recreate waves with frequencies up to 20 kHz in order to preserve everything we can hear. Therefore, a sample rate of 40 kHz should technically do the trick, right?

This is true, but you need a pretty powerful—and at one time, expensive—low-pass filter to prevent audible aliasing. The sample rate of 44.1 kHz technically allows for audio at frequencies up to 22.05 kHz to be recorded. By placing the Nyquist frequency outside of our hearing range, we can use more moderate filters to eliminate aliasing without much audible effect. And we used Frame size (size of FFT) calculated by multiplying the sample size in bytes by the number of channels, frame size is 1024 samples where 1024 byte(1KB) which is suitable for audio FFT, the Duration of each

frame calculated Frame size / sampling rate (sampling frequency) 1024(samples) / 44100KHZ or (samples/sec) 0.022 sec (Suitable from 20 to 30 msec) finally the hop size in our case equal 512. If we want to reduce the difference between neighboring frames, we can allow overlap between them. Usually, the overlap is 1/2 to 2/3 of the original frame. The more overlap, the more computation is needed .and you have to note that Number of frames increases as duration of audio file increases.[5]

And finally, we have to specify the duration in seconds so that it stops recording after that duration. In our case we sit the duration to 2 secs.

4.3.2. Reading the audio file & plotting the signal

Librosa is a python library that has almost every utility you are going to need while working on audio data. This rich library comes up with a large number of different functionalities.

We will use librosa.load() function from this library ,which supports lots of audio codecs, to read audio files. This function returns two things — 1. An array of amplitudes. 2. Sampling rate.

We have got amplitudes and sampling-rate from librosa. We can easily plot these amplitudes with time. Librosa provides a utility function waveplot()

This visualization is called the time-domain representation of a given signal. This shows us the loudness (amplitude) of sound wave changing with time.

These amplitudes are not very informative, as they only talk about the loudness of audio recording. To better understand the audio signal, it is necessary to transform it into the frequency-domain. The frequency-domain representation of a signal tells us what different frequencies are present in the signal.

4.3.3. Extracting Features for Speech Recognition

Before we talk about extracting features, we will talk about audio features and how they are categorized.

4.3.3.1. Audio Features:

Audio features are description of sound or an audio signal that can basically be fed into statistical or ML models to build intelligent audio systems.[6] Audio applications that use such features include audio classification, speech recognition, automatic music tagging, audio segmentation and source separation, audio fingerprinting, audio denoising, music information retrieval, and more.[7]

Different features capture different aspects of sound. Generally audio features are categorized with regards to the following aspects: [8][9]

- Level of Abstraction: High-level, mid-level and low-level features of musical signals.
- Temporal Scope: Time-domain features that could be instantaneous, segment-level and global.
- Musical Aspect: Acoustic properties that include beat, rhythm, timbre (color of sound), pitch, harmony, melody, etc. [10]

- Signal Domain: Features in the time domain, frequency domain or both.
- ML Approach: Hand-picked features for traditional ML modeling or automatic feature extraction for deep learning modeling.

In our paper, we will be interested in discussing the Signal domain and ML Approach categories' features.

4.3.3.2. Signal Domain features:

They consist of the most important or rather descriptive features for audio in general:[11]

- Time domain: These are extracted from waveforms of the raw audio. Zero crossing rate, amplitude envelope, and RMS energy are examples.
- Frequency domain: These focus on the frequency components of the audio signal. Signals are generally converted from the time domain to the frequency domain using the Fourier Transform. Band energy ratio, spectral centroid, and spectral flux are examples.
- Time-frequency representation: These features combine both the time and frequency components of the audio signal. The time-frequency representation is obtained by applying the Short-Time Fourier Transform (STFT) on the time domain waveform. Spectrogram, mel-spectrogram, and constant-Q transform are examples.

4.3.3.3. ML Approach features

Traditional Machine Learning approach considers all or most of the features from both time and frequency domain as inputs into the model. Features need to be hand-picked based on its effect on model performance. Some widely used features include Amplitude Envelope, Zero-Crossing Rate (ZCR), Root Mean Square (RMS) Energy, Spectral Centroid, Band Energy Ratio, and Spectral Bandwidth.

Deep Learning approach considers unstructured audio representations such as the spectrogram or MFCCs. It extracts the patterns on its own. By late 2010s, this became the preferred approach since feature extraction is automatic. It's also supported by the abundance of data and computation power.[7]

Commonly used features or representations that are directly fed into neural network architectures are spectrograms, mel-spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs). We will talk about them and some other features which we used in our study in details.

4.3.3.4. Time domain features:

A. Zero-crossing rate:

A zero-crossing is a point where the sign of a mathematical function changes (e.g. from positive to negative), represented by an intercept of the axis (zero value) in the graph of the function. It is a commonly used term in electronics, mathematics, acoustics, and image processing.

In alternating current, the zero-crossing is the instantaneous point at which there is no voltage present. In a sine wave or other simple waveform, this normally occurs

twice during each cycle. It is a device for detecting the point where the voltage crosses zero in either direction. [12]

Counting zero-crossings and dividing it by the length of the audio frame, which is called zero-crossing rate (ZCR), is also a method used in speech processing to estimate the fundamental frequency of speech which is being a key feature in time domain to classify percussive sounds. [13]

ZCR is defined formally as

$$ZCR = \frac{1}{2(M-1)} \sum_{n=1}^{M-1} |\text{sgn}[\text{x(n + 1)}] - \text{sgn}[\text{x(n)}]|$$

where sgn[…] shows the sign function and the discrete signal and x(n) represents the values ranging from n=1,……,M.

ZCR can be interpreted as a measure of the noisiness of a signal. For example, it usually exhibits higher values in the case of noisy signals.

B. Root-Mean-Square (RMS) Energy:

The square root of the mean of the square. RMS is (to engineers anyway) a meaningful way of calculating the average of values over a period of time. With audio, the signal value (amplitude) is squared, averaged over a period of time, then the square root of the result is calculated. The result is a value, that when squared, is related (proportional) to the effective power of the signal.

The energy of a signal corresponds to the total magnitude of the signal. For audio signals, that roughly corresponds to how loud the signal is. The energy in a signal is defined as:

$$\sum_n |x(n)|^2$$

The root-mean-square energy (RMSE) in a signal is defined as:

$$\sqrt{\frac{1}{N} \sum_n |x(n)|^2}$$

C. Short-Time Energy:

This feature has been used by many researchers in speech classification applications [14]. It has also been used in mispronunciation detection systems. Short time energy can be defined as:

$$\text{Em} = \sum_{n=-\infty}^{\infty} [\text{x(n)}\omega(\text{m} - \text{n})]^2$$

Here input signal is represented by $x(n)$, number of frames by m and window size by $\omega(n)$.

4.3.3.5. Frequency domain features:

A. Spectral Centroid

The spectral centroid indicates at which frequency the energy of a spectrum is centered upon or in other words It indicates where" center of mass" for a sound is located.

It is calculated as the weighted mean of the frequencies present in the signal, determined using a Fourier transform, with their magnitudes as the weights:[14]

$$Centroid = \frac{\sum_{n=0}^{N-1} f(n)\, x(n)}{\sum_{n=0}^{N-1} x(n)}$$

where x(n) represents the weighted frequency value, or magnitude, of bin number n, and f(n) represents the center frequency of that bin.

B. Spectral Roll-off

It can be defined as the action of a specific type of filter which is designed to roll off the frequencies outside to a specific range. This can be used for calculating the maximum and minimum by setting up the roll percent to a value close to 1 and 0.

C. Chroma features

The chroma feature is a descriptor, which represents the tonal content of a musical audio signal in a condensed form. Therefore, chroma features can be considered as important prerequisite for high-level semantic analysis, like chord recognition or harmonic similarity estimation. It provides a robust way to describe a similarity measure between music pieces.

4.3.3.6. Deep Learning approach features:

A. Mel-frequency Cepstral Coefficients (MFCCs)

It is the most widely used feature in speech and music classification applications. Different sounds can be easily classified by using MFCCs because of its discriminative ability. This discriminative property has led its use in CALL systems. It can be calculated for frames as well as for speech segments. [15] Steps to calculate MFCCs can be explained as; first of all, an audio signal is divided into frames to take Fourier transform.

Group 1	ب	ت	ث	ح	خ	ر	ز	ط	ظ	ف	هـ	ي				
Group 2	أ	ج	د	ذ	س	ش	ص	ض	ع	غ	ق	ك	ل	م	ن	و

Table 4.3.3.6. shows details of Arabic phonemes divided into two groups based on their similarity

Then periodograms are estimated of the power spectrum for each frame, the logarithm of all energies is then taken followed by a Discrete Cosine Transform (DCT) of each Mel log power which gives MFCCs.

$$\sqrt{\frac{2}{k}} \sum_{K=1}^{K} (\log \mathrm{Sk}) \cos[\frac{\mathrm{n(k-0.5)\pi}}{k}] \ , where\ n = 1,2,3 \ldots . L$$

Here, K represents the number of band pass filters and L represents the number of MFCCs.

The cepstrum conveys the different values that construct the formants (a characteristic component of the quality of a speech sound) and timbre of a sound. MFCCs thus are useful for deep learning models.

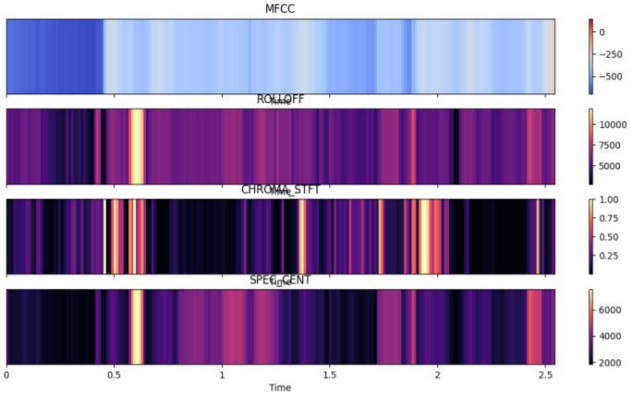


Fig 4.3.3.6. shows comparison between 3 features of frequency domain features and MFCC

4.4. Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is a well-known technique to find an optimal alignment between two given (time-dependent) sequences under certain restrictions. It is used for measuring similarity between two temporal sequences, which may vary in speed. Originally, DTW has been used to compare different speech patterns in automatic speech recognition. Other applications include speaker recognition and online signature recognition. It can also be used in partial shape matching applications. In fields such as data mining and information retrieval, DTW has been successfully applied to automatically cope with time deformations and different speeds associated with time-dependent data [3]. The optimal match is denoted by the match that satisfies all the restrictions and the rules and that has the minimal cost, where the cost is computed as the sum of absolute differences, for each matched pair of indices, between their values. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension. This sequence alignment method is often used in time series classification. Although DTW measures a distance-like quantity between two given sequences, it doesn't guarantee the triangle inequality to hold. In the following figure, Red lines connect corresponding time positions in the input signals.

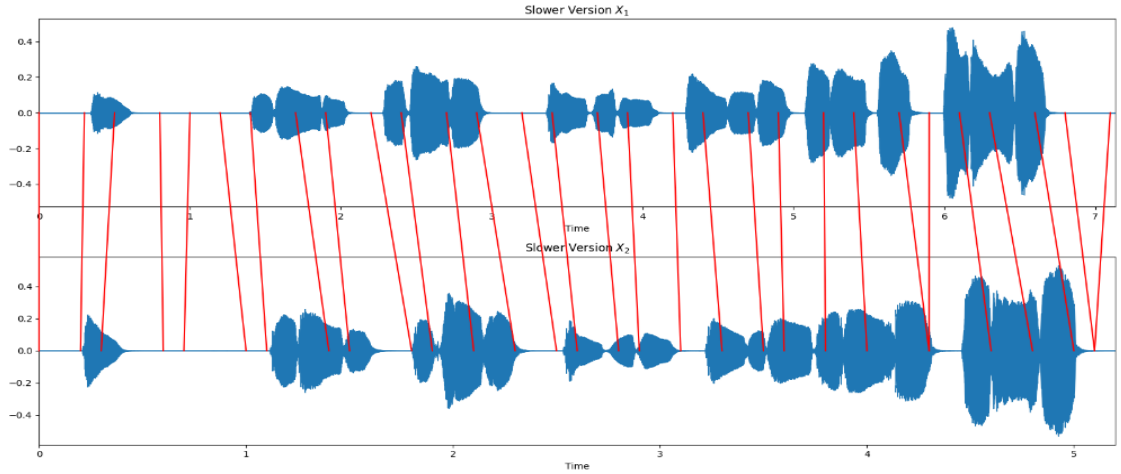


Fig 4.4. shows DTW matching of 2 signals vary in speed

Chapter 5 : Results

Our end result is an application with two modes characterized high accuracy and performance. We reached this result after applying stages of pre-processing and processing on the audio files to get the best form of signal that is amplificated without noise. And then we normalized that signal.

5.1. The main page

Figure 5.1 The main page

5.2. Login & create account

At this step, the user is required to provide his personal info in case of using the app for the first time:

- Gender
- Role
- Name
- E-mail
- Password

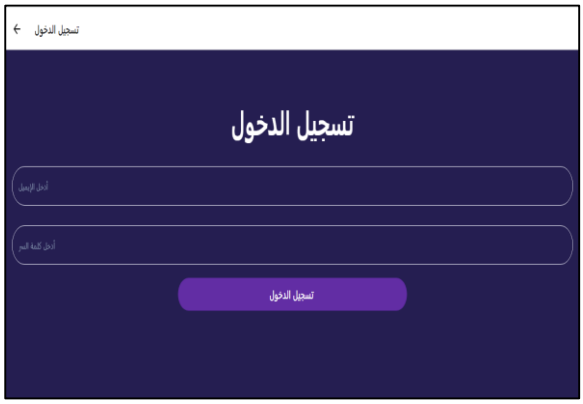


Figure 5.2.1 Login page

Figure 5.2.2 Create new account page

In case of patient: the user will be able to create a consistent system and once he successfully created an account that enables him to select the application mode he wanted. Now, user can start using the app.

Figure 5.2.3 Mode Selection page

5.3. Speech Coach Mode

This mode aims to help users who have problems in pronouncing some letters correctly to improve their pronunciation in some steps

1. Picking random word or name.
2. Playing this word to hear the correct pronunciation.
3. Starting to record this word with his/her own voice for 2 secs.
4. Saving it to make comparison between the reference file and the recorded file.
5. The app analyzes the input and compare between the recorded voice of the user and the correct pronunciation from the database.
6. Displaying the ratio of similarity between the 2 records in form of percentage.
7. Repeating this process until reaching acceptable percentage.

Figure 5.3 Speech Coach Mode

The user can show his/her progress for every word in pronouncing to know how much he improved and keeps track of his improvement. Also, the time of his practicing is recorded to know how frequently he practiced.

5.4. Signal Preprocessing and processing

During this process, we faced a problem in mapping the signals of speeches due to silence durations (in the beginning and end of the audio) so we trimmed it out of the audio signal to get more accurate result. Then we normalized the input signal as a pre-processing stage.

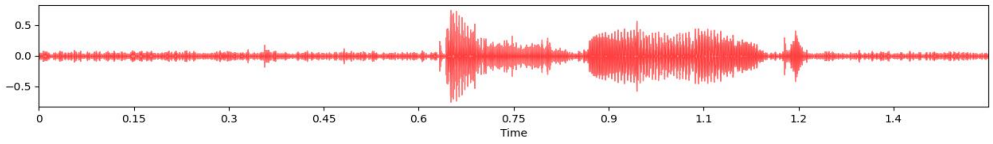


Figure 5.4.1 shows plotting of Ahmed (audio signal) without normalization

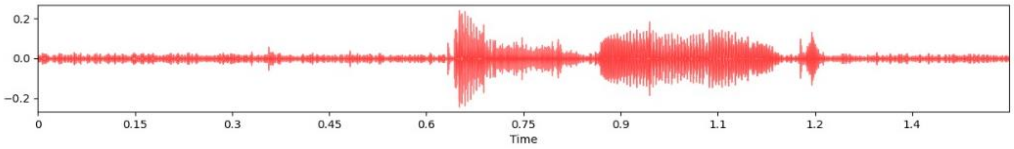


Figure 5.4.2 shows plotting of Ahmed (audio signal) with normalization

- Our code plays and plots the recorded signals and the similar to them in our database before getting the DTW.

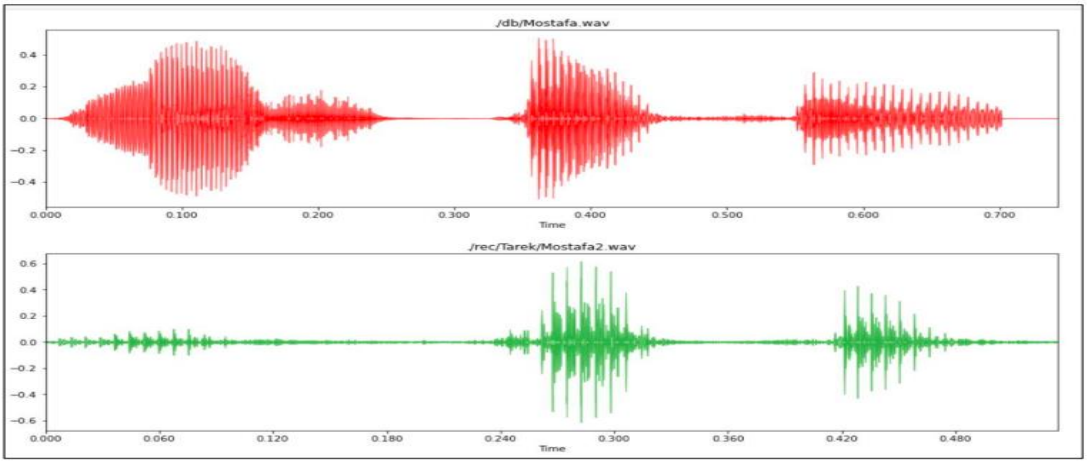


Figure 5.4.3 shows the plotting of two signals for the name مصطفى

- After we get the plotted signal of our recorded audio, we get the MFCC for it.
- Applying the DTW to our audio signal and all the audio files recorded by our voices with different names (حسن – أحمد – عيشة – أميرة – أشرقت – آية – مصطفى – نسرين إبراهيم) in the database to get the similarity (distance) between our file and the others.
- After getting the most similar audio to our recorded file, we will get the cost matrix and warping distance between their MFCCs features and plot it.

The left output shows the accumulated distance between them and the cost matrix and the warping path matrix.

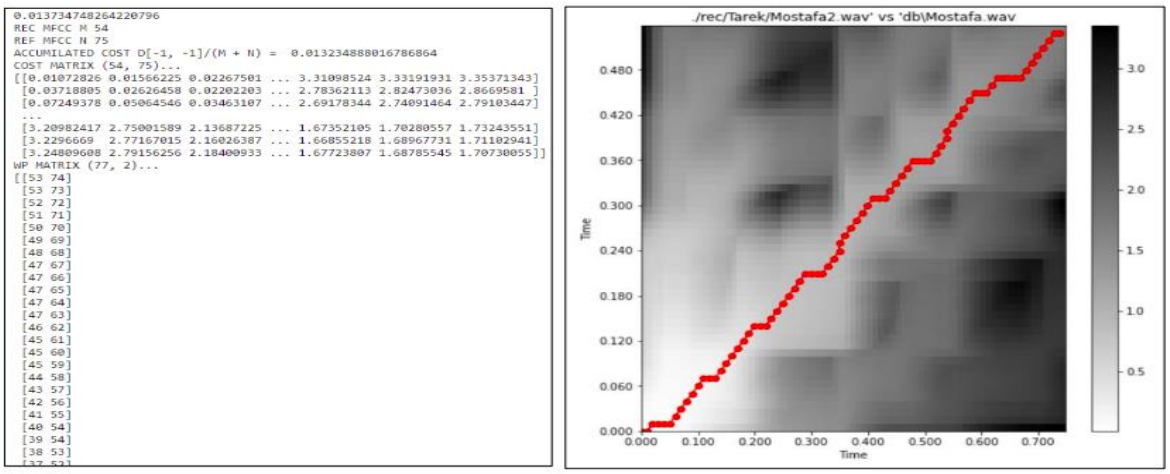


Figure 5.4.4 shows the plot of the best alignment between the two signals at the right side

Finally, we match the two plotted signals with each other to visualize similar durations (areas of every single phoneme in each plot)
Then we get the user's score and level of pronunciation of the word

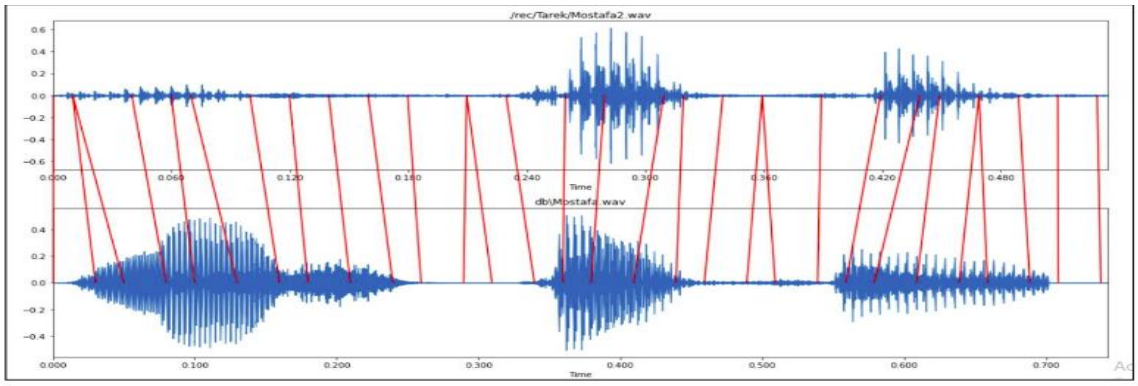


Figure 5.4.5 shows the similarity matching between two records of word (مصطفى) with different speed

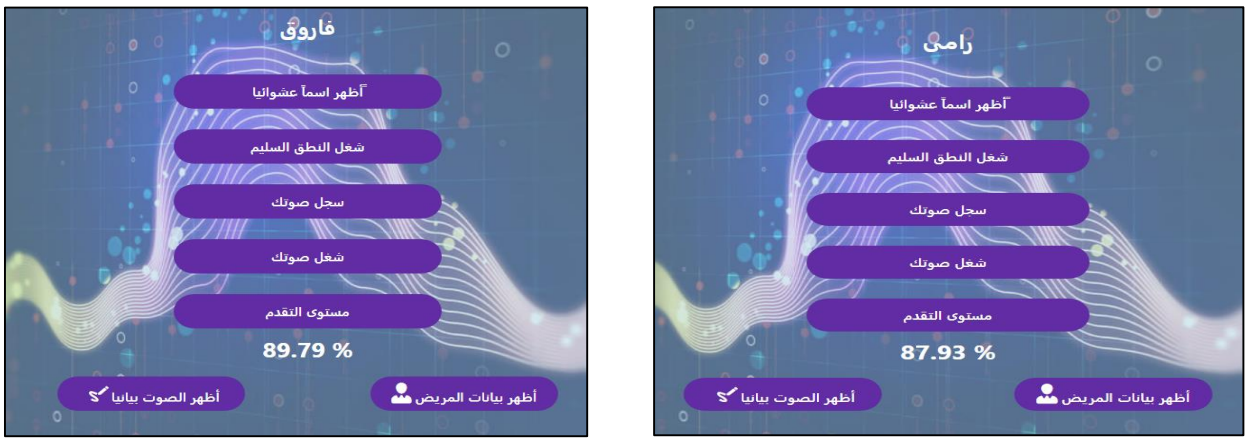


Figure 5.4.6 shows screen shots from our coach mode screen application shows progress of patient in recording word "فاروق"/"رامي"

5.5. Speech Assistant Mode

This mode aims to help users who have problems in remembering names of people they deal with daily and the words describing their need, especially users who have Aphasia (speech disorder after stroke), and help them to form a sentence from categories provided by the application. In this case the application doesn't take input voice from the user.
These categories consist of words from different activities which they are using it daily to help them improve communication with the society and fulfill their needs.
These categories are:

- Colors
- Bed room
- School
- Preposition
- Electric devices
- Rooms
- Kitchen
- Living room
- Persons
- Verbs
- Clothes
- Food

This application targets people who suffering from Alzheimer's. Alzheimer's is a general term for memory loss and other cognitive abilities serious enough to interfere with daily life. Alzheimer's disease accounts for 60-80% of dementia cases. People suffering from this disease don't have the ability to form complete sentence in a correct way.

The user has to choose his words from the categories that appear in front of him to collect his sentences which had been stored in database as mentioned previously. And then playing the audio associated with the category selected in order to help stroke or Alzheimer people to remember things, family and different things.
After forming his sentence, he can listen to it or delete it and then create a new sentence.

Figure 5.5.1 Sentence forming page

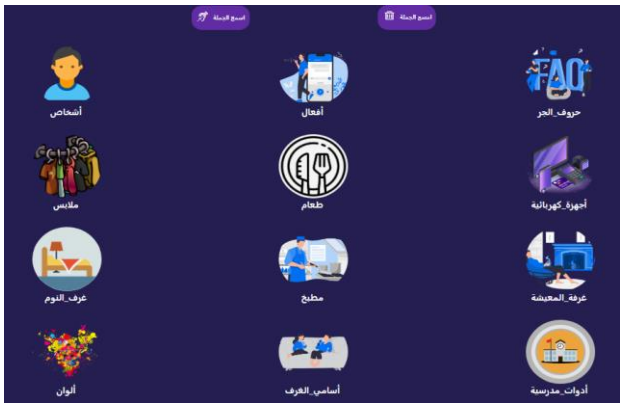


Figure 5.5.2 Category select page

5.6. speech therapist’s view:

He will do the same login steps as the patient had done but with one difference: the role is doctor. And he can access all patient's personal profiles, and edit on them. Then, he can view his patient progress and determine what the next step will be taken for him. The use of the application will allow the therapist to keep track of his patients easily and decrease the number of offline therapy sessions per patient. As a result, the therapist will be available for receiving more patients, increasing his profits and productivity.

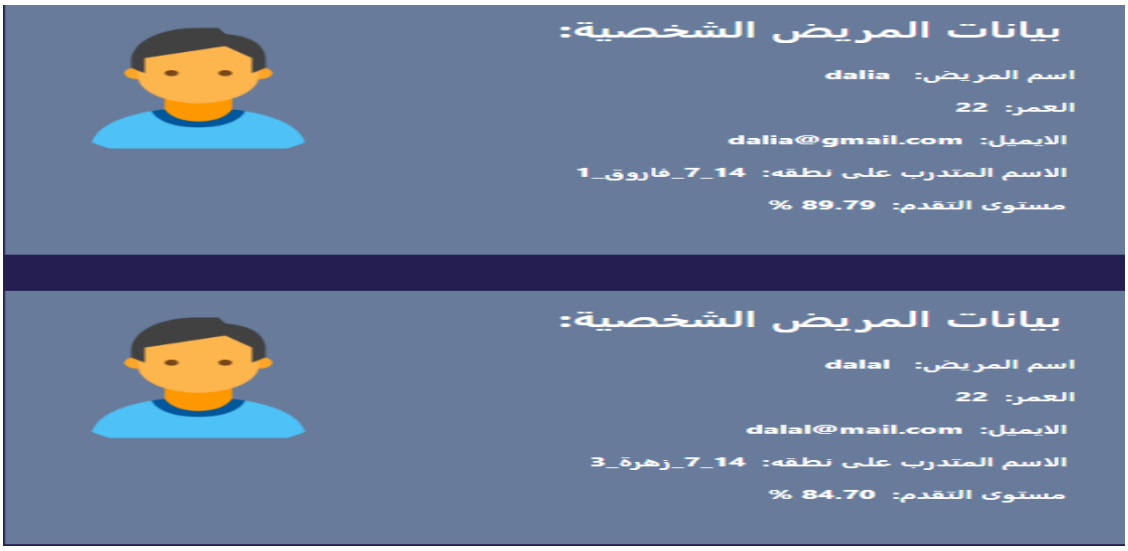


Figure 5.6 The doctor page

Chapter 6 : Discussion

6.1. Previous work

Due to the importance of the subject, intensive studies have been conducted on speech segmentation employing different features. Wang et al. [7], Fu et al. [8] introduced zero crossing rate "ZCR", pitch and energy profile as features for the segmentation of speech. In [9], a survey on Punjabi speech segmentation into syllables is presented using negative derivative of Fourier transformations. In [10], a syllable-based recognition system based on pseudo articulatory method is presented which contributes of more plausible style of speech recognition and new modeling of speech behavior. In [11], a group delay-based approach is proposed which the short-term energy is processed for determining segment boundaries. An attempt is made by Sarada et al. [12] to automate the syllable transcription task

for Indian languages. The method does not require any manual segmentation and a new feature extraction strategy is explored using multiple frame sizes and rates for both training and testing datasets. A technique based on short term energy was implemented in [13] for the automatic segmentation of speech signals in Punjabi speech into syllables. In [14], biological inspired auditory attention cues are proposed for syllables segmentation from continuous speech. The method achieved 92.1 % accuracy of syllable boundary detection at frame level, then using The Mel-frequency Cepstral (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. MFCCs are commonly derived as follows:

1. Take the Fourier transform of (a windowed excerpt of) a signal.
2. Map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows.
3. Take the logs of the powers at each of the Mel frequencies.
4. Take the discrete cosine transform of the list of Mel log powers, as if it were a signal.
5. The MFCCs are the amplitudes of the resulting spectrum.

Then they developed semi-automatic segmentation of speech to Automatic segmentation of speech using Delta-MFCCs based segmentation.
Depending on the variation occurred in the MFCC, we can detect the beginning of each phonetic unit boundary. First, we get the whole utterance from the user and then apply the segmentation algorithm by the following steps:

1. Extract the 12 coefficients of MFC through the whole utterance.
2. Process the MFCCs to get the best coefficients that give a good detection representation.
3. Get the delta function of the best MFC coefficients.
4. Passing the delta function through a law pass filter to eliminate the ripples, although this will smooth the sharp peaks.
5. Scan all local maxima in the delta function, as a represent of the feature variation.
6. Determine the first variable which is the number of the phonetic units present in the input utterance and the order of the target segment as the second variable.
7. Finally, the phonetic unit can be detected by take a size of 300 ms from the letter beginning to be verified. A segmentation success rate of about 91.5% was reached.

6.2. Our work

Our work extended to apply technique called Dynamic Time Warping (DTW) after extracting MFCC. Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences, which may vary in speed. any data that can be turned into a linear sequence can be analyzed with DTW. A well-known application has been automatic speech recognition, to cope with different speaking speeds. Other applications include speaker recognition and online signature recognition. It can also be used in partial shape matching applications. The optimal match is denoted by the match that satisfies all the restrictions and the rules and that has the minimal cost, where the cost is computed as the sum of absolute differences, for each matched pair of indices, between their values. The sequences are "warped" non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension.
This sequence alignment method is often used in time series classification. A segmentation success rate of about 60% was reached.
our drawbacks for our accuracy rate occurred for many reasons:

1. First, we have created our own database so the database was small in relative to other papers' database. This because of non-existence for database available for Arabic language.
2. Second, we didn't have static or balanced reference audio for database as automatic bot. So, our database needed many processing stages and audios were varying in time and speed.

In this paper, we produced comparable results to the existing systems(database) and input audio file. we benchmark our proposed feature-based on Dynamic time warping and Mel-Frequency Cepstral Coefficients (MFCC). MFCC is more preferred in Feature Extraction technique as it generates the training vectors by transforming speech signal into frequency domain, and therefore it is less affected by noise. We extracted MFCC features from the dataset and forwarded these features as a parameter to dynamic time warping for mispronunciation detection. In time series analysis, dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences, which may vary in speed.

MFCC are the most important features, which are required among various kinds of speech applications. It gives high accuracy results for clean speech and can be regarded as the "standard" features in speaker as well as speech recognition.

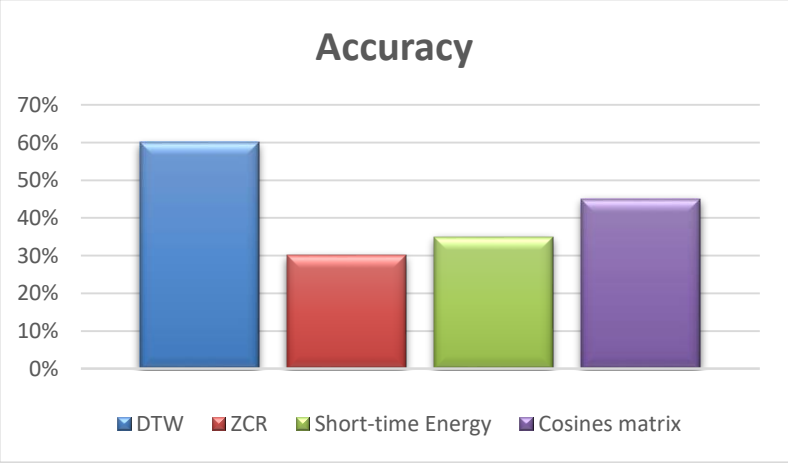


Figure 6.2 shows a chart illustrating the accuracy of different similarity techniques

6.3. Challenges faced us

We faced many challenges in collecting the data like finding relevant data to our project and deciding what data to collect. Collecting data that isn't needed adds time and complexity to the process. But leaving out useful data can limit our dataset's value and affect our results. Other challenges include training people to collect the data and creating sufficient quality assurance procedures to ensure that the data is accurate. Data quality issues as raw data typically includes errors and other issues. As a result, we did some preprocessing on the data to filter them, remove silence and improve its quality.

Chapter 7 : Conclusion and future work

Speech impairments make it hard for people to communicate properly, and they can happen in both children and adults. These disorders can cause frustration and embarrassment to the person suffering from them. It can begin in childhood and carry on through your adult years. Others can happen due to trauma, or after a medical event like a stroke. In our paper, we designed a mobile/web medical application to help people who suffer from speech disorders. There is no such application for Arabic language speakers so our application support mainly the Arabic language. Our application has two modes. The Coach mode depends on comparing an input voice of the user to referenced database that we created and detecting the similarity between the two records. This mode helps patients with articulation disorders who are unable to produce certain word sounds. The Assistant mode depends on forming complete sentences to help patients with cognitive-communication disorders to memorize the words which describe their needs. Signal processing techniques are applied to the speech signal in order to extract the features that distinguish different phonemes from each other. We applied different methods to measure the similarity between recorded files and compared their results to detect the most accurate technique. The dynamic time warping (DTW) technique was the most suitable one for our application which had shown an acceptable accuracy.

Many different adaptations, tests, and experiments have been left for the future due to lack of time (i.e., the experiments with real data are usually very time-consuming, requiring even days to finish a single run). Future work relates to a deeper analysis of particular mechanisms, new proposals to try different methods, or just curiosity.

We can perform an AI system for speech recognition by combining large sets of data with intelligent, iterative processing algorithms to learn from patterns and features in the data that they analyze. Training an algorithm or machine learning model to predict the outcome you design your model to predict.

The AI hardware includes expensive CPUs to handle scalable workloads, special purpose built-in silicon for neural networks, neuromorphic chips, etc.

Advanced deep learning technique such as Transformer model, which allows parallelization and also has its own internal attention, has been widely used in the field of speech recognition. The great advantage of this architecture is the fast-learning speed, and the lack of sequential operation, as with recurrent neural networks. Some studies have shown that the Transformer model improves system performance for low-resource languages. Based on previous experiments, it was revealed that the joint use of Transformer and connectionist temporal classification models contributed to improving the performance of the Kazakh speech recognition system and with an integrated language model it showed the best character error rate of 3.7% on a clean dataset.

We can search for other techniques to improve similarity not just AI systems as future work.

Adding feature that user can make his own database for familiar words he used daily such as family member names by adding their names and the application will convert this name text to record and added to (personal database). So, each user can have his own database to be used in his profile.

Also, therapist can sit a program by choosing specific words for his patient to practice daily according to patient’s progress.

References

1. Abdo, Mohamed & Kandil, Ahmed & Fawzy, Sahar. (2014). MFC peak based segmentation for continuous Arabic audio signal. 224-227. 10.1109/MECBME.2014.6783245.
2. Akhtar, Shamila, Fawad Hussain, Fawad R. Raja, Muhammad Ehatisham-ul-haq, Naveed K. Baloch, Farruh Ishmanov, and Yousaf B. Zikria. 2020. "Improving Mispronunciation Detection of Arabic Words for Non-Native Learners Using Deep Convolutional Neural Network Features" *Electronics* 9, no. 6: 963. https://doi.org/10.3390/electronics9060963
3. Müller, Meinard. (2007). Dynamic time warping. Information Retrieval for Music and Motion. 2. 69-84. 10.1007/978-3-540-74048-3_4.
4. Maqsood, Muazzam & Habib, adnan & Nawaz, Tabassam. (2019). An Efficient Mispronunciation Detection System Using Discriminative Acoustic Phonetic Features for Arabic Consonants. International Arab Journal of Information Technology. 16. 242-250.
5. https://www.izotope.com/en/learn/digital-audio-basics-sample-rate-and-bit-depth.html, by Griffin Brown, iZotope Content Team May 10, 2021
6. Singh, Jyotika. 2019. "An introduction to audio processing and machine learning using Python." Opensource.com, Red Hat, Inc., September 19. Accessed 2021-05-23.
7. Doshi, Ketan. 2021. "Audio Deep Learning Made Simple (Part 1): State-of-the-Art Techniques." Towards Data Science, on Medium, February 12. Accessed 2021-05-23.
8. Velardo, Valerio. 2020a. "Audio Signal Processing for Machine Learning." Playlist on Youtube, The Sound of AI, October 19. Accessed 2021-05-23.
9. Knees, Peter, and Markus Schedl. 2016. "Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies." The Information Retrieval Series, vol. 36., Springer-Verlag Berlin Heidelberg. doi: 10.1007/978-3-662-49722-7. Accessed 2021-05-23.
10. Knees, Peter, and Markus Schedl. 2013. "Music Similarity and Retrieval." Tutorial, SIGIR, July 28. Accessed 2021-05-23.
11. Schutz, Michael, and Jonathan M. Vaisberg. 2012. "Surveying the temporal structure of sounds used in Music Perception." Music Perception: An Interdisciplinary Journal, vol. 31, no. 3, pp. 288-296. doi: 10.1525/mp.2014.31.3.288. Accessed 2021-05-23.
12. Chen, C. H., Signal processing handbook, Dekker, New York, 1988
13. ^ Gouyon F., Pachet F., Delerue O. (2000),On the Use of Zero-crossing Rate for an Application of Classification of Percussive Sounds, in Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00 - DAFX-06), Verona, Italy, December 7–9, 2000. Accessed 26 April 2011.

14. A Large Set of Audio Features for Sound Description - technical report published by IRCAM in 2003. Section 6.1.1 describes the spectral centroid.
15. Zahid S., Hussain F., Rashid M., Yousaf M., and Habib H., “Optimized Audio Classification And Segmentation Algorithm by Using Ensemble Methods,” Mathematical Problems in Engineering, vol. 2015, pp. 1-11, 2015.