République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Chadli Bendjedid
Faculté des Sciences et de la Technologie
Département d'Informatique

الجمهورية الجزائرية الديمقراطية الشعبية
وزارة التعليم العالي والبحث العلمي
جامعة الشاذلي بن جديد
كلية العلوم و التكنولوجيا
قسم الاعلام الالي

MEMOIRE

Présenté par

ABBACI Bochra

Pour l'obtention de diplôme de

MASTER

Filière : Informatique

Spécialité : Systèmes Informatiques Intelligents

Thème

Création d'une base de données parole et l'extraction de ses paramètres vocaux

Building a speech database and then extracting its vocal parameters

Soutenu le :30 /06 /2022

Devant le Jury composé de :

Qualité	Nom et Prénom	Grade	Université
Président	Mr. Benmachiche A.	MCA	Chadli Bendjedid El-Tarf
Rapporteur	**Mme. Makhlouf A.**	**MCB**	**Chadli Bendjedid El-Tarf**
Examinateur	Mr. Chemam C.	MAA	Chadli Bendjedid El-Tarf

Année Universitaire : 2021/2022

Acknowledgment

- First, I would like to thank my supervisor: Dr. Makhlouf Amina, for all her advice, support, help, and guidance.
- I would like to thank the jury members Dr. Chemam C, and Dr. Benmachiche A for accepting to judge my work.
- I would like to thank all the teachers of the 2021-2022 Master's degree in computer science (Intelligent system).
- I would like to thank my little family because they have been by my side throughout my academic journey and without them, I would not have reached here.
- I would like to thank all the people who supported me directly or indirectly in the realization of this work.

Dedication

I dedicate this modest work

To the two dearest beings in the world, the candles that have always guided me on the right path:

My mother,

the light of our life did everything for my success and my happiness

My father,

I must offer all respect and love for his support and tenderness.

To my brother Mounir, and my sisters Oummaima and Dina.

To my Uncle Muhammad and his wife.

To all relatives and friends who in one way or another shared their support either morally, financially, or physically, thank you.

Table of contents

List of figures

List of tables

List of acronyms

DCT	Discrete Cosine Transform
HMM	Hidden Markov Models
LPC	Linear Predictive Coefficients
LPCC	Linear Predictive Cepstral Coefficients
MFCC	Mel Frequency Cepstral Coefficients
MLP	Multi-Layer Perceptron
MSA	Modern Standard Arabic
PLP	Perceptual Linear Predictive
ASR	Automatic Speech Recognition.
RASTA	Relative Spectral
RASTA-PLP	Relative Spectral-Perceptual Linear Predictive
J-Rasta-PLP	Jah-*Relative Spectral -Perceptual Linear Prediction*
CAASA	in-Car Arabic Acoustic Signal Analysis
DFT	Discrete Fourier Transform
Log	Logarithms

General Introduction

Speech is an important way of communication for human, and its simplicity makes it the most popular means of communication in human society, for that, the researchers are always trying to make it possible between the human and machines. In 1922 the first automatic speech recognition (ASR) system was introduced which recognizes just few words is Radio REX. Nowadays we use ASR in a wide range of applications including health care, home automation, telephone and in other areas.

The acoustic signal of speech has characteristics that complicate its interpretation and increase the amount of data to be processed. We need extract the acoustic parameters relevant to the automatic speech recognition by the Feature extractions that play an essential step in the ASR process. The first step of ASR is pre-processing a signal, after that we find the feature extraction phase that is preserved a set of predefined features of the speech processor, using extraction techniques. The problem of automatic speech recognition (ASR) consists in extracting, using a computer, the lexical information contained in a speech signal (and not semantic as in the case of speech understanding).

Automatic speech recognition is not an easy problem, mainly due to the specificities of the speech signal. The acoustic signal of speech has characteristics that complicate its interpretation and increase the amount of data to be processed.

The Arabic language is one of the most morphologically complex languages. For this reason, the extraction of the relevant acoustic parameters is a very difficult task; it is characterized by the high number of dialects used in daily communications.

Arabic is one of the most widely spoken languages around the world with an estimated number of over 313 million speakers with 270 million as a second language speaker of Arabic ranked as the fourth after Mandarin, Spanish and English **[1]**. Moreover, it is the language of the Islamic holy book “Quran” with 1.8 billion Muslims around the world in 2015 and projected to increase to 3 billion in 2060**[2]**. Also, all this motivate us to create an Arabic database, in addition to the lack of databases in the Arabic language, due to the difficulty of processing them, as we mentioned previously.

Our goal is to build an Arabic database and extraction of its vocal parameters with three methods (MFCC, Rasta-PLP and J-Rasta-PLP). To verify the performance of our proposed methods we used the obtained results as inputs in a speech recognition system with Multilayer Perceptron-like Artificial Neural Networks (MLP).

In our work, we will study the extraction of acoustic parameters in automatic speech processing; our thesis will include a general introduction, three chapters, and general conclusion.

General concept on the production of speech, and its automatic processing, the description of the different methods of extraction of the acoustic parameters is introduced in the first chapter.

The second chapter will present the design of our system, some results, and discussion of the results.

In the third chapter, we will see the stages of implementation of the three methods of extraction of the acoustic parameters are summarized within the framework of an application under MATLAB. We end with a general conclusion that will close the comparison between the implemented methods.

Chapter 1: State of the art

1. Introduction

Speech is an important way of communication for human, and its simplicity makes it the most popular means of communication in human society (it is easier to talk to someone than to write or make him a diagram). Nevertheless, this simplicity for the human being contains a very complex processing done by our brain, from the production of speech to its perception and understanding, which makes speech difficult to automate for a machine.

In this chapter we will present the characteristics of speech, production of speech and its processing, as well as the auditory apparatus of the human being, the perception of speech, the advantages of spoken communication and the complexity of the speech signal, the different approaches used for speech recognition, some methods used for the extraction of acoustic parameters for ASR, and we will present some existing databases for ASR systems and conclusion.

1.1. Speech

The vocal signal represents the combination of simple and short elements of the sound signal called phonemes, which make it possible to distinguish the different words. The speech signal is an acoustic vector carrying information of great complexity, variability and redundancy, whose speech signals are differentiated by a set of characteristics.

1.1.1 Characteristics of speech

- The characteristics of this signal are called acoustic features. Among these characteristics are: The fundamental frequency, the frequency spectrum, the stamp, the pitch, and Intensity.

The fundamental frequency

It is the first acoustic feature; it is the frequency of vibration of the vocal cords. For voiced sounds. Corresponds to the period of the wave. The frequency of this wave allows us to evaluate, globally, the pitch of the sound. The waves that accompany the fundamental are called the harmonics.

The frequency spectrum

This is the second acoustic feature on which the timbre of the voice mainly depends. It results from dynamic filtering of signals coming from the larynx or glottis signal by the vocal duct.

The stamp

The Stamp is the set of characteristics that make it possible to differentiate a voice. It comes from the resonance in the chest; the throat, the oral cavity, and the nose are the relative amplitudes of the harmonics of the fundamental which determine the timbre of the sound.

The physical elements of the stamp include:

- The relationships between the parts of the spectrum, harmonic or not.
- The noises existing in the sound (which have no frequency, but whose energy is limited to one or more frequency bands).
- The global dynamic evolution of the sound.
- The dynamic evolution of each of the elements in relation to each other

The pitch

The variation of the fundamental frequency defines the pitch which constitutes the perception

pitch (where sounds are ordered from low to high). Only quasi-periodic sounds (voiced) create a sense of pitch.

Intensity

Intensity, also called volume, makes it possible to distinguish a loud sound from a weak one. Intensity is linked to the air pressure upstream of the larynx, which varies the amplitude of the sound vibrations.

1 .1.2. Speech production and perception

The production of speech sounds is generally presented in three stages in phonetic literature.

Lung wind tunnel: the word is grafted on the breathing which considerably modifies the rhythm.

Laryngeal vibrator*:* the larynx constitutes the vocal source. It contains the vocal cords in the vibration that generates the voice and the intonation.

Supra-Glottis resonator: their number, shape, volume, determines the different speech sounds.

-Speech sounds are produced by the phonatory apparatus described in Figure 1, which includes the diaphragm, lungs, trachea, pharynx, larynx, oral and nasal cavities...etc.

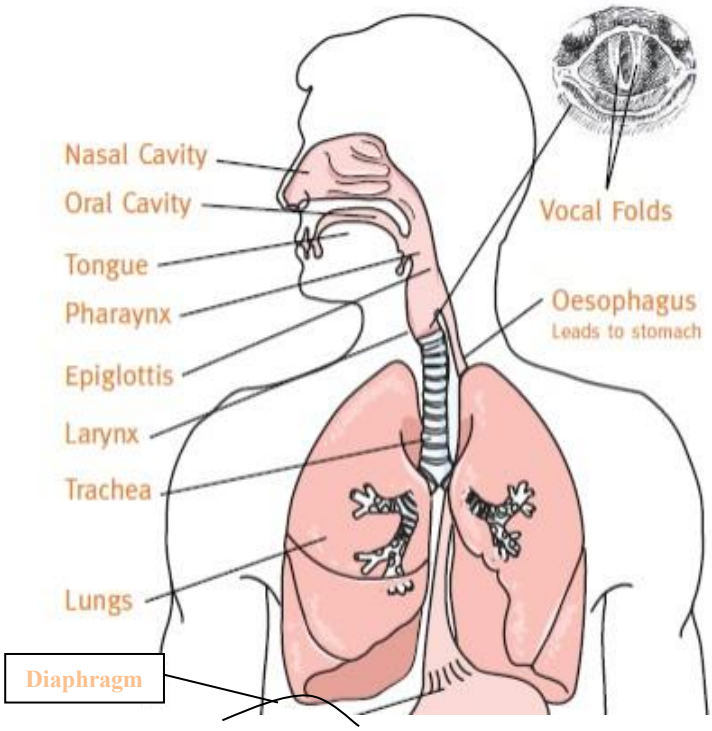

Figure 1. Representation of the phonatory apparatus **[3].**

Voiced speech sounds then arise from different phenomena described in these steps:

1- The brain triggers the five steps below.

2- Diaphragm muscles help inflate and deflate the lungs.

3- The lungs send air to the mouth.

4- The vocal cords vibrate air from the lungs.

5- The nose, mouth, lips, and tongue modify the sound wave to form the words.

6- Finally, a word is pronounced.

Figure 2 illustrates the mechanism of speech production and perception in humans.

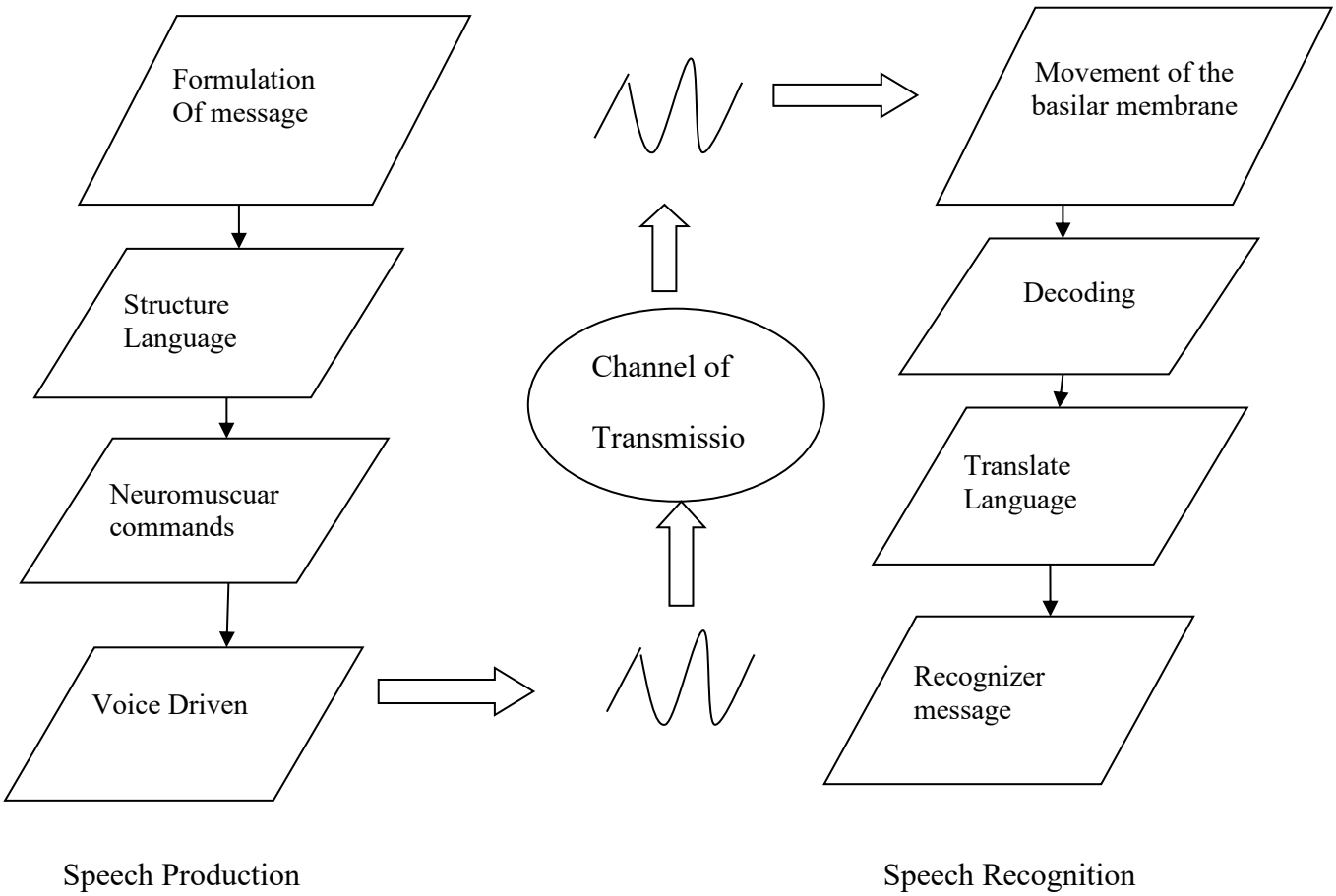

Figure 2. Mechanism of speech production and perception in humans.

Figure 3 describes the main steps of human perception (right part), also illustrates the transcription of these steps in the field of signal processing (left part).

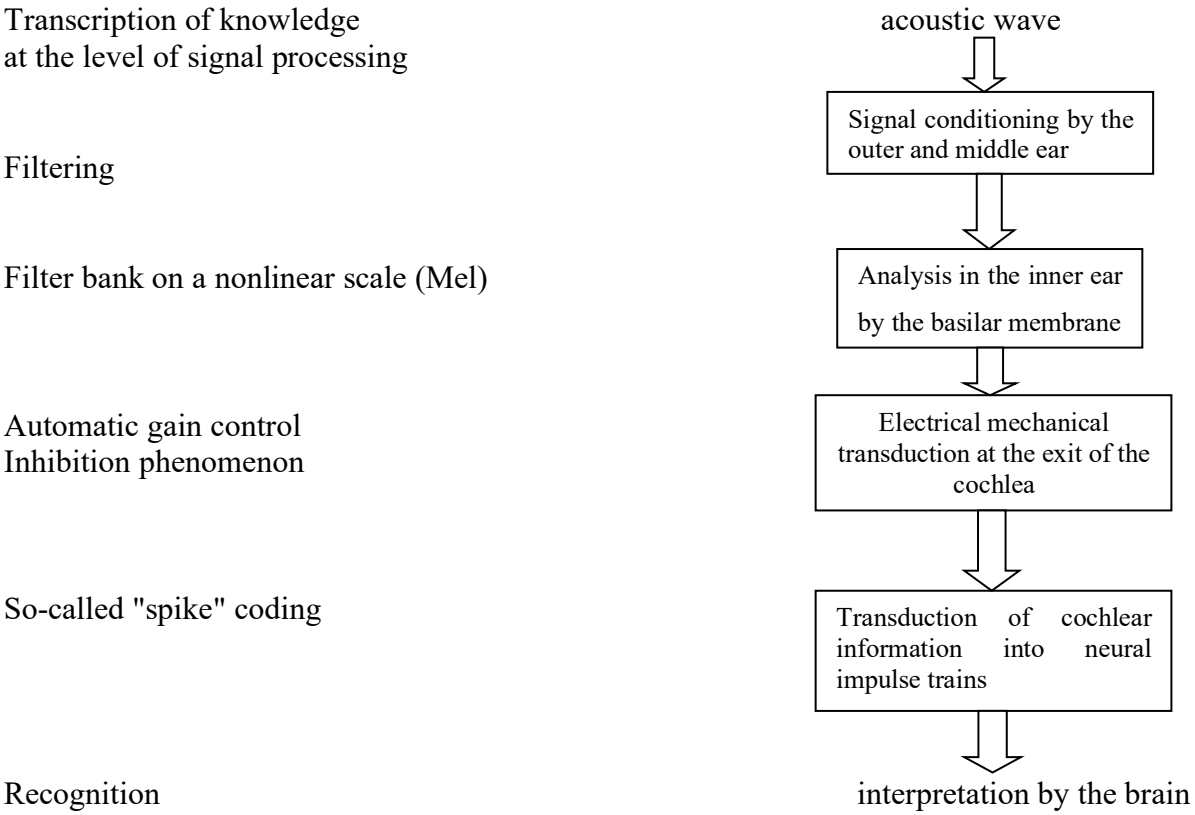

Figure 3. Between human and machine perception.

1.2. The benefits of spoken communication

The advantages of human-machine oral communication are multiple. This mode of the relationship completely frees the use of sight and hands and leaves the user free to move. The transmission speed of information is higher in the man-machine direction than that which allows the use of the keyboard. Moreover, the voice informs about the identity of the speaker and can moreover be transported by existing simple means, such as the telephone network.

Its advantages are reflected in the existence of a wide variety of applications related to automatic speech recognition, as an example, we can cite in Table 1.

Domaine	Description	Examples
Medicine	Help for the disabled (For example, change TV channels without a remote control, answer the phone without hearing any sound)	Voice Access**,** Google Gboar, Voice Assistant of Google
Automotive	Automatic booking (allow drivers to find routes, send emails, make phone calls, and listen to music, all using the sound of their voice.)	Apple CarPlay, Google Android Auto, Manufacturer-Specific Setups, Nuance & BMW
Office automation	office tasks that digital assistants are or will be able to perform: • Launch software • Enabled or disabled WIFI •Do an internet search	Cortana from Microsoft, Siri from Apple.

Table 1. Examples of automatic speech recognition applications.

The development of such applications aims to improve user comfort, increase communication efficiency, and provide new possibilities.

1.3. What makes speech recognition hard?

The speech signal is not an ordinary signal: it is part of spoken communication, a very complex phenomenon. To highlight the difficulties of the problem, we will essentially highlight a few notorious characteristics of this signal:

1.3.1. Noisy data

Speech is spoken in an audio environment such as someone coughing in the background, another speaker in the background, or a second person speaking over the primary speaker, etc. In ASR system this unwanted information in the audio signal is known as noise and we need to identify this noise and filter it out of the speech signal.

1.3.2. Speaker variables

ASR systems often need to include people of different races, parts of the world, and backgrounds. Here are the ways in which speech can differ from person to person: language, dialect, accent, speed, volume, gender, and pitch. An ASR system to work well for all of its users, it is should be able to understand and interpret a large variance in speech.

1.3.3. Homophones and contextual difficulties

In the Arabic language alone, there are many words that are homophones, or words that sound the same but have a different meaning. An ASR system needs a very precise NLP algorithm that powers it to interpret the context of what each speaker is saying.

1.3.4. The continuity

The production of a sound is strongly influenced by the sounds that precede and follow it due to the anticipation of the articulator gesture. Correct identification of a segment of speech isolated from its context is sometimes impossible. Obviously, it is easier to recognize isolated words well separated by periods of silence than to recognize the sequence of words constituting a sentence. Indeed, in the latter case, not only is the boundary between words no longer known but, moreover, the words become strongly articulated.

1.3.5. Bad material

Companies often don't have high-quality hardware to use for capturing audio, resulting in the noisy data.

2. Automatic speech recognition

Indeed humans are now communicating with the machines as much as they communicate with each other or more ,communication with a machine capable of distinguishing a few command words spoken in isolation or even with a computer to which one could dictate a text, or ask it for information, obviously requires the creation of a speech recognition system .The main purpose of an Automatic Speech Recognition (ASR) system is to "simulate" the human listener who can "understand" a spoken language and "respond". This means that the automatic speech recognition must first convert the speech to another medium such as text.

2.1. Types of speech recognition system based on utterances

ASR systems can be classified into several different categories depending on the types of utterances they are able to recognize. These categories are based on the fact that one of the difficulties of ASR is the ability to determine when a speaker begins and ends an utterance.

2.1.1. Isolated words: word recognition systems isolated accept only one word at a time, where they require the speaker to pause between the words. Single-word recognition is suitable for situations where the speaker is required to give the ASR system only one response or isolated words denoting commands.

2.1.2. Connected word: recognition systems connected words make it possible to process words separated by pauses. They are similar to single words but allow separate utterances to run together with minimal pause between them.

2.1.3. Continuous speech: Recognition systems continuous speech allows users to speak almost naturally and deal with a speech where words are connected the instead of being separated by pauses. As a result, word boundaries unknown, co articulation and speaking speed affect their performance. These systems are among the most difficult to create, as they use special methods for determining word boundaries.

2.1.4. Spontaneous speech: spontaneous speech can be considered as natural speech whose content is not known beforehand. An ASR system dealing with spontaneous speech should be able to handle a diversity of natural speech features such as words running at the same time (slight stutters and non-words such as: "um", "ah", etc.)

2.2. Operating modes of a recognition system

A recognition system can be used in several modes:

2.2.1. Speaker dependent (mono-speaker)

In this case, the recognition system is configured for a specific speaker. This is the case with most speech recognition systems available on the market. The main current voice dictation systems have a learning phase recommended before any use in order to adapt the parameters to the user's voice.

2.2.2. Multi-speaker

The recognition system is designed for a limited group of people. The passage from one speaker to another of the same group is done without adaptation.

2.3. Functioning of speech recognition system

The ASR process has two phases as we see in (figure 4):

- **Training**

– Preprocessing + Feature Extraction.

- Model Generation.

– Input: learning speech sounds (70 to 80%) from the sound base.

- **Testing**

– Preprocessing + Feature extraction.

– Classification and take the decision.

– Input: Test speech sounds (30% to 20%) of the sound base.

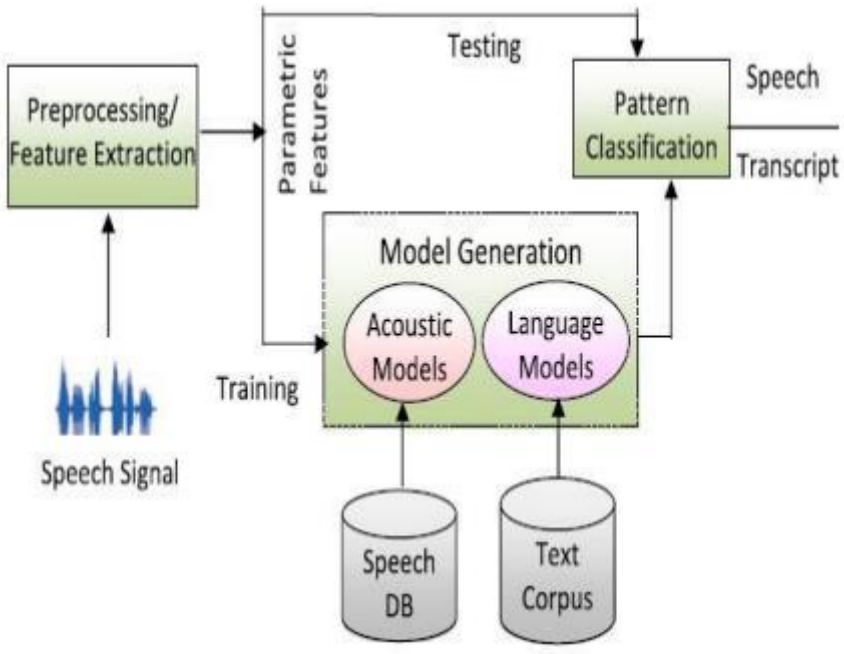

Figure 4. System architecture for automatic speech recognition system **[4].**

2.3.1. Pre-processing/Digital Processing: The recorded acoustic signal is an analogy signal. An analogy signal cannot directly transfer to the ASR systems. So, these speech signals need to transform in the form of digital signals and then only they can be processed. These digital signals

are move to the first order filters to spectrally flatten the signals. This procedure increases the energy of signal at higher frequency. This is the pre-processing step **[4]**.

2.3.2. Filter the data with feature extraction: This phase makes it possible to extract parameters that characterize the information hidden behind this signal, which is also called a feature vector or descriptor that can be used for voice signal processing for recognition. The main purpose of the feature extractor is to keep the relevant elements
information and discard irrelevant ones. To act on this operation, the feature extractor divides the acoustic signal into 10-25ms(segmentation). The data acquired in these frames are multiplied by window a function. There are many types of window functions that can be used as a Rectangular Hamming, Blackman, Welch or Gaussian etc in this way, in this way features have been extracted from every frame **[4]**. For this phase there are approaches and for each approach, there are several techniques (which explained in section 3).

2.3.3. Acoustic modelling: is one of the most important knowledge sources for ASR. It takes the speech waveform and breaks it down into small fragments and predicts the most likely phonemes of speech.
In an empirical approach, the acoustic model is created from large speech DB. The model that is used in most systems is based on Hidden Markov Models (HMMs), which are finite state transducers to determine the probability of a series of observations. Formally, to find the most probable sentence $W*$, the equation (1) that maximizes all sequences of words W knowing the acoustic observation X, is written as follows:

$$\boldsymbol{W = arg_w\ max\ P\left(\frac{W}{X}\right) = arg_w\ max \frac{P(W)/P(\frac{X}{W})}{P(X)}} \qquad \textbf{(1)}$$

2.3.4. Language modelling: form mapping phone to words and phrases is called the language model as huge table that contains probabilities of two words following each other. For example, “hello mine aim is” and “hello my name is” have same pronunciation but mean very different things.

2.3.5. Pattern Classification: Pattern Classification is the process of comparing the unknown test pattern with each sound class reference pattern and computing a measure of similarity between them. After completing training of the system at the time of testing patterns are classified to recognize the speech **[4]**.

3. Speech feature extraction techniques

Feature Extraction is the most important part of speech recognition because it plays an important role to separate one speech from other.

This part presents the main principles that have led to the different speech signal analysis methods currently used for automatic speech recognition. Those methods are cepstrum, linear prediction coding and methods based on hearing models. It should be noted here that before any analysis (extraction of parameters), the speech signal undergoes shaping operations.

Figure 5 illustrates these different operations. The signal is first filtered and then sampled at a given frequency Fe. A pre-emphasis is carried out to raise the high frequencies, and then the signal is fragmented into frames. Each frame consists of a number N of speech samples. In general N is fixed in such a way that each frame corresponds to approximately 30ms of speech. Finally, processing a small piece of signal leads to problems in the filtering (edge effects). To avoid this, windows of weights are used. These are functions that are applied to all the samples taken from the window of the original signal to reduce the edge effects. From most common windows we have the Hamming window. In general, successive windows overlap and they must be of sufficient length.

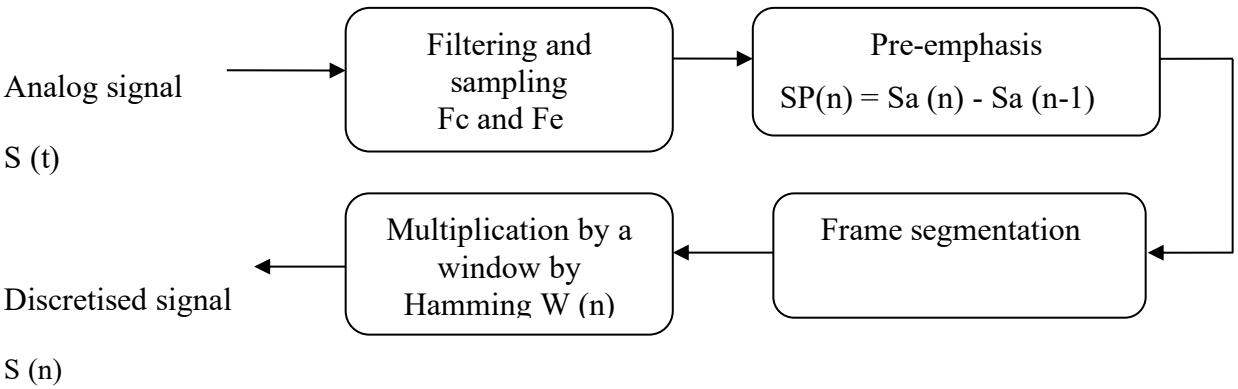

Figure 5. Shaping of the speech signal.

The extracted parameters must be:

- *Relevant:* extracts from sufficiently finished measurements, they must be precise, but their number must remain reasonable so as not to have too high a computational cost in the module of decoding.

- ***Discriminating:*** they must give a characteristic representation of the basic sounds and make them easily separable.

-Robust: they must not be too sensitive to variations in sound level or to background noise.

- Following are some feature extractions (Table 2) and their properties and the procedure of implementation.

Sr. No	**Method**	**Property**	**Procedure of Implementation**
1	Principal Component analyses (PCA)	Nonlinear feature extraction method, Linear map, fast, eigenvector-based	Traditional, eigenvector base method, also known as karhuneu-Loeve expansion; good for Gaussian data
2	Liner Discriminante Analysis (LDA)	Nonlinear feature extraction method, Supervised linear map; fast, eigenvector-based	Better than PCA for classification.
3	Independent Component Analysis (ICA)	Non linear feature extraction method, Linear map, iterative non- Gaussian	Blind course separation, used for de-mixing non- Gaussian distributed sources(features)
4	Linear Predictive coding	Static feature extraction method,10 to 16 lower orders coefficient,	It is used for feature Extraction at lower order
5	Cepstral Analysis	Static feature extraction method, Power spectrum	Used to represent spectral envelope.

6	Mel-frequency scale analysis	Static feature extraction method, Spectral analysis	Spectral analysis is done with a fixed resolution along a Subjective frequency scale i.e., Mel-frequency Scale.
7	Filter Bank analysis	Filters tuned required frequencies	/
8	Mel-frequency cepstrum (MFFCs)	Power spectrum is computed by performing Fourier Analysis	/
9	Kernel based feature extraction method	Non linear transformations	Dimensionality reduction leads to better classification, and it is used to redundant features, and improvement in classification error.
10	Wavelet	Better time resolution than Fourier Transform	It replaces the fixed bandwidth of Fourier transform with one proportional to frequency which allow better time resolution at high frequencies than Fourier Transform
11	Dynamic feature extractions i)LPC ii)MFCCs	Acceleration and delta coefficients i.e., II and III order derivatives of normal LPC and MFCCs coefficients	It is used by dynamic or runtime Feature
12	Spectral subtraction	Robust Feature extraction method	It is used basis on Spectrogram
13	Cepstral meansubtraction	Robust Feature extraction	It is same as MFCC but working on Mean statically parameter

14	RASTA Fellering	For Noisy speech	It is found out Feature in Noisy data
15	Integrated Phoneme subspace method (Compound Method)	A transformation base on PCA+LDA+ICA	Higher Accuracy than the existing Methods

Table2. List of technique with their properties For Feature extraction **[5].**

- ❖ The most widely used feature extraction techniques are explained below with their advantages and disadvantages.

3.1. LPC

LPC coding is one of the most powerful speech analysis techniques that have gained popularity as a formant estimation technique. LPC coefficients are calculated by cutting the speech signal into small windows of short duration. The Hamming window is then applied to the different signal portions obtained. The application of the Hamming window makes it possible to reduce the spectral distortion. With the equation:

$$(n) = \sum akpk{=}1 \times (n - k) + (n) \qquad \textbf{(2)}$$

3.1.1. Advantages:

-Computation speed.

-Robust for extracting features from speech signals with a low bit rate.

3.1.2. Disadvantages:

-Highly correlated feature coefficients.

- Unable to distinguish words with similar phonemes.

3.2. LPCC

The Linear Prediction Cepstral Coefficients (LPCC) technique is mainly derived from linear predictive analysis, where the LPCCs parameters (the first p cepstral coefficients Cn).

LPCC was created to address the limitations of LPC by delivering less correlated coefficients instead of the strongly correlated ones provided by LPC. It should be noted that all the calculation

steps of the LPCs are maintained in the calculation of LPCCs. The LPCC characteristics are calculated by introducing the cepstral coefficients in LPC parameters.

3.2.1. Advantages:

-Decor relates feature coefficients by the cepstral analysis.

- Robust than LPC.

3.2.2. Disadvantages:

- Unable to analyse local events accurately.

3.3. MFCC

MFCC is the most obvious example of a feature the group that is widely used in speech recognition. Such as Frequency bands are set logarithmically in MFCC Approaches the response of the human system closer than any other system. MFCC calculation technology is based on in the short term, and therefore from each MFCC framework vector is calculated. To extract transactions Speech sample is taken as input and methods window It is applied to reduce interruptions in the signal. Then DFT It will be used to create a Mel filter bank. According to mil Frequency torsion, width of the triangular filters and are different so the total logarithmic energy is in a critical range around the center Frequency is included. After distorting numbers Transactions are obtained. Finally, a discrete Fourier inverse the converter is used to calculate the cepstral coefficients. Converts quefrency domain transaction log to the frequency domain where N is the length of the DFT. MFCC can be calculated using the formula. All that shown in figure 6.

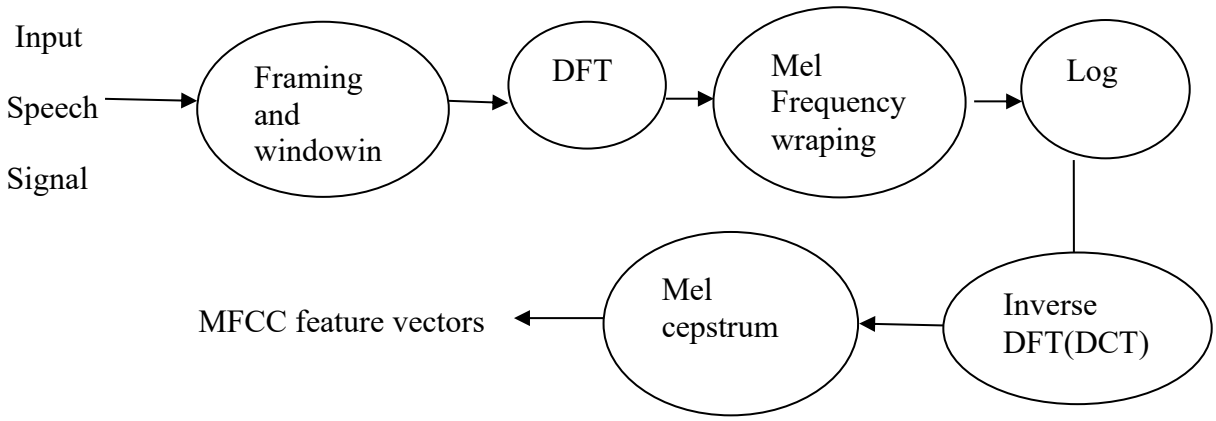

Figure 6. Steps involved in MFCC feature extraction.

3.3.1. Advantages:

- High recognition accuracy.

-Good discrimination and low coefficients correlation.

3.3.2. Disadvantages:

- Inaccurate recognition in noisy speech.

- High dimensional features vectors.

3.4. PLP

Another analysis, called perceptual linear prediction (PLP), uses the same basic principle as the LPC technique since it uses the short-term spectrum of the signal. In addition, PLP uses the knowledge from the psychoacoustics of the human auditory system to optimize the use of the spectrum. This aspect made this analysis closer to human hearing which allowed it to provide more robust parameters. This technique represents an alternative for the cepstral coefficient technique (MFCC).

3.4.1. Advantages:

- Low-dimensional feature vector.
- Reduce the gap between voiced and unvoiced speech.

3.4.2. Disadvantages:

- Altered Spectral balance.

3.5. Rasta-PLP and J-Rasta-PLP

RASTA is derived from PLP analysis and proposed in. The basic design is to suppress variations that are too slow or too fast by filtering on the amplitude spectrum, with the aim of not retaining as the variations related to the signal produced by the human being. This technique deals with non-linguistic speech components, which vary slowly over time, due to convolution noise (log-RASTA) and additive noise (JRASTA). The articulators can't move too fast, so if the characteristics change too quickly, they may not have come from the spoken word. Moreover, if the characteristics change too slowly, this change will not be perceived.

RASTA analysis is often combined with PLP analysis, giving RASTA-PLP, to increase the robustness of the parameters used by ASR systems.

The scheme in figure 7 bellow shows the most steps used in the RASTA-PLP method based on the compute of the critical band power spectrum, as in the PLP method.

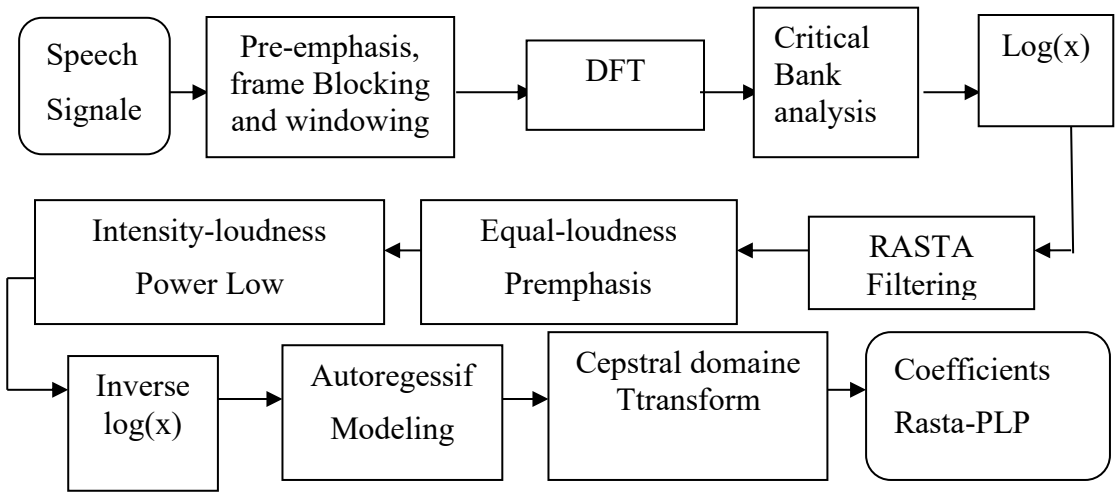

Figure 7. Block diagram of Rasta-PLP method.

To further improve the PLP method, defines the J-RASTA method, plus resistant to additive noise than the RASTA method, by adding low-pass filtering in the spectral domain., the difference between RASTA and J-RASTA is at the logarithm level (5th step): log(x) for RASTA and log (1 + Jx) for J-RASTA. J-RASTA processing addresses the convolutional noise and additive noise.

3.5.1. Advantages:

- Robustness.

-Excludes variations between cepstral components and speech signal.

3.5.2. Disadvantages:

- Low performance for noiseless speech.

❖ Figure 8 illustrates Hybrid feature extraction analysis block diagram.

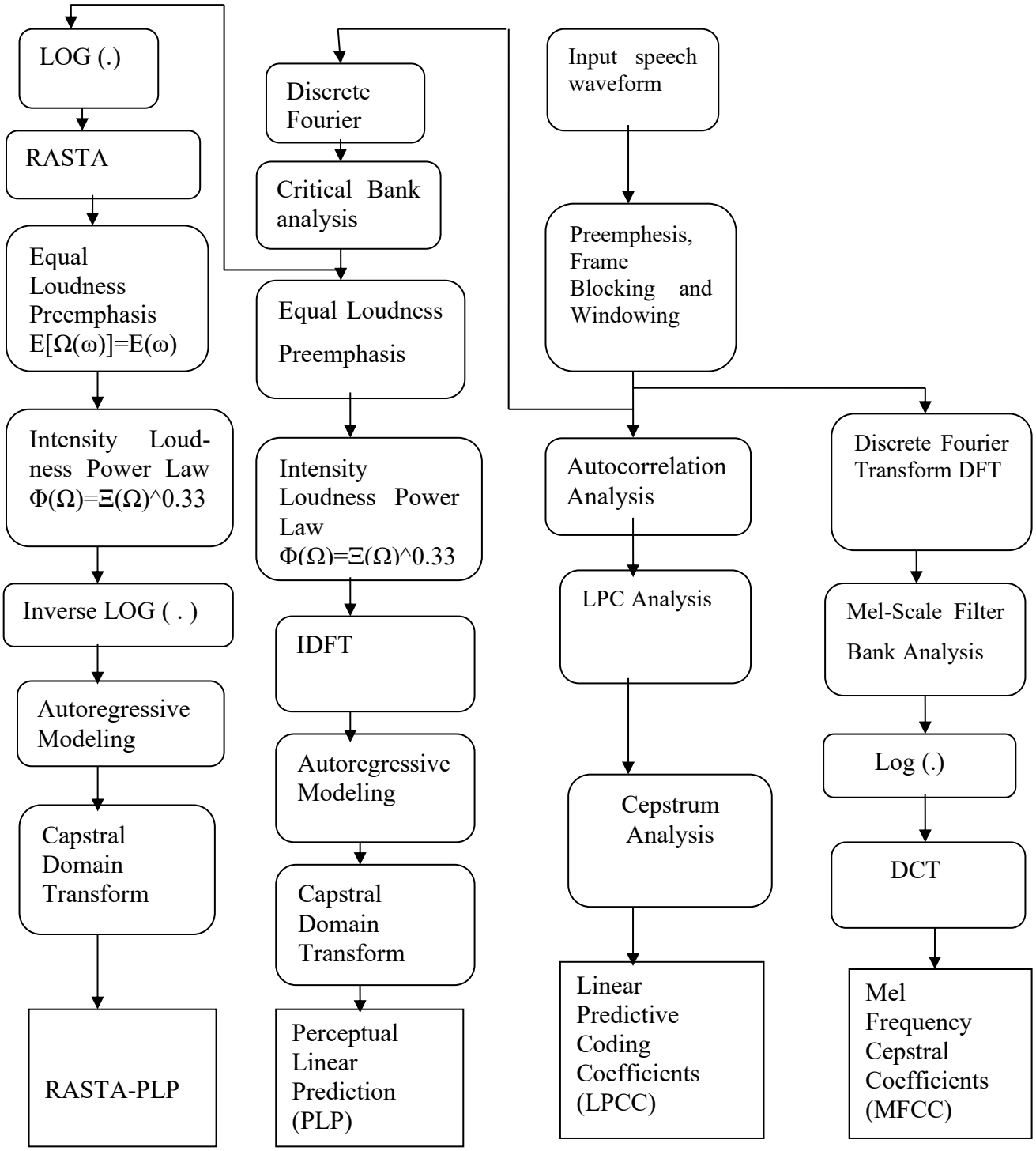

Figure 8. Feature extraction analysis block diagram.

4. Existing speech database

Table 3 is a summary of our survey of other speech databases in different languages and different dialect different environments and different types of instructions.

Corpus	Language /Dialect	Number of speakers	Environment	The type of instruction
Egyptian Arabic Speecon **[6]**	Egyptian	550 adults and 50 Childe	Office, Entertainment , car, public Location	Spontaneous + Read (words, Sentences)
ALGASD **[7]**	Algerian Arabic	300	_	Sentences
SAAVB **[8]**	Saudi	1033	Interior, Exterior, car	Numbers, words, Sentences, alphabets
BREF-80 **[9]**	French	80	_	The texts come from the newspaper LE MONDE and allow to reach a vocabulary of more than 20000 words
TIMIT **[10]**	eight different dialects of American English	630	_	450 phonetically compact sentences (SX) and 1890 phonetically diverse sentences
OrienTel Morocco MCA **[11]**	Moroccan	772	Fixed & mobile phones	Digits, words, sentence + spontaneous

FABIOLE**[12]**	French	130 men	_	French radio or TV shows
OrienTel UAE MCA **[13]**	UAE	880	Fixed & mobile phones	Digits, words, sentence + spontaneous
OrienTel Tunisia MCA**[14]**	Castilian Spanish	104 men Between 28 and 42 years old	_	microphone and telephone channels; read and spontaneous speech
QSDAS **[15]**	Quran recitation	77	Quranic verses	Microphone 1-channel

Table 3. Comparison of speech databases.

5 Conclusion

Automatic speech recognition is not an easy problem, mainly due to the specificities of the speech signal.

The acoustic signal of speech has characteristics that complicate its interpretation and increase the amount of data to be processed. There are several methods to extract the acoustic parameters relevant to the automatic speech recognition step, such methods as MFCC, RASTA-PLP and J-RASTA-PLP.

In this chapter we have presented we explained everything related to the speech, the ASR system, and the different feature extraction methods most used in it with their advantages and disadvantages, and the databases for existing speech synthesis.

Chapter 2: Conception

1. Introduction

The speech appears as a variation of the air pressure in the human phonatory apparatus different acoustic traits of the speech signal are in particular: its fundamental frequency, its energy, and its spectrum... Each of these elements being itself even initially linked to a perceptible magnitude: pitch, intensity, timbre...

The objective of a parameterization system is to extract the characteristic information from the speech signal by eliminating redundant parts as much as possible. Such a system thus takes a signal as input and returns a parameter vector (called observation vector). The parameter vectors must be relevant (precise, of restricted size, and without redundancy), discriminating, and robust to different noises and/or speakers.

While significant research activities have been devoted to English ASR, less attention has been paid to Arabic ASR. Arabic is the official language in 21 countries known as the Arab countries and it is the 6 the most widely used language based by several first language speakers **[16]**. Modern Standard Arabic (MSA) is the current standard form of Arabic, and it is used in writing, news broadcast, formal speeches, and movie subtitling **[17]**. Arabic as a language poses some inherent challenges in conducting ASR research; furthermore, such a study can also bump into other logistical challenges, for example, accessibility to resources and benchmarks and morphological complexity. The lack of a unified large Arabic language isolated speech corpus is a major obstacle that might limit research in this promising field **[18].**

This chapter proposes the architecture of an in-Car Arabic Acoustic Signal Analysis (CAASA) for the selection of acoustic parameters and presents three methods MFCC, Rasta-PLP, and J-Rasta to extract the relevant acoustic parameters of the speech signal, then their applications to the Arabic in-car commands database and descript it, the we explain the three methods of analyses and give an application example, then we present the performance discussion of the three methods for a speech recognition system, finally a conclusion.

2. System architecture

This paper proposed in-Car Arabic Acoustic Signal Analysis (CAASA) system to extract the characteristic information of the speech signal by eliminating as much as possible the redundant parts, our system takes a WAVE signal as input and returns a vector of parameters. The methods

employed based are on Relative Spectral Transform-Perceptual Linear Prediction (RASTA-PLP), Mel-Frequency Cepstral Coefficients (MFCC) and J-Rasta-PLP feature extraction techniques in MATLAB.

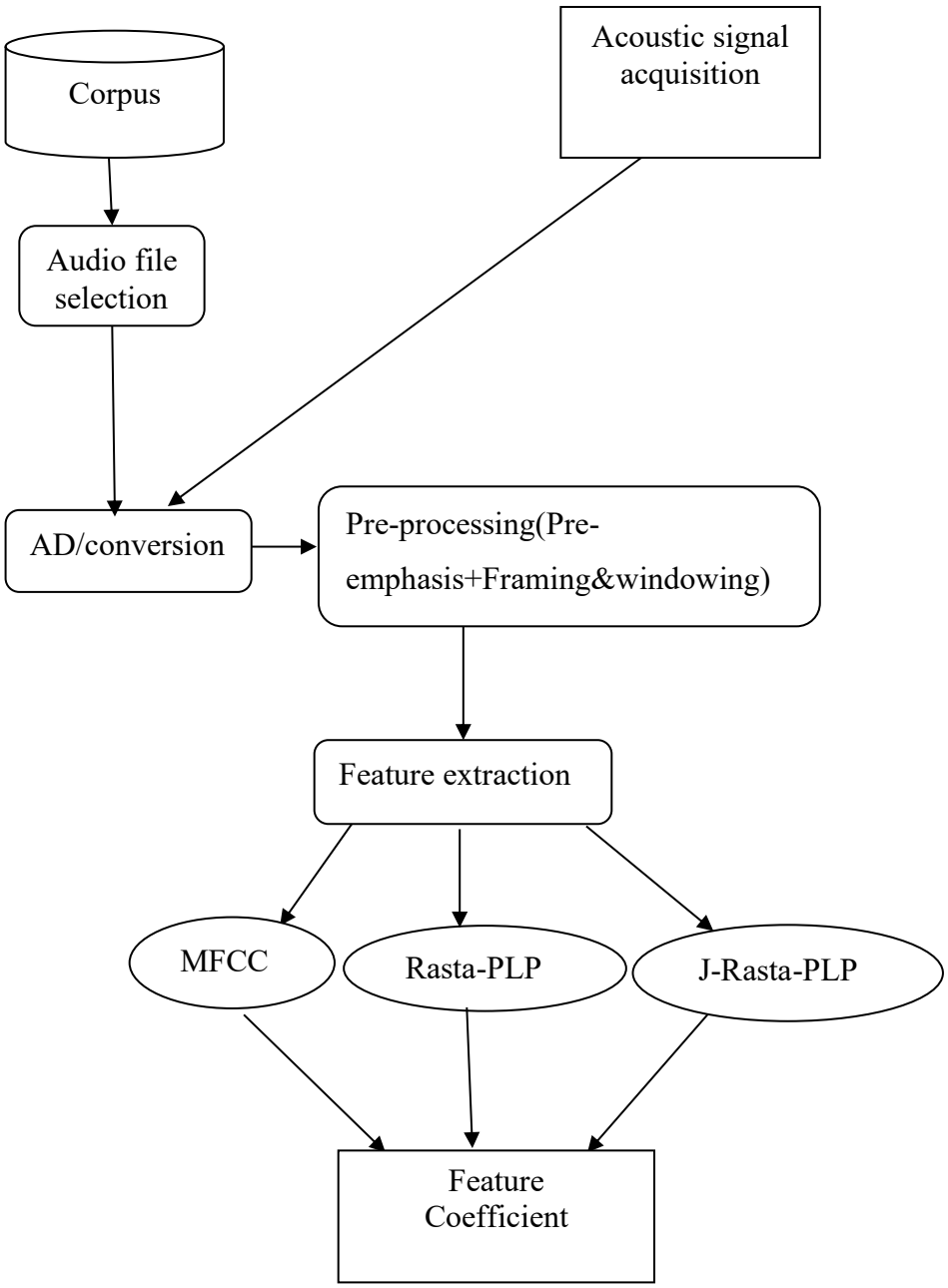

Figure 9. Architecture of CAASA system.

2.1. Acquisition phase

The acquisition was made through a microphone with a wire connected to a Smartphone equipped with software that allows us to record the. WAVE extraction speech signal.

We recorded sounds containing isolated words in the Arabic language.

The distance between the microphone and the mouth is adjustable. The words captured by the distant microphone under two conditions: idling and driving on a city street with a normal in-car environment. In our basic corpus which contains only isolated words, the size of each recording is between 1.5sec and 2 seconds which is a sufficient time to pronounce a word slowly in Arabic.

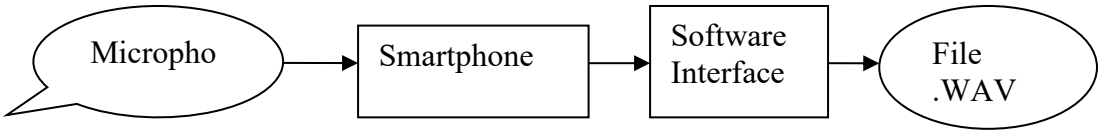

Figure 10. Speech signal acquisition module.

2.2. AD/conversion: we will convert our audio signal from analog to digital format with a sampling frequency of 16 kHz.

2.3. Pre-processing: After the acquisition of the corpus data, we go to the pre-processing step to recover only the speech signal, for this there are common steps include pre-emphasis, frame blocking and windowing to prepare an input speech signal to extract the features.

2.3.1. Pre-emphasis: The goal of this step reduces the dynamic range of the speech spectrum, enabling to estimate the parameters more accurately.

2.3.2. Frame blocking and windowing: To ensure the smoothing transition of estimated parameters from frame to frame, pre-emphasized signal is blocked into samples with 25 ms frame long and 10 ms frame shift, after that we applied hamming windowing technique w (n) as shown in Equation (3) to reduce the effect of discontinuity at the beginning and the end of each frame.

$$W(n) = 0.54 - 0.46 \cos(2\pi n / (N-1)) \quad 0 \leq n \leq N-1. \qquad (3)$$

3. Corpus description

3.1. Dataset

In-car speech recognition systems aim to eliminate the distraction of looking at your mobile phone while driving. Alternatively, a head-up display allows drivers to keep their eyes on the road and

their minds on safety. But while safe driving behaviours (and in many places, the law) require us to ignore constant phone calls, emails, and texts while driving, this kind of disconnect isn't a reality.

For this reason, we made a corpus of fifteen Modern standard Arabic (MSA) words we give it the name “corpusData”, each of which is pronounced by 12 speakers with different ages varying between 18 and 60(The legal driving age starts from 18 years), are invited to pronounce the ten words 50 times. Thus, the corpus consists of 9000 tokens (15 words. 50 repetitions. 12speakers).

The speech corpus is recorded under real conditions when the speaker is driving a car, which noise is existed in the street with a different degree.

The choice of the words of this corpus is not arbitrary; indeed, we took as application, a subset of command the most used that we need it when we use the phone while we are driving our cars (In-car speech recognition system). They are represented in order in the following table 4.

class label	Arabic words	English translation
1	طقس	Weather
2	خريطة	Map
3	يوتيوب	YouTube
4	رسالة	Message
5	اكتب	Right
6	اقر	Read
7	هاتف	Phone
8	مكيف	Air conditioner
9	غناء	Singing
10	ابحث	Research
11	افتح	Open
12	اغلق	Close
13	قف	Stop

14	يسار	Left
15	يمين	Right

Table 4. The words used in our corpus.

3.2. Collect data

Our speech database and oriented towards ASR systems, recognition paradigms such as HMM, neural networks of the MLP (Multi-Layer Perceptron) type, and hybrid systems requires optimization on the processed data side, this necessity is our main objective.

We also see the diversity of data for learning to improve the voice signals are collected from 12 speakers, these speakers are from different dialect regions, these speakers are 5 men and 7 women, each speaker pronounces each word 50 times with different modes of pronunciation (normal, slow, and fast).

4. Methods of analysis

in this work step, we try to develop and optimize three acoustic analysis techniques: MFCC, Rasta-PLP, and J-Rasta, the first of which performs Mel-based filtering to eliminate a large part of the noise, and the second concerns a linear band pass filter and thus keep more information relevant and the differences between the speakers (male, female), the coefficient description and well detailed in chapter 1, and the third method J-RASTA processing addresses the convolutional noise and additive noise.

4.1. MFCC method:

The MFCC technique includes pre-processing, framing, and windowing, applying the DFT, taking the log of the magnitude, and then warping the frequencies on a Mel scale, followed by applying the inverse DCT. The detailed description of various steps involved in the MFCC feature extraction is explained below.

- After the AD/conversion, pre-emphasis, and framing & windowing (common steps in all the feature extraction methods) that already explained in system architecture, we going to explain the rest steps of MFCC.

. DFT (Discrete Fourier Transform): We are applying DFT to convert each windowed frame into magnitude spectrum.

$$X(K) = \sum_{n=0}^{N-1} x(n) e^{\frac{-j2\pi nK}{N}} \quad 0 \leq k \leq N - 136 \qquad (4)$$

Where N is the number of points used to compute the DFT.

. *Mel-Filter Bank:* The way our ears will perceive the sound is different from how the machines will perceive the sound. Our ears have higher resolution at a lower frequency than at a higher frequency. So, if we hear sound at 200 Hz and 300 Hz, we can differentiate it easily when compared to the sounds at 1500 Hz and 1600 Hz even though both had a difference of 100 Hz between them. Whereas for the machine the resolution is the same at all the frequencies. It is noticed that modelling the human hearing property at the feature extraction stage will improve the performance of the model **[19]**.

For that we will use the Mel scale to map the actual frequency to the frequency that human beings will perceive. The formula for the mapping is given below.

$$\text{mel(f)} = 1127 \ln\left(1 + \frac{\text{f}}{700}\right) \qquad \text{Where f is the real frequency (Hz)} \quad (5)$$

. *Applying Log:* Humans are less sensitive to change the energy of an audio signal at higher power compared to lower power. Log function also has a similar property, at a low value of input x gradient of log function will be higher but at high value of input gradient value is less **[19]**. For that to mimic the human hearing system we need to apply log to the output of Mel-filter.

. IDFT (DCT):

The inverse transform of the output are doing from the previous step. Like we see in the chapter 1 how the sound produced by human beings and each sound produced will have its unique position of the tongue and other articulators.

The below figure shows the signal sample before and after the idft operation.

The MFCC model takes the first 12 coefficients of the signal after applying the idft operations. Along with the 12 coefficients, it will take the energy of the signal sample as the feature. It will help in identifying the phones. The formula for the energy of the sample is given below in equation (6) **[19]**.

$$Energy = \sum_{t=t1}^{t2} x^2[t] \quad (6)$$

. Dynamic Features:

With these 13 features, the MFCC technique will consider the first order derivative and second order derivatives of the features which constitute another 26 features.

Derivatives are calculated by taking the difference of these coefficients between the samples of the audio signal and it will help in understanding how the transition is occurring **[19]**.

So, the comprehensive MFCC technology will generate 39 features from each audio signal sample that are used as inputs to the speech recognition model.

4.2. Rasta-PLP method:

Special band-pass filter added to each frequency channel of the old PLP algorithm, the implementation of this filtering allows, when it is carried out in the logarithmic spectral domain, to remove the constant spectral components, thus removing the effects of convolution of the communication channel.

The steps of RASTA-PLP are as follows (see **[20]** for a comparison with the conventional PLP method) **[21]**:

For each analytical framework:

1) Calculate the spectrum of the critical band (as in the PLP) and take its logarithm.

2) Estimate the time derivative of the logarithmic spectrum of the critical band using the regression line through five consecutive spectral values.

3) Nonlinear processing (such as applying a threshold or median filtering) can be performed in this domain. Currently we are not doing anything here.

4) Reintegrate the log critical band time derivative using a first order IIR system. The pole the position of this system can be adjusted to define the effective size of the window. Currently we set this value at 0.98, providing an exponential integration window with a 3 dB point after 34 frames.

5) In accordance with conventional PLP, add the equal loudness curve and multiply by 0.33 to Simulate the power law of hearing.
6) Take the inverse logarithm (exponential function) of this relative logarithmic spectrum, which gives a relative auditory spectrum.
7) Calculate an all-pole model of this spectrum, following the classic PLP technique.
It can be shown that if the derivative of step (2) is estimated by a first simple difference, and if full integration in step (4) is performed (pole at z = 1.0), then all intermediate terms cancel and the technique is equivalent to subtracting the log spectrum of the first analysis frame from each new frame. In this case, the RASTA technique is like spectral or blind subtraction deconvolution techniques.

However, in the general case presented here, the whole derivative-reintegration process is equivalent to band-pass filtering each frequency channel through an IIR filter with the transfer function.

$$H(z) = 0.1 \times \frac{2+z^{-1}-z^{-3}-2z^{-4}}{z^{-4}\times(1-0{,}98z^{-1})} \qquad (7)$$

4.3. J-Rasta-PLP method:

The deference between Rasta-PLP and J-Rasta-PLP is in the first one there is Rasta and in the second one there is J-Rasta, and the common think is PLP that we can understand it in **[20]**.
In this section we go to review J-RASTA processing. We assume the observation signal y(t) as follows:

$$y(t) = h(t) * (x(t) + d(t)) \qquad (8)$$

Where x(t) is a pure speech signal, h(t) is convolutional noise and d(t) is additive noise. In the logarithmic magnitude spectral domain, we can linearly separate the convolutional noise as follows:

$$\log Y(\omega) = \log H(\omega) + \log (X(\omega) + D(\omega)) \qquad (9)$$

Since the power spectrum of real-world convolutional noise is like changing the distortion of the communication channel relatively slowly compared to speech, we can remove the first term "log H (ω)" in equation (9) with a high-pass filter. In Log-RASTA processing, the following band-pass filter is applied to the logarithmic size spectrum (that is, log Y(ω)). This band Passing the filter also prevents a rapid and invisible spectral change in the speech signal **[22]**.

$$R(z) = 0.1z^4 \times \frac{2+z^{-1}- z^{-3}-2z^{-4}}{z^{-4}\times(1-0{,}98z^{-1})} \quad (10)$$

Next, we can extract the term log (X (ω) + D (ω))", which was affected by the added noise D (ω).

In the power spectral domain, we can linearly separate speech that only it is affected by convolutional noise.

$$Y(\omega) = H(\omega)\, X(\omega) + H(\omega)\, D(\omega) \quad (11)$$

We can reduce the added noise by using the same Log-RASTA processing scheme,

(For example, H (ω) D (ω)) which changes relatively slowly or quickly, but the output is affected by convolutional noise and additional noise that comes through the filter R (ω).

There are two problems that arise in this case. The first is that we need to reduce both additives and convolutional noise at the same time. The other is the noise that comes through the R (ω) in equation (11) **[22]**.

The J-RASTA treatment is a solution to the previous problem. Balance Effects equation (9) and equation (10) by setting the input spectrum as follows:

$$X'(\omega) = \ln(1+JX(\omega)) \quad (12)$$

J is set to be close to zero if the additive noise is dominant, otherwise J must be greater value for convolutional noise suppression. In other words, convolutional noise is reduced When J is set to a larger value, because in this case X' (ω) Looks like a log **[22]**.

5. Application example

In this part, we will represent the result of the application of the three methods on an Arabic sample (the word طقس).

5.1. MFCC method:

MFCC extracts 13 spectral coefficients; 13 parameters for the elements are created during the windowing and sampling of the signal, the sampling is fixed at 16 kHz, the size of each window is 25 ms and the windowing step is 12 ms.

The resulting vector is a concatenation of 9000 matrices.

Number of cepstral coefficients = 13.

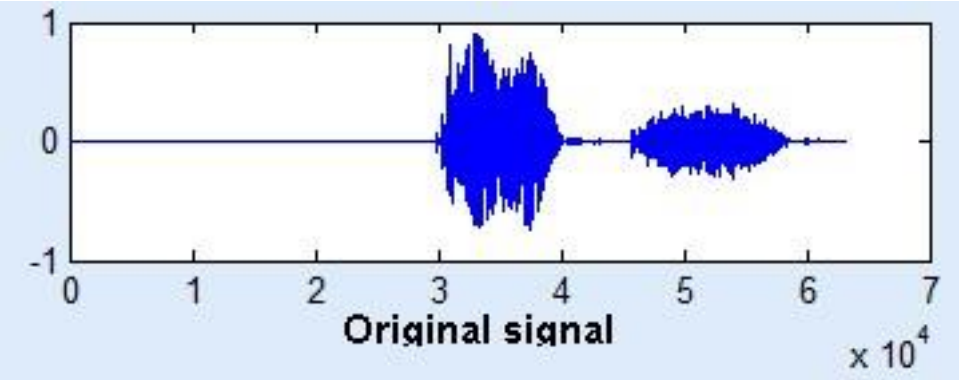

Figure 11. Original signal of the word (طقس)

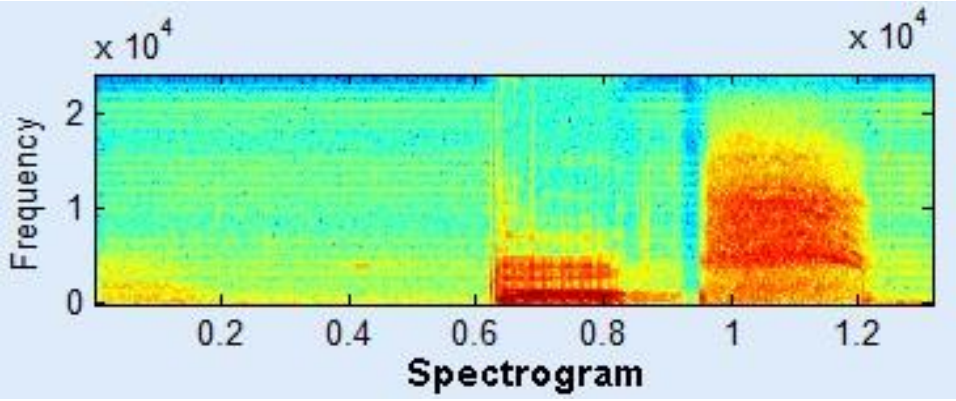

Figure12. Spectrogram for the word (طقس) of MFCC.

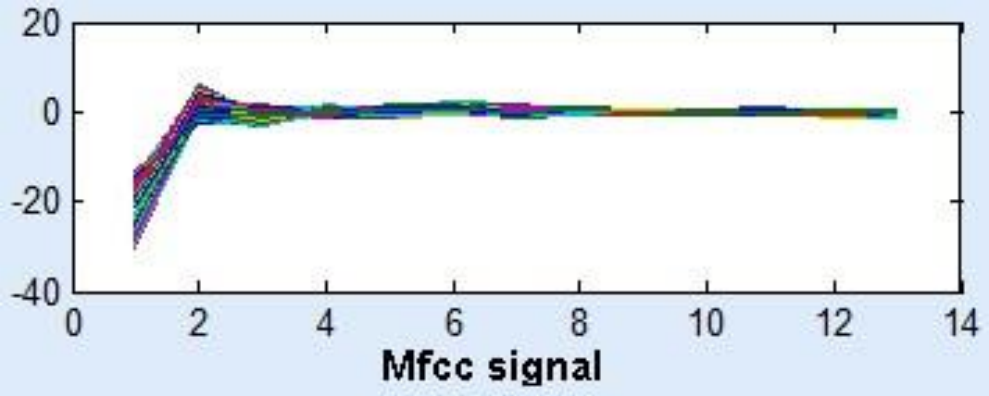

Figure 13. MFCC result for the word (طقس).

	1	2	3	4
1	-28.9286	-27.3783	-25.7139	-25.1802
2	1.2205	0.9002	1.4544	1.2088
3	-0.2270	-0.4209	-0.4669	-0.2032
4	-1.0421	-1.0743	-0.8299	-1.0649
5	0.0640	-0.1131	0.4358	0.1429
6	0.1231	0.3715	0.6270	0.7179
7	-1.1453	-0.6933	-0.6204	-0.4618
8	-0.5609	-0.3559	-0.3118	-0.1727
9	0.0371	-0.1654	-0.2958	0.0654
10	-0.0261	-0.0885	-0.2427	0.3452
11	0.1087	0.4159	0.1955	0.3574
12	0.1708	0.5943	0.1922	0.2940
13	0.5777	0.4752	0.4584	0.1355

Figure 14. Result matrix with MFCC.

5.2. Rasta-PLP method:

By Rasta-PLP, we extract 9 parameters from the acoustic signal of sampling frames at 16kHz, and with a window size of 0.025 seconds and a step of 0.010 seconds.

For “corpusData”, we obtained 9000 matrices concatenated by our corpus.

Number of cepstral coefficients = 9.

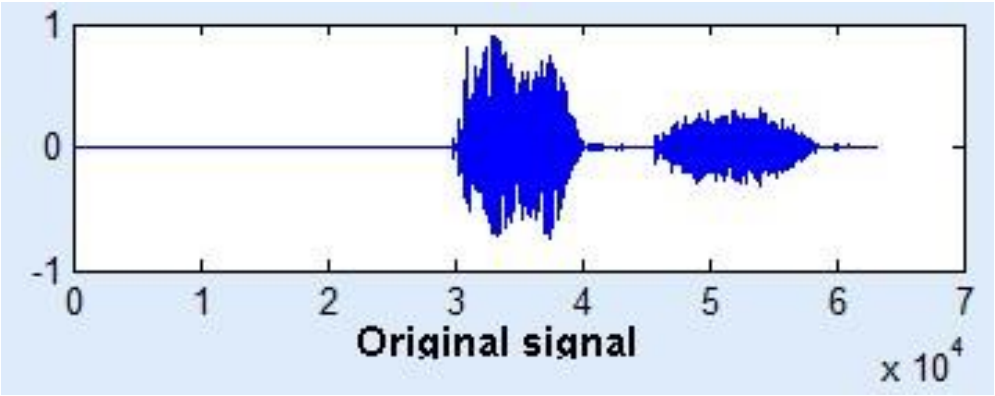

Figure15. Original signal of the word (طقس).

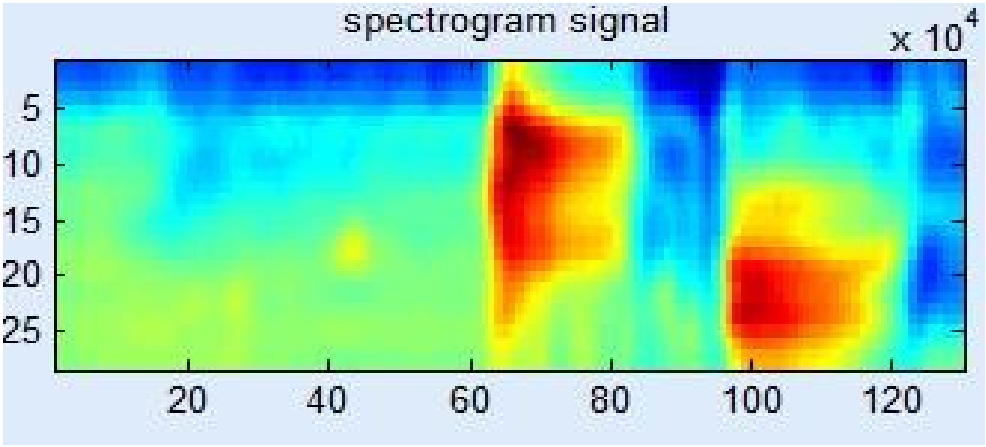

Figure16. Spectrogram for the word (طقس).

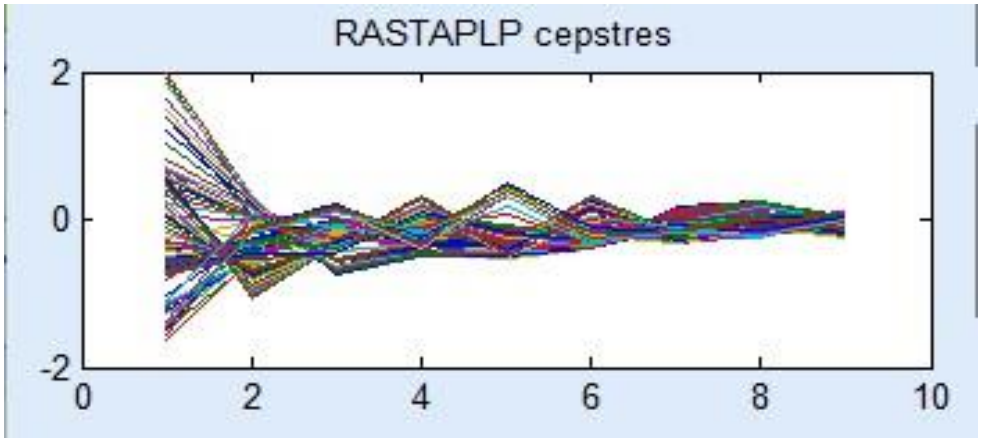

Figure 17. Rasta PLP results for the word (طقس).

	1	2	3	4
1	-0.4163	-0.4163	-0.4163	-0.
2	-0.3263	-0.3263	-0.3263	-0.
3	-0.2731	-0.2731	-0.2731	-0.
4	-0.1979	-0.1979	-0.1979	-0.
5	-0.1723	-0.1723	-0.1723	-0.
6	-0.1555	-0.1555	-0.1555	-0.
7	-0.1267	-0.1267	-0.1267	-0.
8	-0.0870	-0.0870	-0.0870	-0.
9	-0.0385	-0.0385	-0.0385	-0.

Figure 18. Result matrix with Rasta PLP.

- For this method, we observe that the number of parameters obtained is lower than the MFCC with small values between -0.41 and -0.038.

5.3. J-Rasta-PLP method:

By J-Rasta-PLP, we extract 8 parameters from the acoustic signal of sampling frames at 16 kHz, and with a window size of 0.025 seconds and a step of 0.010 seconds.

For “corpusData”, we obtained 9000 matrices concatenated by our corpus.

Number of cepstral coefficients = 8.

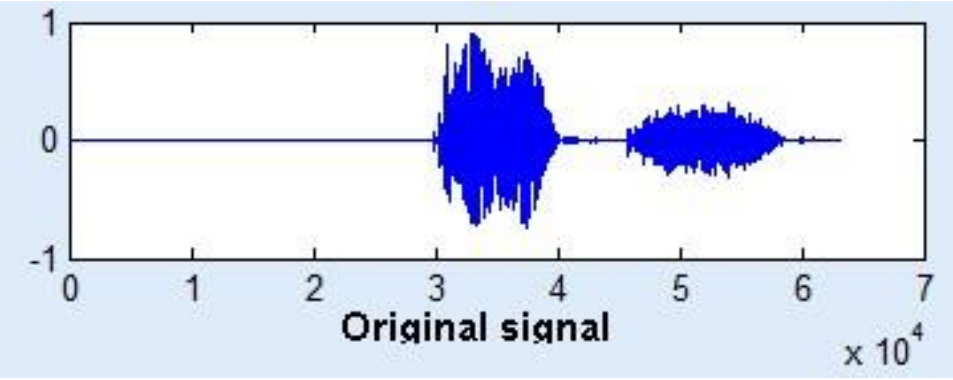

Figure 19. Original signal of the word (طقس).

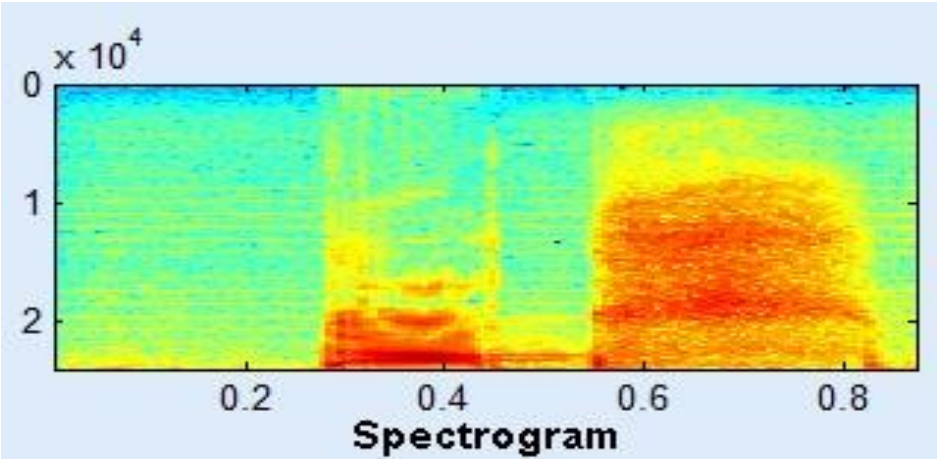

Figure 20. Spectrogram for the word (طقس) of J-Rasta-PLP.

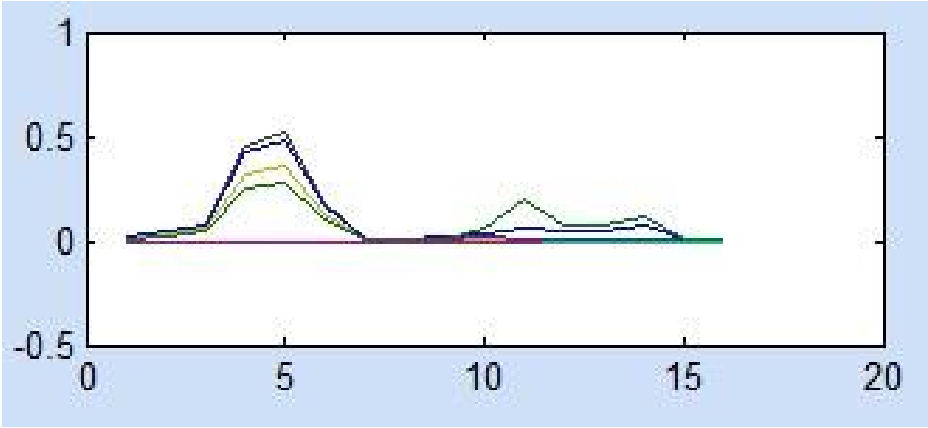

Figure 21. Rasta PLP results for the word (طقس).

	1	2	3
1	-0.4716	-0.4716	-0.4716
2	-0.3411	-0.3411	-0.3411
3	-0.2474	-0.2474	-0.2474
4	-0.1835	-0.1835	-0.1835
5	-0.1660	-0.1660	-0.1660
6	-0.1422	-0.1422	-0.1422
7	-0.1104	-0.1104	-0.1104
8	-0.0804	-0.0804	-0.0804

Figure 22. Result matrix with J-Rasta PLP.

For this method, we observe that the number of parameters obtained is lower than the Rasta-PLP with small values between -0.47 and -0.080.

6. Recognition result

6.1. Recognition method used

To verify the performance of our proposed methods (MFCC, RASTA-PLP and J-RASTA-PLP) we used the obtained results as inputs in a speech recognition system with Multilayer Perceptron-like Artificial Neural Networks (MLP).

We use a multilayer perceptron with only one hidden layer.

• The input layer is made up of X neurons, such that X represents the classes resulting from the clustering step (we used K-means).

• The hidden layer made up of a set of neurons whose number will be deduced by experimentation.

6.2. Results obtained for Arabic database

	MFCC	Rasta-PLP	J-Rasta-PLP
AVERAGE RATE	**14.8%**	**21.6%**	**23.2%**

Table 5. Recognition rate of the Arabic corpus extracted with MFCC, Rasta-PLP and J-Rasta-PLP.

6.3. Discussion

The result of tables 5 represents the speech recognition rate of corpus Data, in these results we noticed that the MFCC method gives 14.8%, Rasta-PLP gives 21.6%, and J-Rasta-PLP gives 23.2%.

Rasta-PLP method gives higher average corpusData rate compared to MFCC method.

J-Rasta-PLP method gives higher average corpusData rate compared to both of MFCC and Rasta-PLP methods.

- From these results we can conclude that the parameters extracted by the Rasta-PLP method give a better performance than the MFCC method and J-Rasta-PLP gives a better performance than MFCC and Rasta-PLP methods for an automatic speech recognition system.

- We observe that the performance of systems is low, this is due to the poor quality of the microphone and the environment (in-car) in which we recorded it contains a lot of noise.

7. Conclusion

In this chapter, we explained the architecture of CAASA system, corpus description, the method of analysis used and the explication of their steps, application example for the word (طقس), recognition result that explained the recognition method used, and Results obtained for Arabic database then we do the discussion of those results.

The results show that J-Rasta-PLP performs better than Rasta-PLP and Rasta-PLP performs better than MFCC in speech recognition rate. Furthermore, the J-Rasta-PLP gives consistent results and is more robust to noise because it is based on the principle of linear prediction, which follows the frequency scale observable in the ear.

In the next chapter we will see what we use to build this system and his final interfaces.

Chapter 3: Implementation

1. Introduction

In this chapter, we will present the implementation of our application using the MATLAB language. We start first with a presentation of development tools that contain the materials used in our work and the chosen programming language (MATLAB) and why we choose it specifically. Then we present screenshots of the execution of our application with an example of a sample Arabic corpus that is already presented in the previous chapter and completed the chapter with a conclusion.

2. Presentation of development tools

2.1. Material:

2.1.1. Computer

The hardware produced is Hp i5 personal PC with 8GB memory capacity, Intel(R) Core (TM) i5-6300U CPU 2.40 GHz 2.50GHz processor, with Windows 10 Professional edition, 64-bit system type.

2.1.2. Microphone

- Microphone Type: BOYA (BY-M1DM).

- Frequency Range: 65 Hz to 18 kHz.

- Impedance: 1000 Ohm or less.

-Power requirement: LR44 (included).

-Sensitivity: -30 dB +/ -3 dB/0dB=1V/P a, 1 KHz.

-Signal/Noise: 74 dB SPL.

2.1.3. Smartphone

We used a Redmi 9 Smartphone (version MIUI V: MIUI Global 12.0.4) Stable for the acquisition of our speech signal.

2.2. Software

2.2.1. Recorder software: Wave Editor1.101 software is mainly for audio editing. It has all the necessary tools to allow us to modify all or part of the song. It can add a personal touch to the sound by using the various effects built into it. Files are saved with the extension (.wave)

The files are then prepared for organization and renaming, thus moving them to the workspace to apply the analysis methods developed.

2.2.2. MATLAB: MATLAB R2008b

The name MATLAB stands for Matrix Laboratory. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. Furthermore, MATLAB is a modern programming language and problem-solving environment: it has sophisticated data structures, contains built-in editing and debugging tools, and supports object-oriented programming. These factors make MATLAB an excellent tool for teaching and research **[23]**.

2.3. Why we use MATLAB

• It is easy to use programming software

- Several predefined functions to analyse and represent data: you can do elaborate things with very little code.

• Particularly suitable for speech signal analysis

- There is a special signal (and image) analysis module.

- Several predefined functions (spectral analysis, filtering, etc.).

• Creation of beautiful figures.

- Stable and aesthetic figures.

- Automation of figure creation.

• Creation of interfaces to analyse various data.

- Ex. Alignment of acoustic data and articulator data.

• **CAASA Window:** Represents the first window of the application where we can save new recordings with the Record button or select a file for the extraction of the acoustic parameters with the Browser button or extracting result matrix with process button and exit the application with the Exit button as shown in Figure 23.

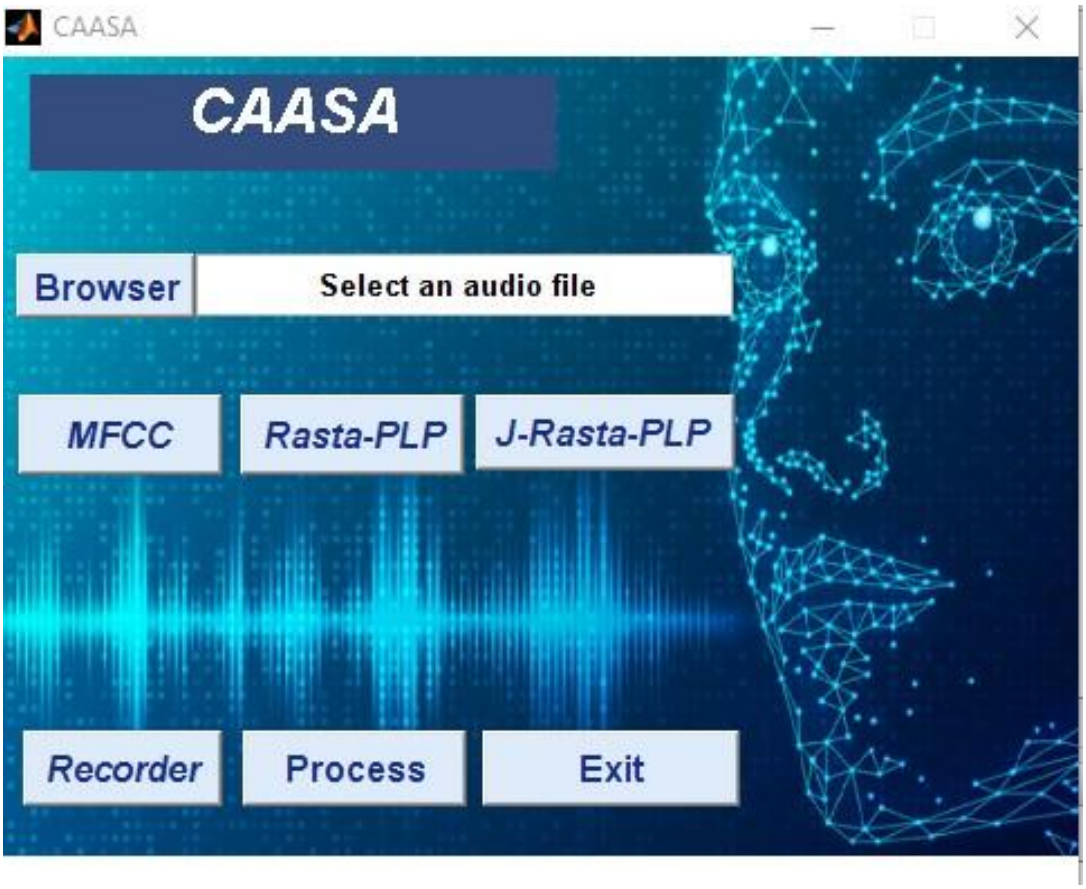

Figure 23. CAASA interface.

- Browser button: Allow us to search for the location of our speech database that we gave it the name 'corpusData' to select the under folder (simple) which contains repetitions of a particular word

we take like an example the 'Airconditioner' folder (simples) as shown in (Figure 24).

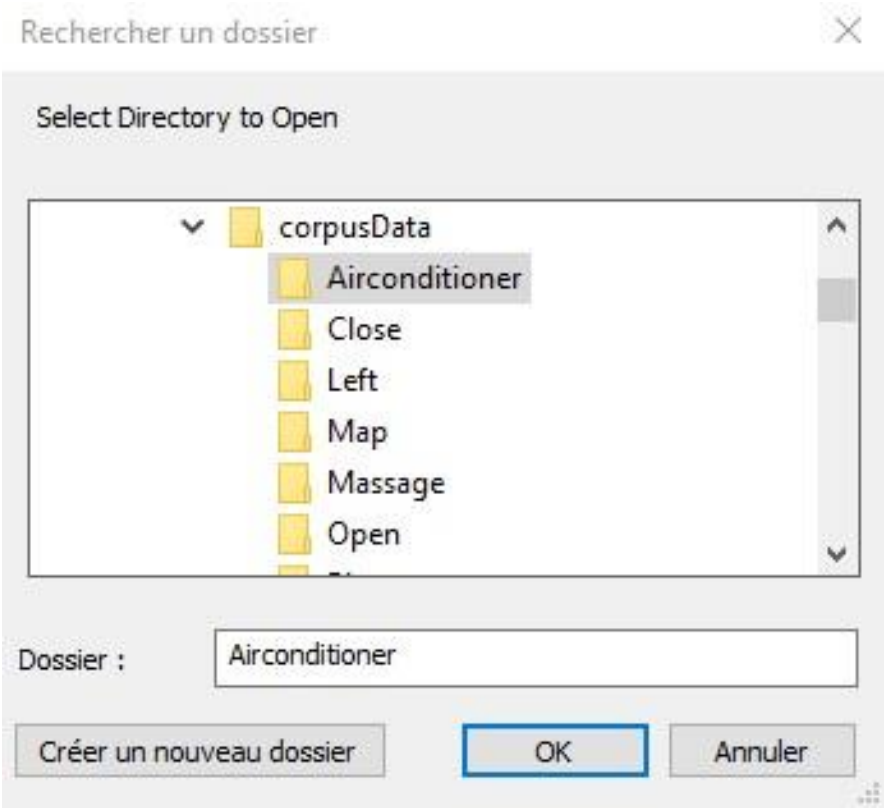

Figure 24. Search window.

- After we choose a simple by “Browser” button (Figure25) we choose an audio file from this under folder(simples), as shown in figure bellow (Figure26).

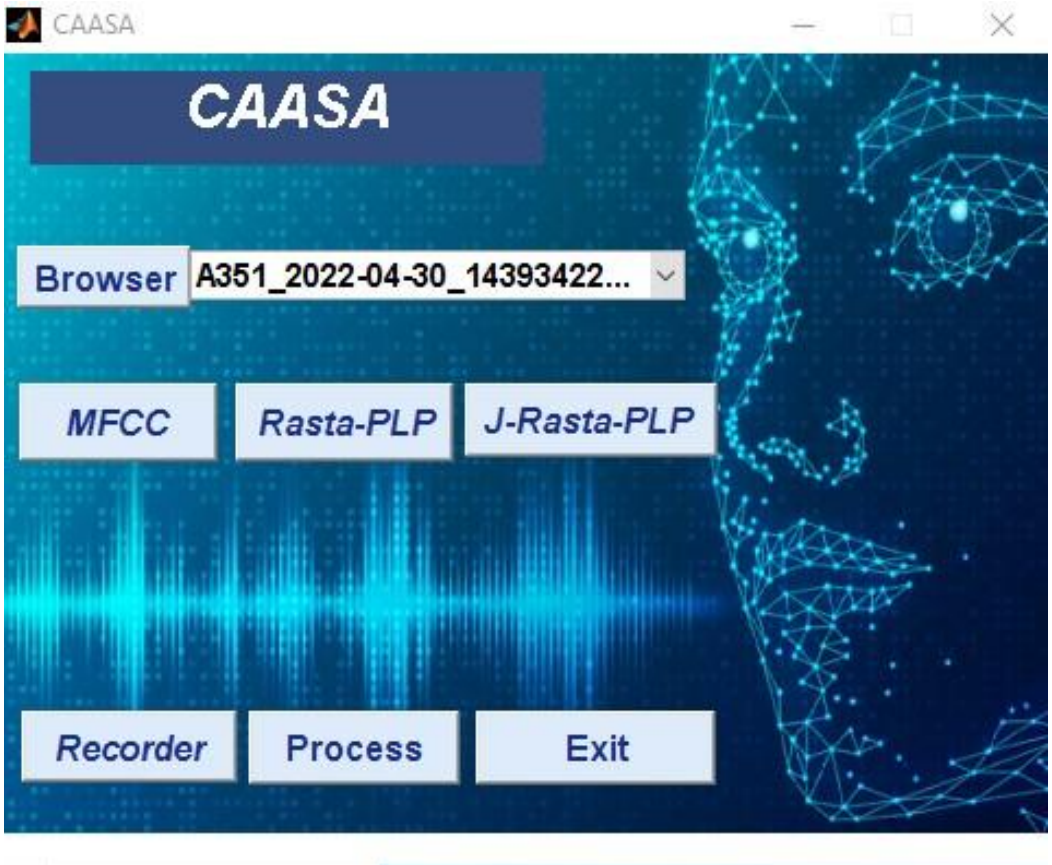

Figure25. CAASA interface when we choose an audio corpus.

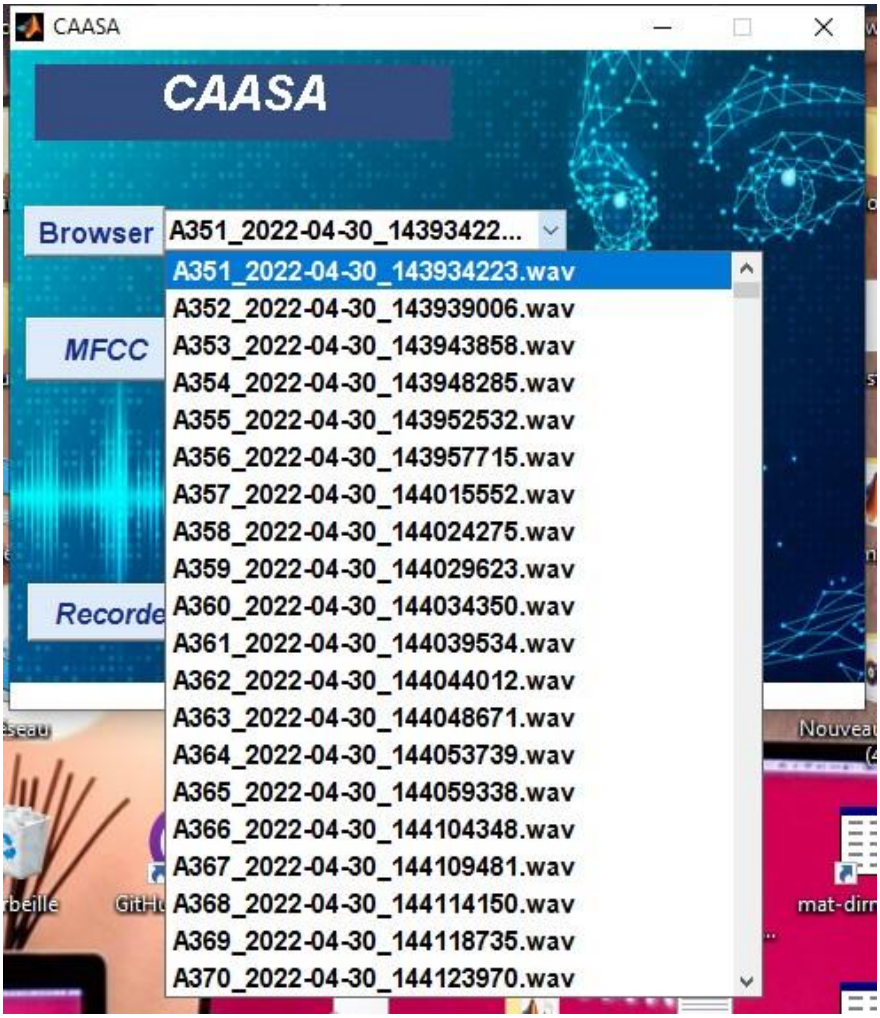

Figure26: choose an audio file from the simple that we chose.

- Record button: a window is displayed, which allows us to record when we press the “Record” button for the first time, the second press saves what we recorded in “Aircondioner” folder that we selected before. We can delete the recording with “Delete” button, and “Back” button that return us to the CAASA interface (Figure 27).

.

.

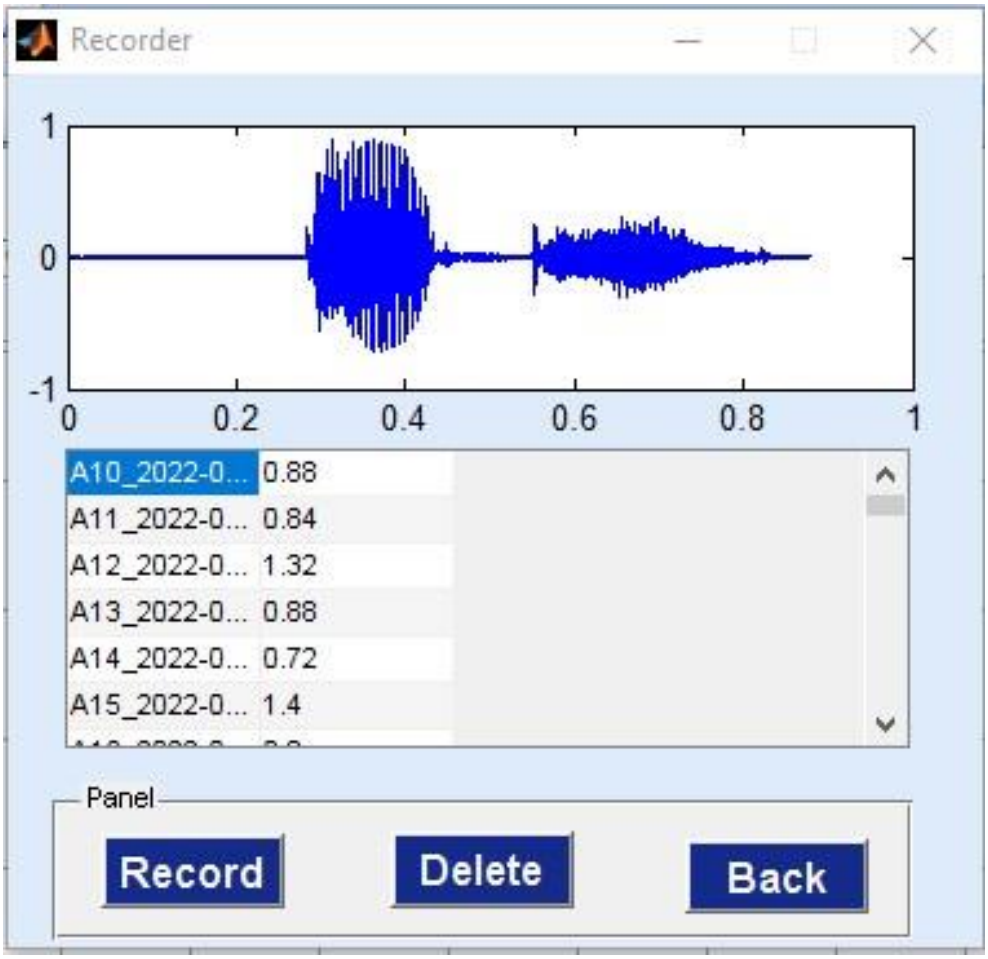

Figure 27. Recorder Window.

-Process button: Allow us to extract the matrixes result after we select a corpusData that contain simples (Figure28) we get an output matrices of the three methods is called "corpusDataoutput" (Figure30), for use these results in the training of the ASR system and test performance and the figure (Figure 29) shown the start of the processing.

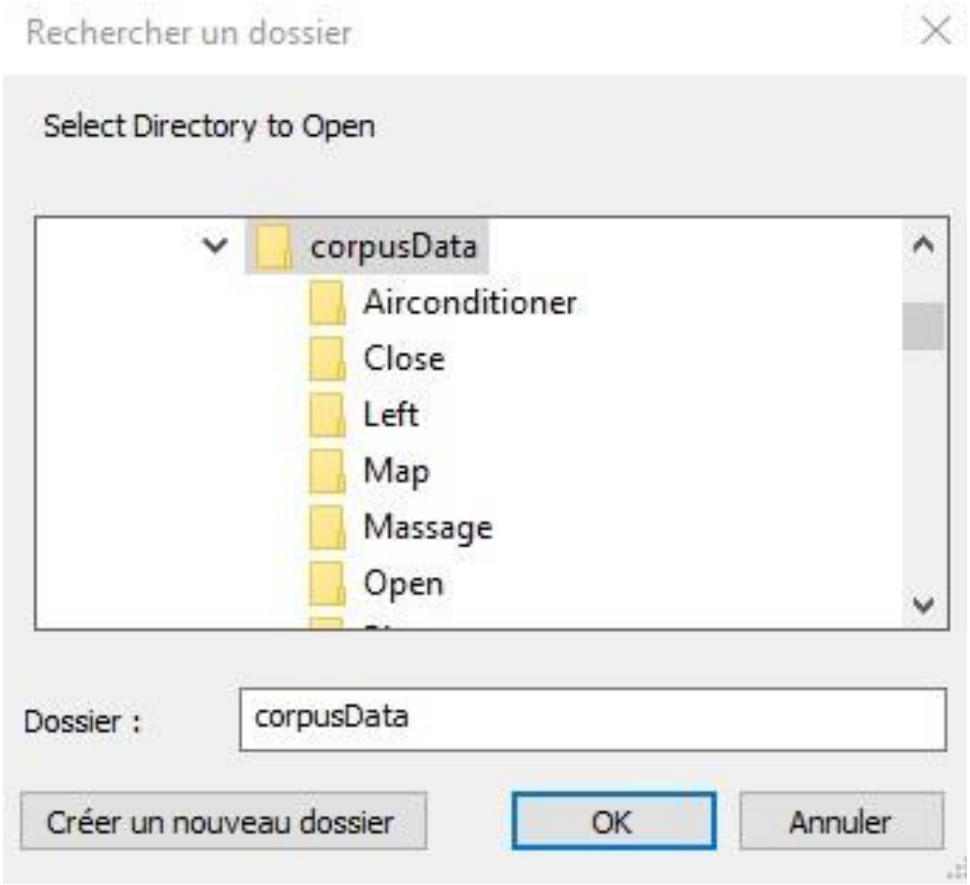

Figure28. Select corpusData.

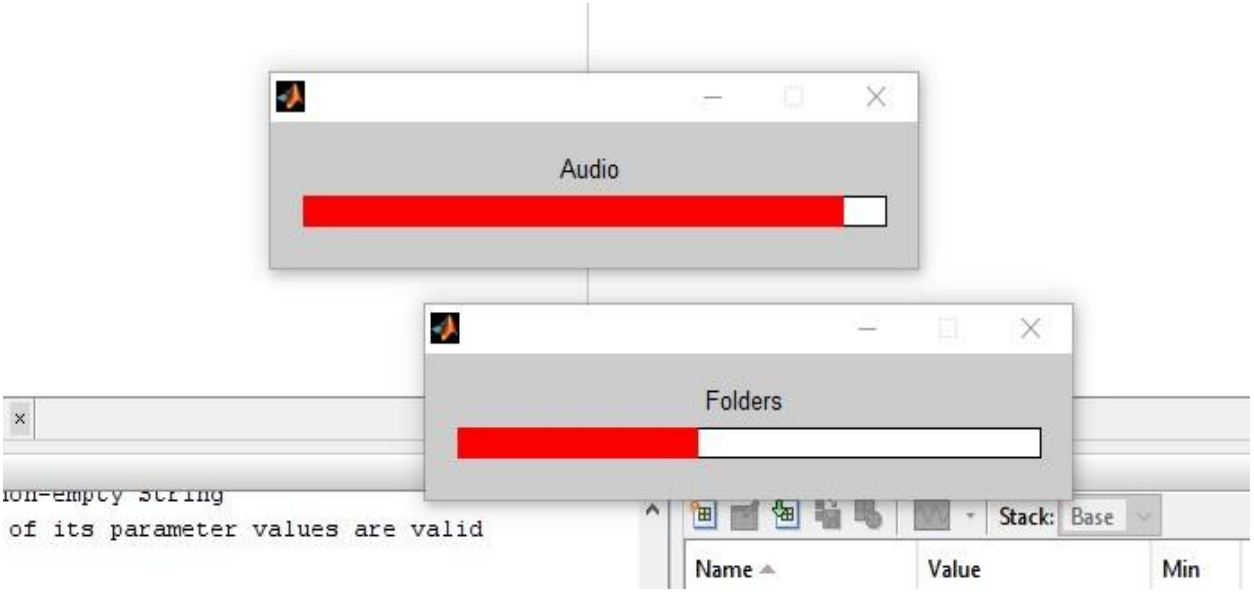

Figure29. Start the processing to get the output matrices.

The corpusDataoutput (Figure29)contain:

-The matrix “mat-dirname_stat_end” is for knowing where the beginning and the end is of each simple. The matrix “mat-resultJrasta plp” is J-Rasta-PLP’s matrix.

-The matrix “mat-resultJmfcc” is MFCC’s matrix.

-The matrix “mat-resultrasta plp” is Rasta-PLP’s matrix.

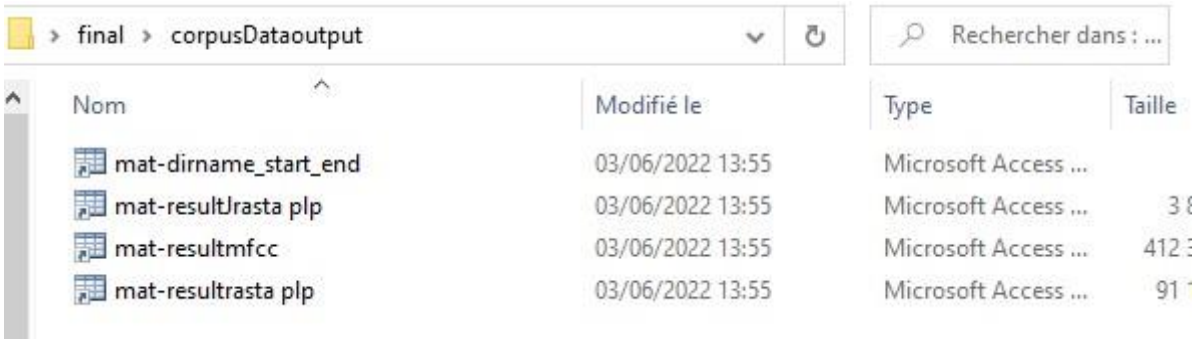

Figure30. corpusDataoutput.

• **MFCC window:** It contain three button: “play” button that play the record file that chosen, “Analyse“ button that represents the results of analysis by the MFCC method (figure30), and the “Back” button that returns us to the CAASA interface.

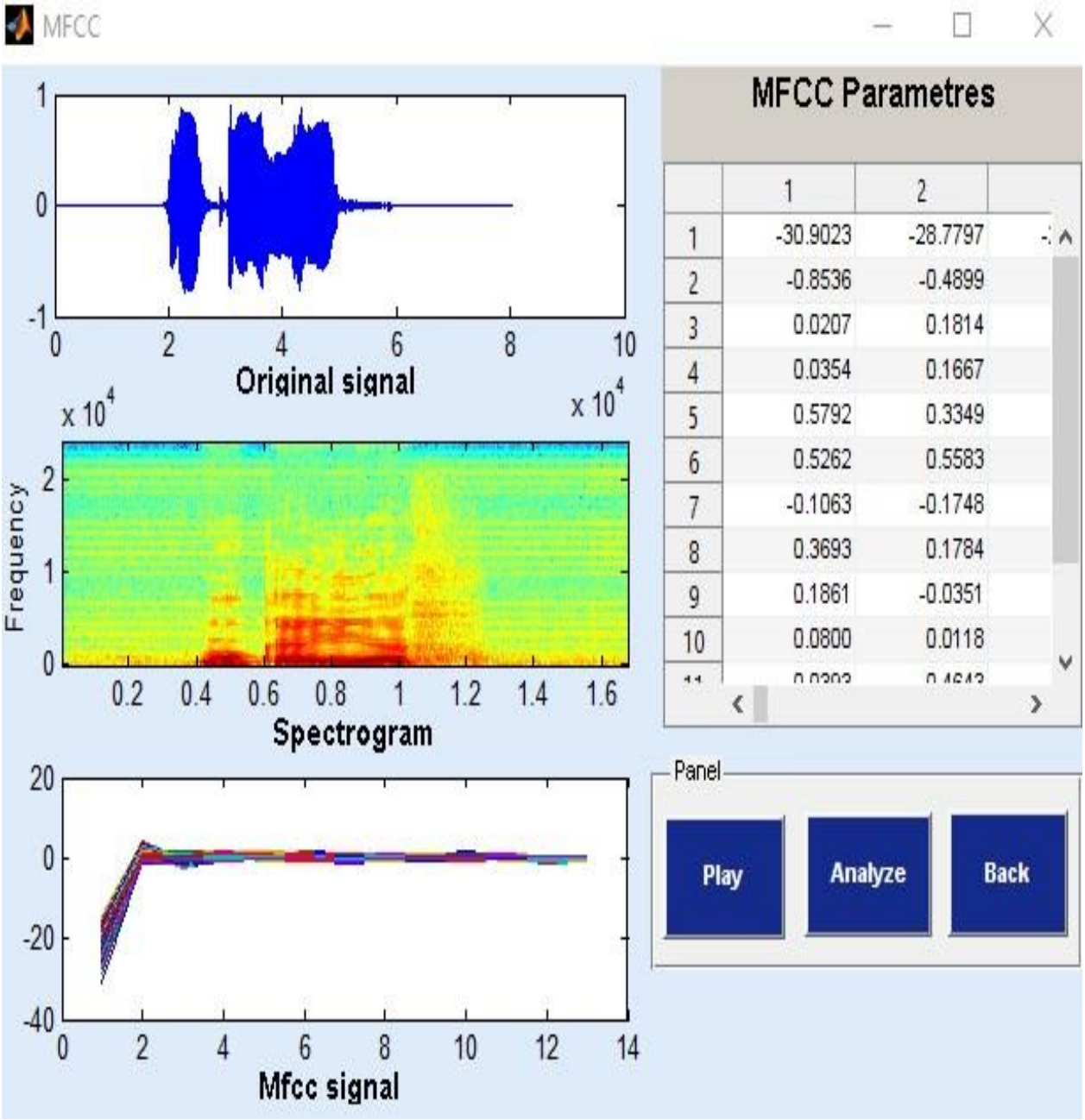

Figure 31. MFCC analysis method execution window.

• **Rasta-PLP window:** It contains three buttons: the “play” button that plays the record file that is chosen, the “Analyse“ button that represents the analysis results by the Rasta-PLP method (Figure 31), and the “Back” button that returns us to the CAASA interface.

.

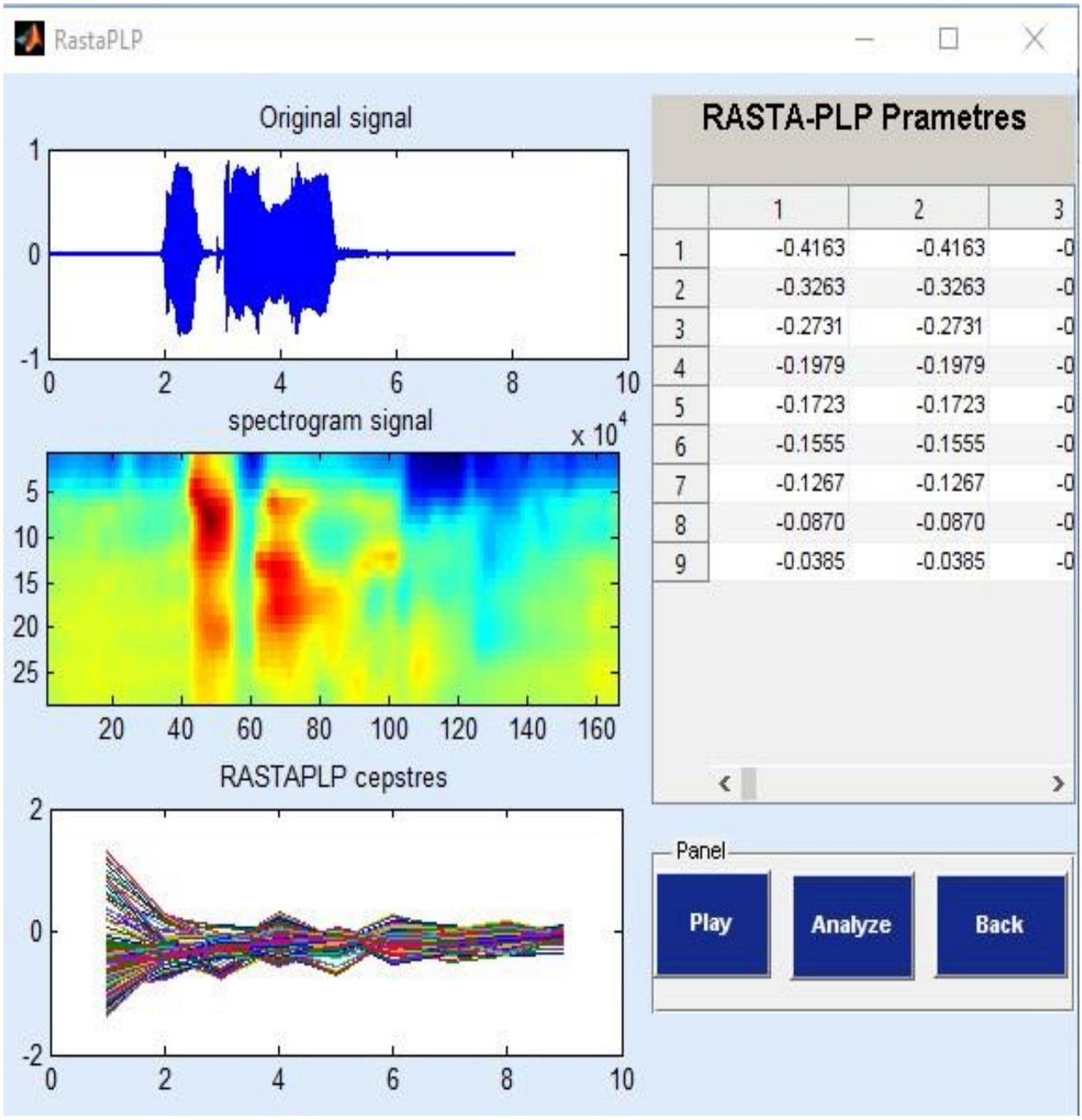

Figure 32. Rasta-PLP analysis method execution window.

•J- Rasta-PLP window: It contain button: the "play" button that play the record file that chosen, the "Analyse "button that represents the analysis results by the Rasta-PLP method (Figure 32), and the "Back" button that return us to the CAASA interface.

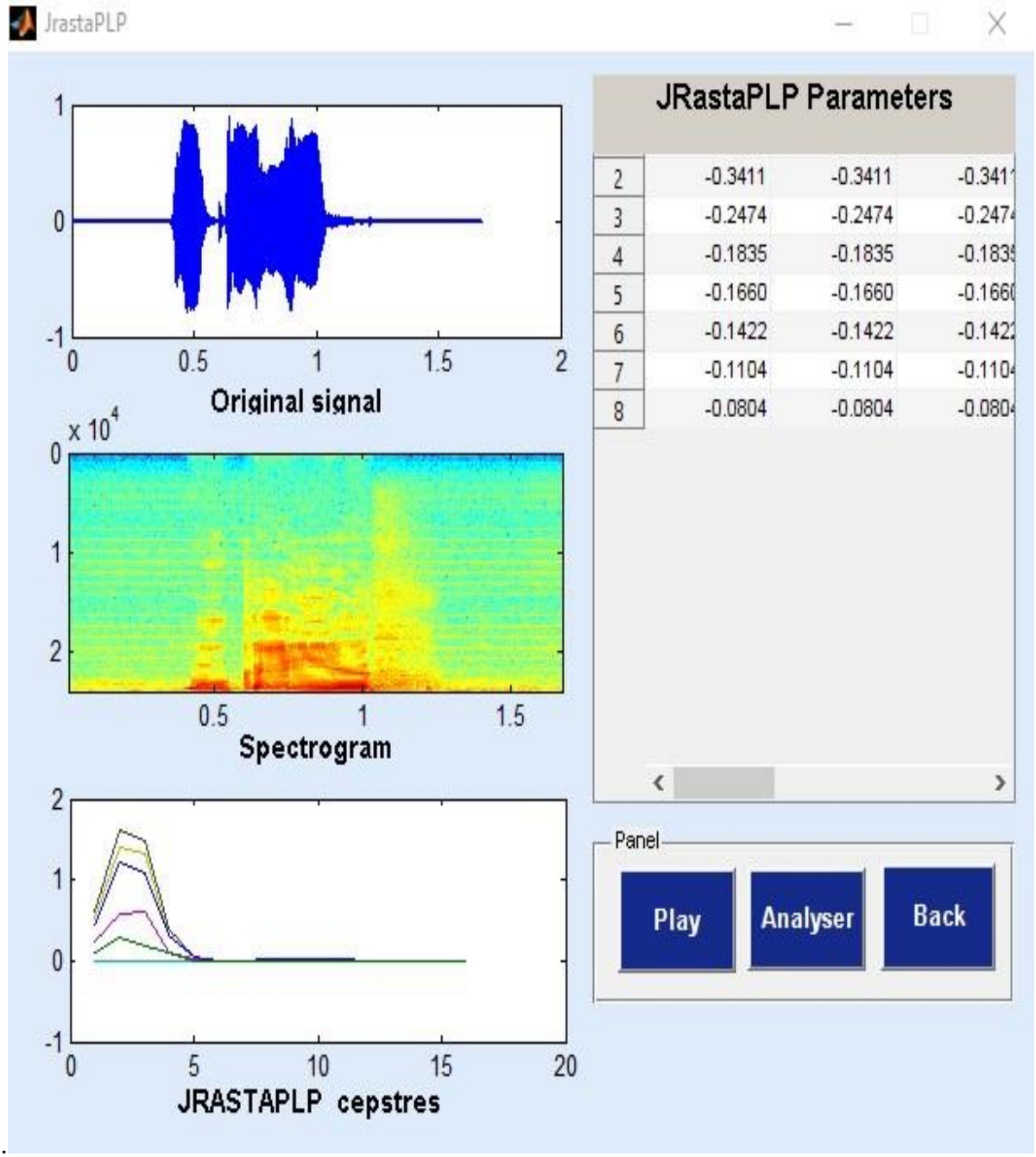

Figure 33.J-Rasta-PLP analysis method execution window.

4. Conclusion

After completing our design, we gave the necessary tools for the realization of our work. We also presented the development environment and defined it; also, we explained why we chose MATLAB.

At the end we presented our application giving some screenshots that explain the progress and operation of our work, as well as the results obtained.

We also gave an example of the extraction of acoustic parameters for an Arabic corpus by the three methods implemented MFCC, Rasta-PLP and J-Rasta-PLP.

Conclusion and Perspectives

Speech is one of the natural means of communication between human beings, several machines have been developed to analyse, recognize, and produce speech. Speech technology is evolving rapidly, and a few tools have been developed for better implementation. ASR allows the computer to convert the voice signal into text or commands through the process of identification and understanding. Speech recognition is connected to many fields of physiology, psychology, linguistics, computer science and signal processing, and it is even linked to the body language of the person, its objective is to achieve natural language communication between beings. Human and computer tools. The acoustic signal of speech has characteristics that complicate its interpretation and increase the amount of data to be processed.

ASR systems are essentially based on the extraction of acoustic parameters as one of the first phases of automatic recognition, allowing these methods the MFCC, Rasta-PLP and J-Rasta-PLP which are the most used in ASR systems.

The work presented in this master's thesis focuses on collect an Arabic speech database then build CAASA system to extract the characteristic information of the speech signal by eliminating as much as possible the redundant parts, our system takes a WAVE signal as input and returns a vector of parameters with three methods (MFCC, Rasta-PLP and J-Rasta-PLP). To verify the performance of our proposed methods we used the obtained results as inputs in a speech recognition system with Multilayer Perceptron-like Artificial Neural Networks (MLP). The result of parameters show that Rasta-PLP method gives higher average corpusData rate compared to MFCC method. J-Rasta-PLP method gives higher average corpusData rate compared to both of MFCC and Rasta-PLP methods, this goes back to the advantages and disadvantages that I mentioned in the first chapter.

In this work, we have described how the speech is produced and how we take the idea to make the machine do the same steps that we call it ASR system and its architecture and the most feature extraction method that used, also we mentioned the existing speech database, all that is presented in the first chapter.

Our bibliographical study was directed in the second chapter on all steps and environment to collect a speech database and the steps do building an application that allow us to extract the vocal parameters by input a vector and output a matrix that use it to know the performance of system with each method, and all that in chapter two.

After the study necessary for the collection of speech dataset and extract their parameters, we gave the necessary tools for the realization of our work and presented the development environment and we presented our application by giving some screenshots, and all that we do it in chapter three.

We conclude the principle of each method:

-The first MFCC method reduces the frequency information of the speech signal to a small number of the coefficients, it is easy and relatively fast to calculate more low frequency information at high frequencies due to the Mel filter banks, therefore it behaves more like the human ear than other techniques.

-Rasta-PLP's method applies a bandpass filter to each spectral component of the critical band spectrum estimate, these features are best used when there is a mismatch in the analog input channel between development systems and when fielded, a lower-order analysis yields best estimates of the data in extracting acoustic parameters from speech.

-J-Rasta-PLP's method to further improve the PLP method, defines the J-RASTA method, plus resistant to additive noise than the RASTA method, by adding low-pass filtering in the spectral domain., the difference between RASTA and J-RASTA is at the logarithm level (5th step): log(x) for RASTA and log (1 + Jx) for J-RASTA. J-RASTA processing addresses the convolutional noise and additive noise.

The difficulties that i encountered when collecting corpus data were that we don't have professional microphone and studio to record with high quality and the environment of recording (car) contain a lot of noise.

The main objective of this multi-criteria comparison is to help researchers determine the feature extraction method according to the most important standards to speech recognition.

This thesis has opened many perspectives that can be summarized as follows:

. Building a complete speech recognition system by completing the stages of its construction that we explained in the first part.

• Applying other methods to this rule and seeing the results to compare them with the results we obtained in this work.

• Increasing the number of iterations and people of the base that we built in this work.

• Collecting a database of people with disabilities.

References

[1] Abdelhamid, A. A., Alsayadi, H. A., Hegazy, I., & Fayed, Z. T. (2020, September). End-to-end arabic speech recognition: A review. In *Proceedings of the 19th Conference of Language Engineering (ESOLEC'19), Alexandria, Egypt* (pp. 26-30).

[2] Algihab, W., Alawwad, N., Aldawish, A., & AlHumoud, S. (2019, July). Arabic speech recognition with deep learning: A review. In *International Conference on Human-Computer Interaction* (pp. 15-31). Springer, Cham.

[3] Ishak, D. (2017). *A speaker recognition system based on vocal cords' vibrations* (Doctoral dissertation, Université de Valenciennes et du Hainaut-Cambresis; Université de Balamand (Tripoli, Liban)).

[4] Saksamudre, S. K., Shrishrimal, P. P., & Deshmukh, R. R. (2015). A review on different approaches for speech recognition system. *International Journal of Computer Applications*, *115*(22).

[5] Gaikwad, S. K., Gawali, B. W., & Yannawar, P. (2010). A review on speech recognition technique. *International Journal of Computer Applications*, *10*(3), 16-24.

[6] The Egyptian Arabic Speecon Database, ELRA catalog ELRA-S0308

[7] Selouani, S. A., & Boudraa, M. (2010). Algerian Arabic speech database (ALGASD): corpus design and automatic speech recognition application. *Arabian Journal for Science and Engineering*, *35*(2), 157-166.

[8] Ajami Alotaibi, Y., Alghamdi, M., & Alotaiby, F. (2010, June). Speech recognition system of Arabic alphabet based on a telephony Arabic corpus. In *International conference on image and signal processing* (pp. 122-129). Springer, Berlin, Heidelberg.

[9] Deroo, O., Ris, C., Malfrere, F., Leich, H., Dupont, S., Fontaine, V., & Boite, J. M. (1997). Hybrid HMM/ANN system for speaker independent continuous speech recognition in French. *Proceedings of PRORISK*, *97*, 137-141.

[10] Zue, V., Seneff, S., & Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. *Speech communication*, *9*(4), 351-356.

[11] OrienTelMorocco MCA, ELRA catalog ELRA-B0004.

[12] Ajili, M., Bonastre, J. F., Kahn, J., Rossato, S., & Bernard, G. (2016, May). Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 726-733).

[13] OrienTel United ArabEmirates MCA, ELRA catalog ELRA-B0010.

[14] OrienTelTunisia MCA, ELRA catalog ELRA-B0005

[15] Harrag, A. and Mohamadi T., " QSDAS: New Quranic Speech Database for Arabic Speaker Recognition", The Arabian Journal for Science and Engineering, vol. 35, no. 2C, pp. 7-13, December 2010. [B] Mihelic, F., Gros, J., Nöth, E., & Warnke, V. (2000, May). Labeling of Prosodic Events in Slovenian Speech Database GOPOLIS. In *LREC*.

[16] Elmahdy, M., Gruhn, R., Minker, W., & Abdennadher, S. (2009, October). Modern standard Arabic based multilingual approach for dialectal Arabic speech recognition. In *2009 Eighth International Symposium on Natural Language Processing* (pp. 169-174). IEEE.

[17] Mohamed Elmahdy, Rainer Gruhn, Wolfgang Minker, and Slim Abdennadher (2009), "Survey on common Arabic language forms from speech recognition point of view," Proceedings of the International Conference on Acoustics (NAG-DAGA), Rotterdam, Netherlands, pp. 63-66.

[18] Al-Anzi, F. S., & AbuZeina, D. (2022). Synopsis on Arabic speech recognition. *Ain Shams Engineering Journal*, *13*(2), 101534.

[19] MFCC Technique for Speech Recognition, Uday Kiran — June 13, 2021, (accessed 1 April 2022), https://www.analyticsvidhya.com/blog/2021/06/mfcc-technique-for-speech-recognition/

[20] Hönig, F., Stemmer, G., Hacker, C., & Brugnara, F. (2005). Revising perceptual linear prediction (PLP). In *Ninth European Conference on Speech Communication and Technology*.

[21] Hermansky, H., Morgan, N., Bayya, A., & Kohn, P. (1991, December). RASTA-PLP speech analysis. In *Proc. IEEE Int'l Conf. Acoustics, speech and signal processing* (Vol. 1, pp. 121-124).

[22] Ogawa, H. (1997). *More robust J-RASTA processing using spectral subtraction and harmonic sieving*. Technical Report TR-97-031, ICSI, Berkeley.

[23] Houcque, D. (2005). Introduction to Matlab for engineering students. *Northwestern University*, (1).

Résumé

De nos jours, les applications de reconnaissance vocale se retrouvent dans de nombreuses activités. Le système de paramètres joue un rôle important dans l'ASR où le but est d'extraire les informations caractéristiques du signal de parole en éliminant autant de parties redondantes que possible. Notre travail consiste également à construire un jeu de données arabe et à extraire ses coefficients audios via trois méthodes d'extraction de caractéristiques : « MFCC » (coefficients cepstraux de fréquence Mel), « RASTA-PLP » (spectre relatif PLP) et J-Rasta-PLP. Ce travail vise à comparer les performances de nos méthodes proposées, nous avons donc utilisé les résultats obtenus comme entrée dans un système de reconnaissance de la parole avec des réseaux de neurones artificiels multicouches (MLP).

Mots clés : Reconnaissance automatique de la parole, MFCC, Rasta-PLP, J-Rasta-PLP, corpus de parole arabe.

Abstract

Nowadays, speech recognition applications can be found in many activities. Parameter system plays an important role in ASR system where the goal is to extract the characteristic information of the speech signal by eliminating as many redundant parts as possible. Our work is also to build an Arabic dataset and extract its audio coefficients via three feature extraction methods: “MFCC” ' (Cepstral coefficients with frequency Mel), 'RASTA-PLP' (PLP relative spectrum), and J-Rasta-PLP. This work aims to compare the performance of our proposed methods, so we used the obtained results as input into a speech recognition system with multilayer artificial neural networks (MLP).

Keywords: Automatic Speech Recognition, MFCC, Rasta-PLP, J-Rasta-PLP, Arabic Speech Corpus.

ملخّص

في الوقت الحاضر ، يمكن العثور على تطبيقات التعرف على الكلام في العديد من الأنشطة. يلعب نظام المعلمات دورًا مهمًا في نظام ASR حيث يتمثل الهدف في استخراج المعلومات المميزة لإشارة الكلام من خلال التخلص من أكبر عدد ممكن من الأجزاء الزائدة عن الحاجة..يتمثل عملنا أيضًا في بناء مجموعة بيانات عربية واستخراج معاملاتها الصوتية عبر ثلاث طرق لاستخراج الميزات: "MFCC" (معاملات Cepstral ذات التردد Mel) و "RASTA-PLP" (الطيف النسبي PLP) و J-Rasta-PLP ا. يهدف هذا العمل إلى مقارنة أداء طرقنا المقترحة ، لذلك استخدمنا النتائج التي تم الحصول عليها كمدخلات في نظام التعرف على الكلام مع الشبكات العصبية الاصطناعية متعددة الطبقات (MLP).

الكلمات المفتاحية: التعرف التلقائي على الكلام ، MFCC، Rasta-PLP، J-Rasta-PLP ، مجموعة الكلام العربية.