DEMOCRATIC AND POPULAR REPUBLIC OF ALGERIA
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
UNIVERSITY OF MOHAMED BOUDIAF-M'SILA

FACULTY OF TECHNOLOGY
ELECTRONICS DEPARTMENT
N° :.....

DOMAIN : SCIENCE AND TECHNOLOGY
FILIERE : TELECOMMUNICATION
OPTION : TELECOMMUNICATION SYSTEMS

Dissertation Submitted in partial fulfilment of the requirements
For the Master Academic Degree
By:

AMEUR Nehal

Entitled

Automatic Speaker Recognition using Mel Frequency Cepstral Coefficients (MFCC) and Fast Fourier Transform (FFT)

Presented in front of the jury composed by :

Dr. KENANE Hadi	University of Mohamed Boudiaf - M'sila	President
Dr. KHENNOUF Salah	University of Mohamed Boudiaf - M'sila	Superviser
Dr. SAHED Mohamed	University of Mohamed Boudiaf - M'sila	Examiner

June : 2023

ACKNOWLEDGMENTS

بِسْمِ اللَّـهِ الرَّحْمَـٰنِ الرَّحِيمِ

قال الله تعالى:" وَآخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ" سورة يونس

In the name of ALLAH, the greatest and the most clement and merciful... Thanks to him for having guided me towards the right path and for having helped me throughout my studies.

My thanks go to my supervisor, Dr Khennouf Salah, for agreeing to direct this work, for his support, and his foresight that have been invaluable to me. I would like to thank, also, the members of the jury who gave me the great honor for evaluating this work.

Thanks to all lecturers who spared no effort to give us their knowledge during my university studies.

Finally, I would like to thank all those who have contributed in any way to the preparation of this thesis and to the success of my university career.

DEDICATION

I dedicate this humble work, which is the fruit of my long studies and deep gratitude

To My support in life and shoulder my dear father laid (Mourad)

To whom it was said that heaven is under her feet, my dear mother, Nabila Chaanbi

For all the support and sacrifices they made

To my wonderful sisters Meryem, Mayssoune and Marwa

To my dear brother Akram

May God protect you and my family, with my wishes for continued health and wellness

To my dear friends

My good friend Khalil Berroubi

My best friend Radja Hamdaoui

My friend, despite the distances and different countries, is Fatima Hussain

My dear friend, Aya Nouasria

And my classmates are all in his name

To my cousins who stood by me, Mohammad and Rida Rahmani

To the person who gave me that last batch at the end of in order to continue of Mr. Abd Essamed

To everyone who participated in this work from near and far

THANK YOU ALL

ABBREVIATIONS LIST

ASD : Automatic Speaker Detection

ASI : Automatic Speaker Identification

ASInd : Automatic Speaker Indexing

ASR : Automatic Speaker Recognition

ASS : Automatic Speaker Segmentation (ASS)

AST : Automatic Speaker Tracking

ASV : Automatic speaker verification

DCT : Discrete Cosine Transform

DFT : Discrete Fourier Transformation

DTW : Dynamic Time Warping

FFT : Fast Fourier Transform

GMM : Gaussian Mixture Model

HMM : Hidden Markov Models

LFCC : Linear Frequency Ceptral Coefficients

LFSC : Linear Frequency Spectral Coefficients

LPC : Linear Predictive Coefficients

LPCC : Linear Predictive Cepstral Coefficients

MFCC : Mel Frequency Cepstral Coefficients

MFSC : Mel Frequency Spectral Coefficients

TFPC : Time Frequency Principal Components

FIGURES LIST

TABLES LIST

CONTENTS

Chapter-2

Proposed Methods for Speaker Verification

Chapter-3

Conducted Experiments and Obtained Results

Introduction

Introduction

Speech is the most natural form of communication for humans, but the development and widespread use of telephones, audio-phonic storage devices, radio, and television have made it even more important. Speech processing has become increasingly important in a variety of applications, such as voice compression, augmentation, synthesis, and recognition.

- **Motivation**

The security system, by which people can be automatically identified, is in great demand for many purposes such as ; telephony, banking, security services, access to information and databases, etc. Among the possible methods that can be used for these systems are those based on biometric measurement. The latter is based on measuring the physiological and/or behavioural characteristics of each individual. These characteristics include fingerprints, retina, iris, face, hand signature, hand geometry and wrist veins [Jai, 1999]. In addition, the voice can also be used as a physiological feature, as it not only conveys the message, but also carries information about the speaker's gender, mood and identification [Cam, 1977]. All these information are embedded in the voice, and the ability to use this information comes naturally to human beings.

Speaking is the most preferred medium for human communication, to transmit thoughts and information. Automatic Speaker Verification systems include discriminating between persons and holding a conversation involving several interlocutors. The ability to verification from speakers inspired us to realize a system that can verification from multiple speakers using artificial intelligence methods.

- **Objectives**

The objectives of this thesis are:

- Conception of a system that automatically verification from two speakers.
- Design a speaker recognition model employing the MFCC and FFT extraction approach.

- **Document organisation**

This document is follows : The first chapter provides a brief overview of the human phonatory apparatus and auditory system to help readers understand how speech is produced and perceived and some characteristics of the speaker recognition systems are described.

The suggested method for verification speakers, using the features extraction technique (Mel Frequency Cepstral Coefficients (MFCC) and Fast Fourier Transform(FFT)), were covered in the second chapter. The speaker verification experiments and the obtained results (verification rate) are well explained in the third and last chapter.

Finally, this document is with a general introduction, that exposes the problem of this thesis, and ended with a general conclusion that presents the main features of our system and the obtained results.

Chapter-1

Speech Production and Perception and Automatic Speaker Recognition

Chapter-1
Speech Production and Perception and Automatic Speaker Recognition

Introduction

Speech is the articulated, symbolic human language used to communicate thoughts, information and ideas. It involves forming audible sounds, words that contain syllables and phonemes. Speech production is a highly complex mechanism, with specific parameters to each individual marking the uttered sounds. Human beings are able to distinguish between sounds picked up by their auditory system, discriminate between speakers and identify them if they are known a priori.

Automatic speaker recognition is a branch of biometric authentication that enables the automatic recognition of a person's identity by extracting intrinsic characteristics from their voices. It is the most natural and economical choice for man-machine communication systems, because collecting speech data is much more practical than other patterns (DNA, IRIS, etc.) on the one hand. On the other hand, speech is the dominant mode of information exchange for humans and is the dominant mode of information exchange for man-machine communication systems.

This chapter, which is dedicated generally to biometrics, will study the mechanisms of speech production and perception to better understand how sounds are processed by the human auditory system. It will also focus on the different automatic speaker systems and diverse speech processing tasks.

1.1 Biometrics

Biometrics is a science derived from biology and statistics, which consists in analysing the biological, physiological and behavioral characteristics of human beings (saliva, iris, voice, etc.) with a view to possible recognition (verification or identification) of the identity of a person. It refers to all the technologies and techniques used to identify individuals. [Khe, 2018]

1.1.1 Physiological characteristics

Physiological characteristics are linked to the body shape. The oldest features, which have been used for over 100 years, are fingerprints. Others characteristics are ; face recognition, hand geometry and iris recognition.

1.1.2 Behavioural characteristics

Behavioural characteristics are related to a person's behaviour. The first characteristic to be used, and still widely used today, is the hand signature. There are also other features, more modern, such as the study of typing and voice dynamics.

1.2 Speech production mechanisms

1.2.1 The human phonatory apparatus

The process of speech production is a highly complex mechanism based on an interaction between neurological and physiological systems. Speech begins with neurological activity [Red ,2012]. Once the idea and the will to speak have arisen, the brain directs the operation of the phonatory organs. The functioning of these organs is physiological in nature [Ama,2009]. A large number of organs and muscles are involved in the production of sounds in natural languages. The functioning of the human phonatory apparatus is based on the interaction between three entities: the lungs, the larynx and the vocal tract (see figure 1.1).

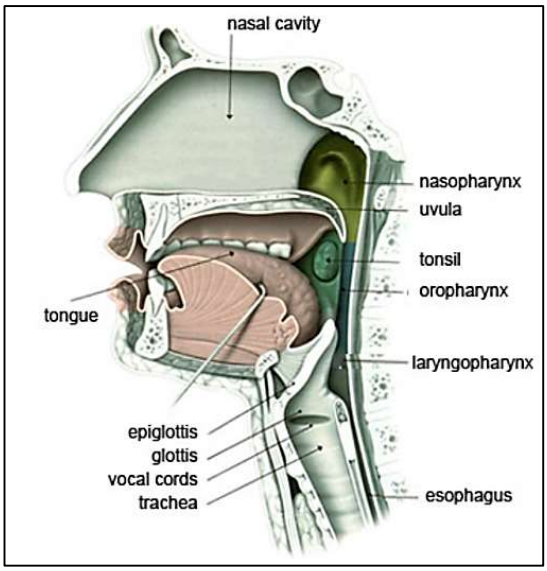


Figure 1.1.: Section of the human phonatory apparatus

The larynx is a cartilaginous structure whose main function is to regulate airflow via the movement of the vocal cords. The vocal tract extends from the vocal cords to the lips in the buccal part, and to the nostrils in the nasal part [Ama,2009].

Air from the lungs is compressed by the action of the diaphragm. This pressurized air then reaches the vocal cords. If the cords are open, the air passes freely, enabling the production of sound. If they are closed, pressure can cause them to vibrate, producing a quasi-periodic sound whose fundamental frequency generally corresponds to the pitch of the perceived voice. The air, which may or may not be set in vibration, continues its journey through the vocal tract and is then propagated into the atmosphere. The shape of the vocal tract, determined by the position of articulators such as the tongue, jaw, lips and soft palate, determines the characteristics of the different speech sounds.

The vocal tract is thus seen as a filter for the different sources of speech production, such as vocal cord vibrations or turbulence generated by the passage of air through the constrictions of the vocal tract [Ama,2009], [Yes,2014]. The resulting sound can be classified as voiced or unvoiced, depending on whether or not the air emitted caused the vocal cords to vibrate [Ama,2009], [Yes,2014], [Out,2016].

1.2.2 Formants

The formants are zones of reinforced harmonics measured in Hertz (Hz). This reinforcement is a function of the size and nature of the resonators. The vocal tract acts like a series of resonators:

- 1st formant frequency F_1 : Depends on aperture (pharyngeal cavity).
- 2nd formant frequency of F_2 : Depends on the position of the tongue and lips (oral cavity).
- 3rd formant frequency F_3 : Depends on lip position.

It should be noted that, in general, the frequency range of a speech signal is considered to be between 100 Hz and 5000 Hz for natural communication and 300 Hz to 3400 Hz for telephony.

1.2.3 The articulation and the acoustic signal

The source of the resonator can be divided into two distinct emissions of different origins : The vocal cords and airflow noise.

1.2.3.1 Vocal cords

In addition to their fundamental frequency, vocal cords produce a spectrum rich in harmonics. This is how they produce voiced sounds.

1.2.3.2 Flow noise

The sound of airflow from the lungs, with a spectrum similar to white noise, creates unvoiced sounds. This action can be modulated by the source-filter model:

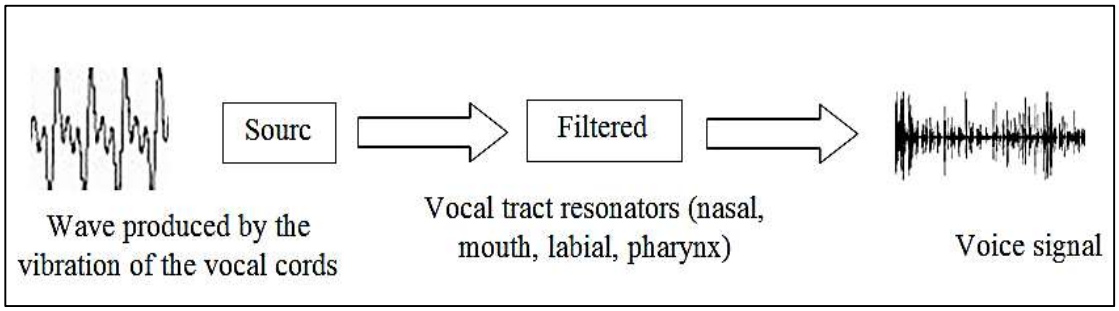


Figure 1.2: Source-filter model of the speech signal

However, a vibrating source placed in front of a resonant cavity will always produce a sound whose frequencies are filtered by the resonator bandwidth.

1.2.4 Voiced and unvoiced sounds

1.2.4.1 Voiced sounds

In this case, vocal cord vibrations modulate the flow of air from the lungs. This results in a variation in the pressure of the air passing through the glottis. This wave then propagates through the vocal tract to the lips.

The frequency of vocal cord oscillation is high for men. It lies between 100 Hz and 150 Hz during normal phonation (i.e., normal conversation). For women, it is higher (200 Hz to 300 Hz), and even higher for children (300 Hz to 450 Hz).

1.2.4.2 Unvoiced sounds

When producing unvoiced sounds, the vocal cords do not vibrate. The sound is produced by the friction of air in the vocal tract between the glottis and the lips. If the air encounters no obstacles in its path, the sound produced is very weak (breath sound). For example, the pronunciation of the letters ‘f’ and ‘t’.

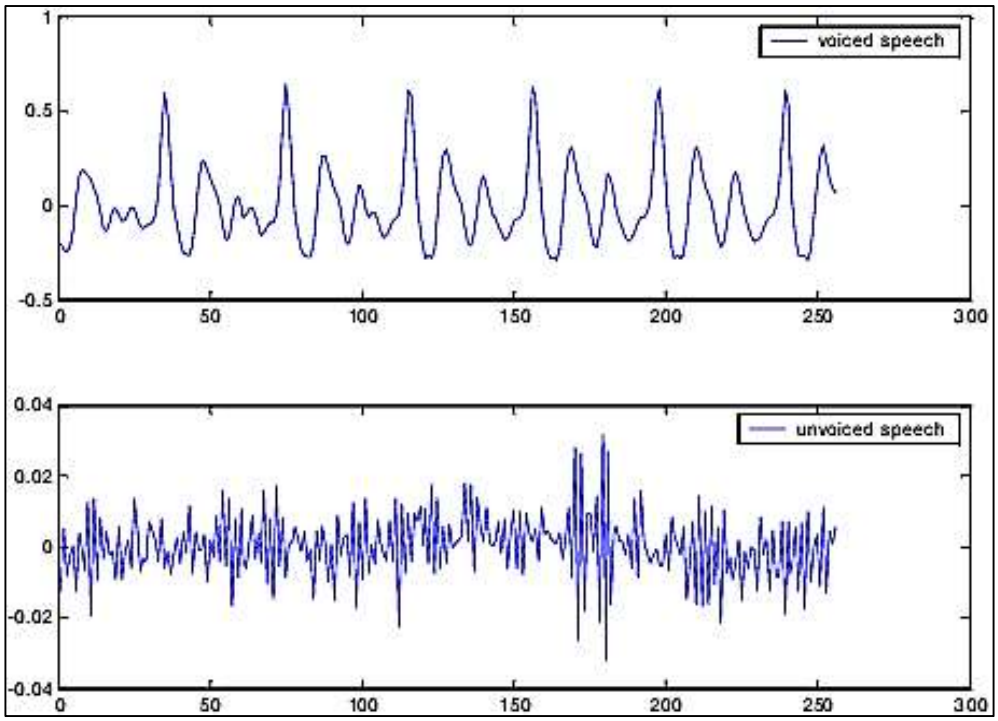


Figure 1.3 : Comparison between voiced and unvoiced speech

1.2.5 Acoustic features of a speech signal

The acoustic features of a speech signal are directly linked to its production in the phonatory apparatus, and are also linked to other perceptual variables such as intensity, rhythm and timbre. These acoustic features are :

- Sound energy, which is linked to air pressure upstream of the larynx.
- The fundamental frequency F_0, which corresponds to the frequency of the opening/closing cycle of the vocal cords, for voiced sounds.
- The spectrum of the speech signal, which results from dynamic filtering of the signal from the larynx by the vocal tract, which can be considered as a succession of acoustic tubes or cavities of various cross-sections.

The temporal evolution of the fundamental frequency F_0 or pitch varies according to the phonemes uttered in a sentence, unvoiced sounds are associated with zero frequency as shown in Figure 1.4 [Vel,2006].

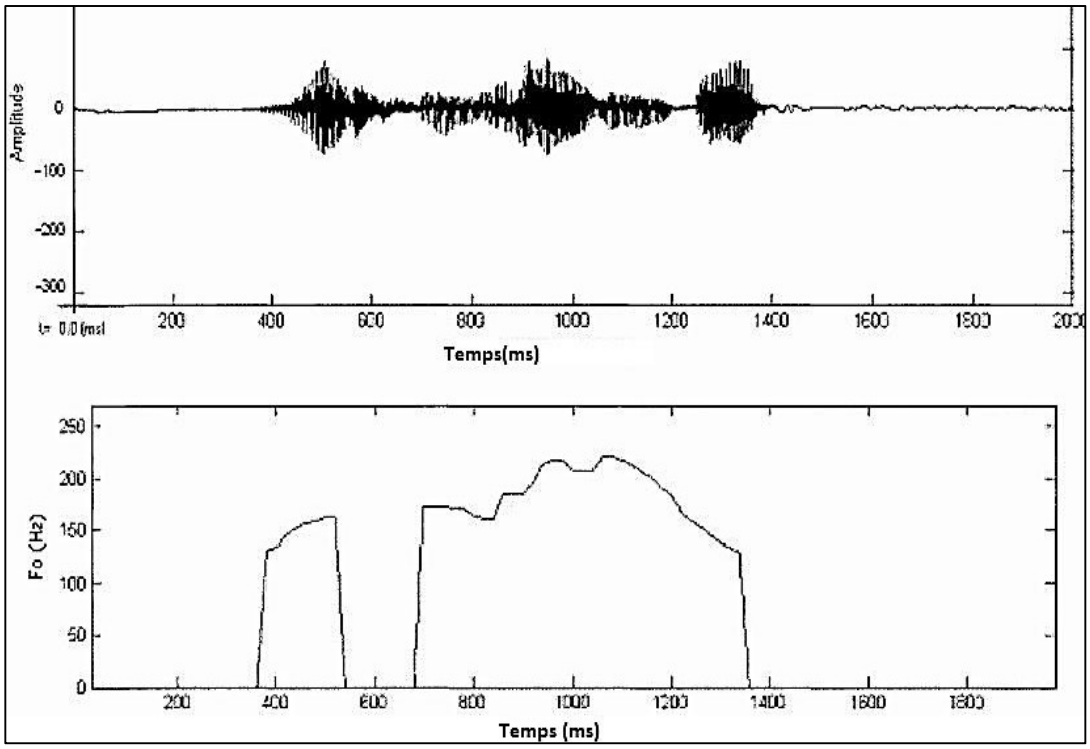


Figure 1.4: Evaluation of vocal cord vibration frequency for the sentence:"I have a cold" .

The fundamental frequency of neighbouring zones evolves slowly over time, with orders of magnitude ranging from 80 Hz to 200 Hz for men, 150 Hz to 350 Hz for women, and 200 Hz to 500 Hz for children.

1.3 Speech perception mechanisms

1.3.1 The human auditory system

The role of the human auditory system is to transform a sound into a series of nervous activities which are transmitted as potentials to the brain. The brain then processes these potentials to guide our behaviour. The ear, which acts as a mechanical transducer by converting pressure information into electrical information, is made up of three parts: the outer ear, the middle ear and the inner ear (see figure 1.5).

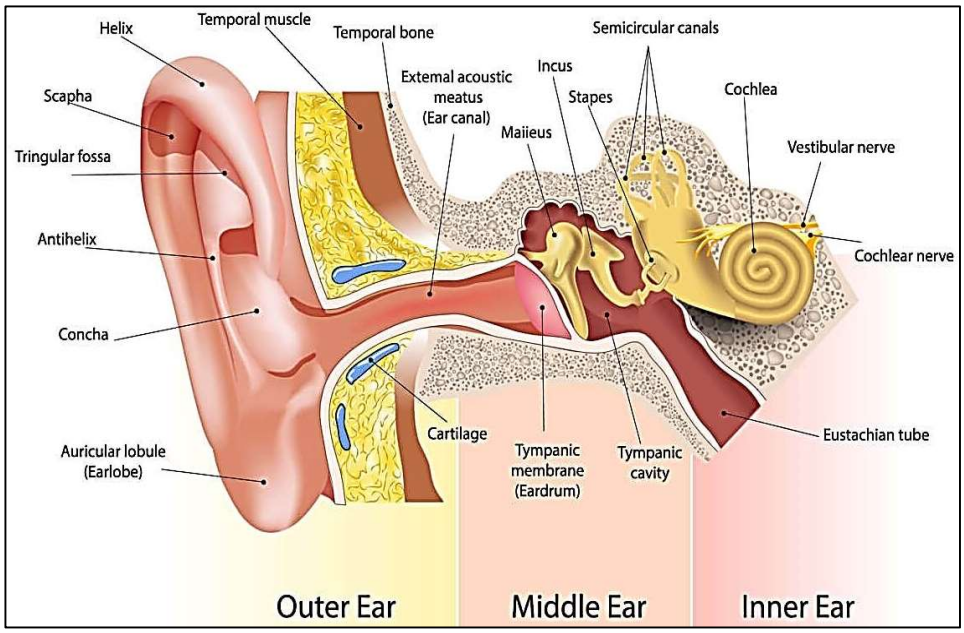


Figure 1.5: Anatomy of the ear

With a range of audible frequencies from 20 Hz to 20 kHz, the ear is able to distinguish a minute variation in pressure caused by a 1W sound with a frequency of 3 kHz at a distance of around 300km, in an ideal completely silent world [Khe,2010].

Human beings are capable of localizing sound with astonishing accuracy. Although this accuracy varies from individual to individual, the localization of sounds coming from the front is more precise. The error increases by up to 20° in other directions. The characteristics of an acoustic wave are modified by the presence of the

head before interacting with the pinna. These modifications provide acoustic cues that are used by the brain to locate a sound source in the azimuthal plane and in elevation.

1.3.2 The limits of auditory perception

The simplest and earliest explorations of psychoacoustics looked at the limits of auditory perception for human beings and distinguished three types of limits: intensity limits, frequency limits, direction limits and time limits.

1.3.2.1 Intensity limit

The most sensitive human ear perceives sounds corresponding to an acoustic pressure of around 20 micros pascals, while pressures exceeding 2 Pa (120 dB SPL) can cause a painful sensation, and ear damage. The intensity perceived is called loudness.

1.3.2.2 Frequency limit

The range of sound frequencies perceptible to the human ear is from 20 Hz to around 20,000 Hz [Nor,2004]. Its frequency limitation is around 16 kHz to 20 kHz, depending on the individual.

1.3.2.3Boundary in direction

Auditory localization is the process of combining sounds arriving at both ears to discern the direction of origin of sound..

1.3.2.4The time limit

Auditory perception takes time. Sounds lasting less than a tenth of a second are poorly determined in terms of their other three characteristics. Perception becomes more refined as the duration increases, until it reaches its greatest finesse at around half a second [Wik,2015].

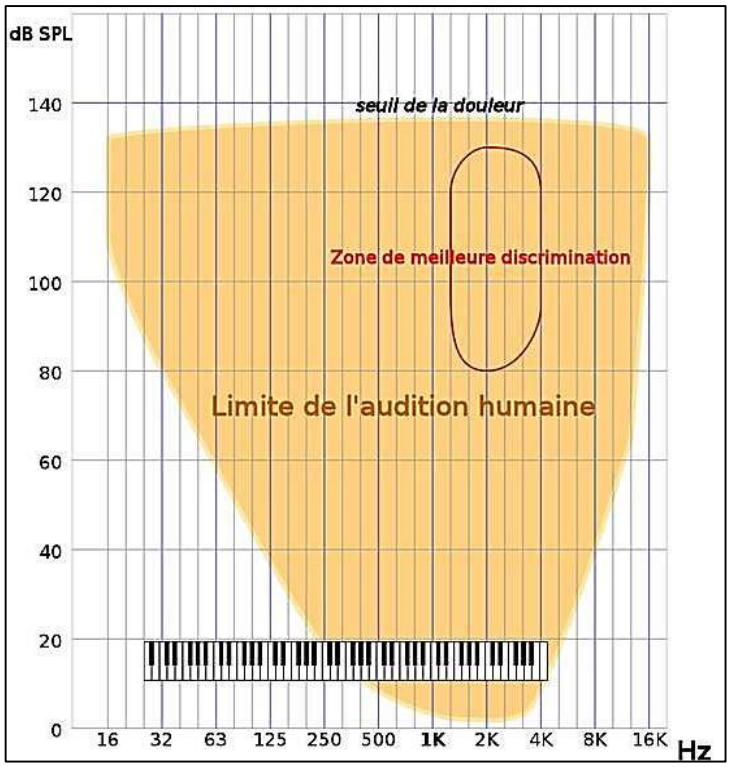


Figure 1.6: Limit of human hearing

Figure 1.6 shows the frequency and intensity limits of audible sounds. The zone of maximum acuity is the one where the distinction between sounds close in pitch or intensity is the finest. The human ear has a maximum sensitivity to 500 Hz-10KHz, outside which sounds must be more intense to be perceived.

1.3.3 Modeling scales

Bell's laboratories made a major contribution to auditory research, demonstrating the existence of critical bands in the response of the cochlea, which are essential for understanding many phenomena such as intensity perception, pitch or timbre[Vel,2006].

In the previous section (frequency analysis of the ear), we saw that each hair cell responds to a frequency that depends on its position throughout the cochlea. The cochlea acts as if it were made up of a bank of filters whose critical bands overlap, and whose central frequencies are continuously graduated. Among the proposed modalizations of these critical bands are the Bark scale and the Mel bank. The latter is more efficient for spectral analysis.

1.3.3.1 Bark scales

Bark is a psychoacoustic tool that creates the x-axis of the loudness curve specific to a complex sound when the 24 Barks are placed end-to-end, corresponding to the 24 critical bands of hearing [Nor,2004].

The ear's critical band is continuous, and any audible frequency is always centred on it. Researchers used the Bark scale to process spectral energy in a way that was as close to the ear as possible.

Bark frequency can be expressed in terms of linear frequency (in Hz) by [Vel,2006] :

$$\boldsymbol{B(f)} = \frac{\mathbf{26.81}\boldsymbol{f}}{\mathbf{1960}+\boldsymbol{f}} - \mathbf{0.53} \text{ (Expressed in Barks)} \quad (1.1)$$

The approximation is accompanied by two correction functions, one for low frequencies to ensure the curve approaches the standard values as closely as possible, and the other for high frequencies to rectify incorrect values for B(f) > 20.1 :

$$\boldsymbol{B`} = \begin{cases} \boldsymbol{B} + \mathbf{0.15}(\mathbf{2} - \boldsymbol{b}) \; \boldsymbol{Si} \; \boldsymbol{B(f)} < 2 \\ \boldsymbol{et} \\ \boldsymbol{B}+.\mathbf{22}(\boldsymbol{B} - \mathbf{20.1}) \; \boldsymbol{Si} \; \boldsymbol{B(f)} > 20.1 \end{cases} \quad (1.2)$$

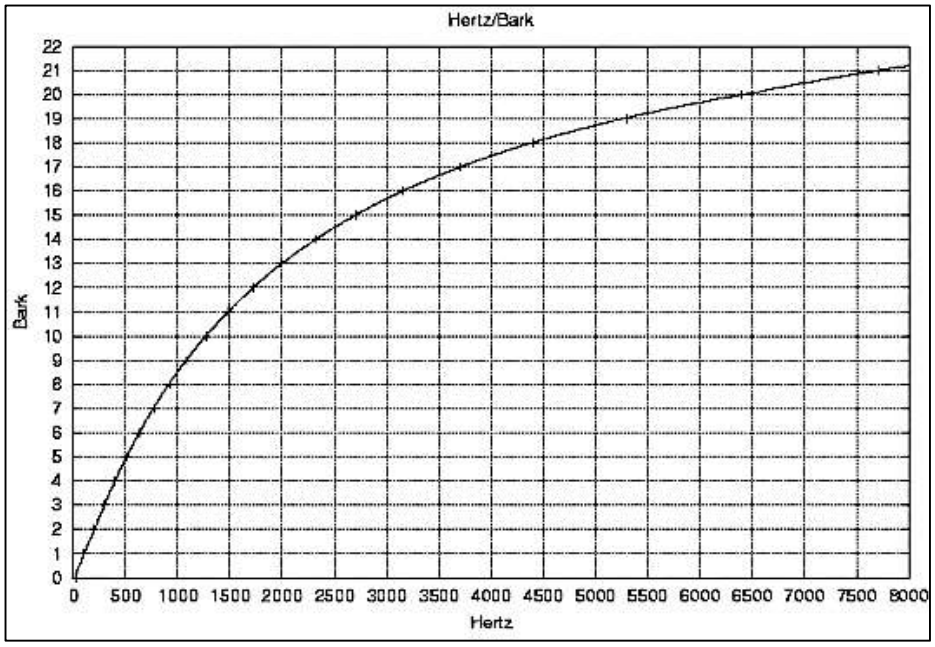


Figure 1.7: Bark scale versus frequency .

1.3.3.2Mel scales

The Mel scale, proposed in 1937 by Stevens et al., is a perceptual scale of pitches judged by listeners to be at equal distance from each other. It is defined by assigning a perceptual pitch of 1000 Mels to a 1000 Hz tone, 40 dB above the listener's threshold. This scale is based on pitch comparisons, meaning that 4 octaves on the Hertz scale above 500Hz are judged to comprise 2 octaves on the Mel scale, this is because increasing intervals are judged to produce equal step increments [Wik,2015] :

$$\boldsymbol{m = 1127ln\left(1+\frac{f}{700}\right) = 2595log\left(1+\frac{f}{700}\right)\ (en\ Mel)} \qquad (1.3)$$

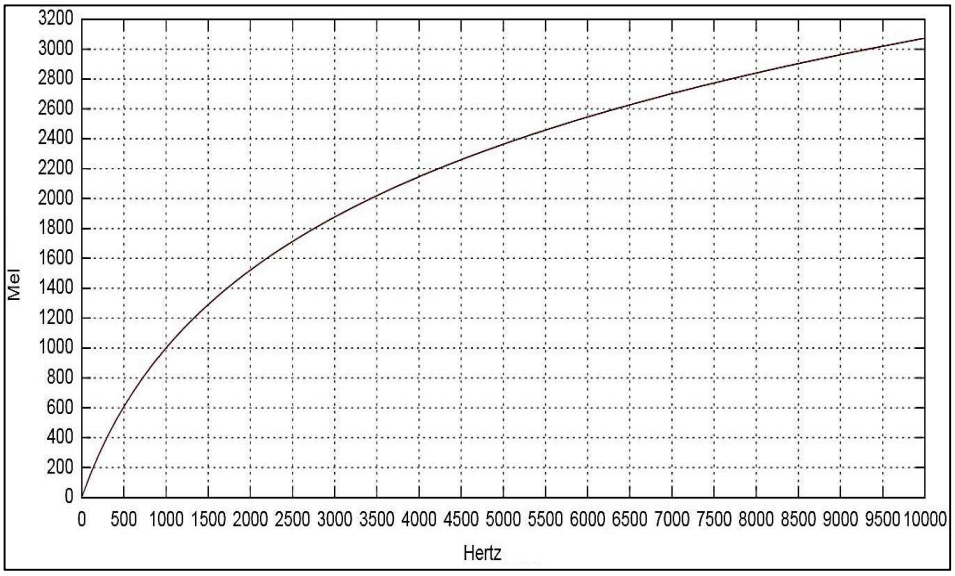


Figure 1.8 : Mel scales

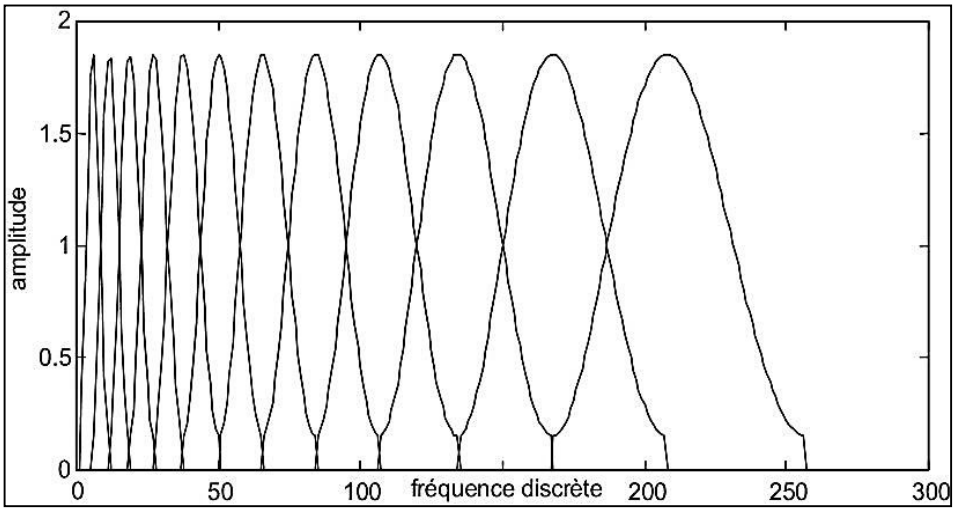


Figure 1.9: Mel scale approximation (Hamming window filter bench)

Mel scaling analysis is used to better model the sensitivity of segments of the human ear, and is widely used in speech recognition systems.

1.4 Automatic Speaker Recognition (ASR)

The ASR is the ability of a machine to identify a speaker based uniquely on the content of the speaker's voice.

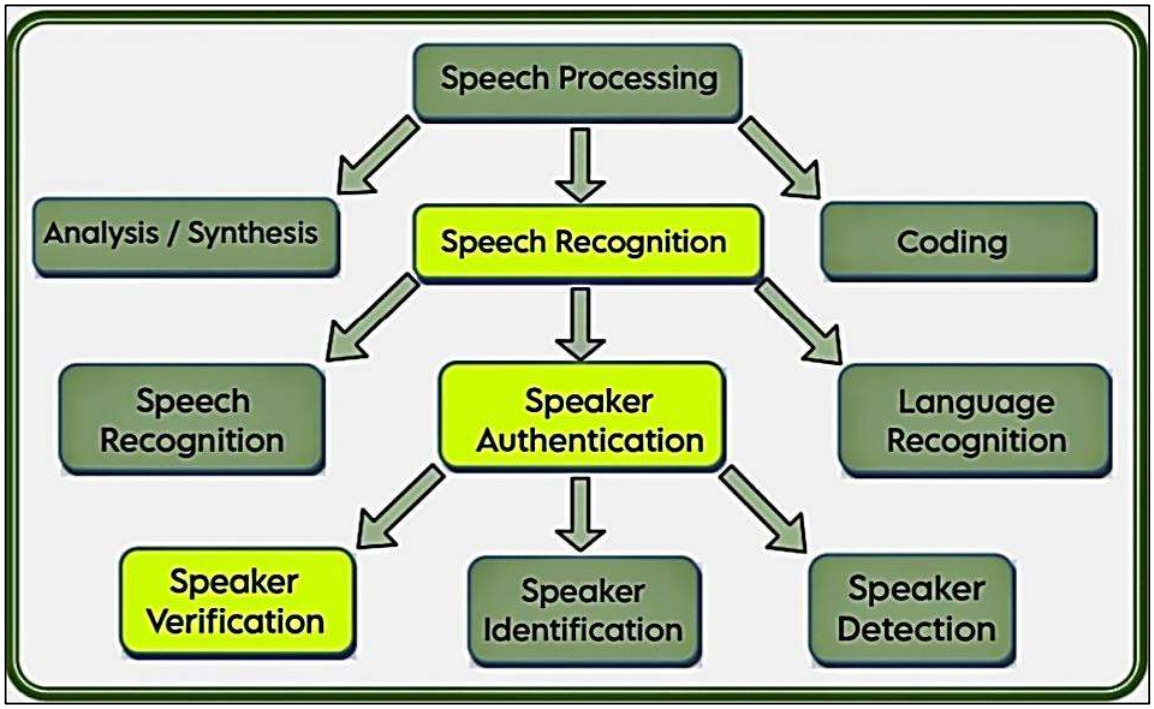


Figure 1.10: Speech signal processing

Automated speaker recognition is frequently advocated as a way to address security issues by utilizing the voiceprint of the speaker :

- Controlling access to secure places.
- Access to information with privileges.
- Telephone-based verification of banking transactions.
- Personalized device use (smartphone),Find potential suspects.
- Determining the speaker's emotional state (stress, weariness, etc.).
- Names of journalists are automatically shown.
- Automated database search for a certain audio recording.

1.4.1 Different ASR tasks

Automatic Speaker Verification (ASV) and Automatic Speaker Identification (ASId) are the most pioneer tasks of ASR. Later, new applications of ASR and other tasks such as Automatic Speaker Tracking (AST), Automatic Speaker Segmentation (ASS) and Automatic Speaker Indexing (ASInd) are emerging. We will mention the different tasks of the ASR with detailed explanations to distinguish the difference between them.

1.4.1.1 Automatic Speaker Identification (ASI)

Automatic Speaker Identification (ASI) is the process of determining, from a population of known speakers, the person who has spoken a given message. Schematically (Figure 1.12), a speech sequence is given as input to the ASI system, and for each speaker known to the system, the speech sequence is compared with a reference characteristic of the speaker: the identity of the speaker whose reference is closest to the speech sequence is given as output by the ASI system. Two modes are available in an ASI system :

- The ASI system is used to determine the most likely identity among known users in a closed set. It should only be used in an environment where all individuals are known.
- The ASI system must decide on the reliability of its judgment by accepting or rejecting the identity it has found in an open set. The performance of ASI systems generally deteriorates as the speaker population increases [Oua,2009].

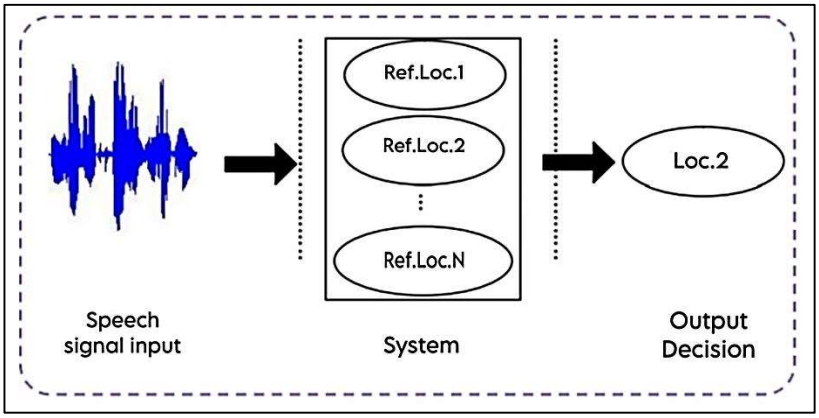


Figure 1.11: Basic principle of the Automatic Speaker Identification task

1.4.1.2Automatic Speaker Detection (ASD)

Automatic Speaker Detection (ASD) is a variant of ASV that considers an audio stream composed of speech sequences produced by several speakers. The task of detection is to determine if a given speaker is involved in the audio document. The ASD is used to detect a single-speaker audio stream, which is the same as the verification task :

- Detection of messages from suspicious persons in telephone conversations [Oua,2009] .
- Detection of whether or not a given speaker is involved in a collection of audio recorded in a TV or radio program.
- Search for newscasters in order to extract themes from their speech or topics covered in the news.

1.4.1.3 Automatic Speaker Indexing (ASInd)

The Speaker Indexing task is a more recent development than automatic speaker identification and verification, involving the targeting of speaker interventions in an audio stream. The only average input to an indexing system is the audio document to be indexed, and the system is based on a blind speaker segmentation phase followed by a grouping phase. The output of such a system is the answer to two questions: "who is speaking?" and "when?" The indexing principle is based on a blind speaker segmentation phase, followed by a grouping phase :

- Speaker 1 spoke from t_0 to t_1 and t_5 to t_6.
- Speaker 2 spoke from time t_1 to t_2.
- Speaker 3 spoke from time t_2 to t_3.

Speaker indexing is used to process audio databases (e.g., to search for sequences of TV or radio programs by tracking the speaker, to estimate the speaking time of each speaker in a debate). The following figure illustrates the input and output parameters for a speaker-based indexing system.

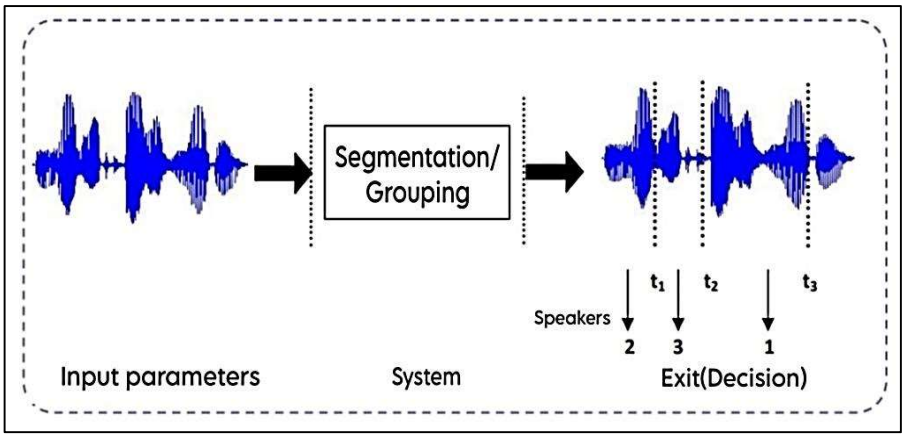


Figure 1.12: Indexing system by speakers of an audio stream

1.4.1.4 Automatic Speaker Tracking (SAT)

The speaker tracking task is a simplified version of the speaker indexing of an audio stream. It involves determining the interventions of one or more speakers, called target speakers, in an audio stream. The speaker-tracking system needs to know which speakers are present in the document to be indexed and has a characteristic reference for each speaker[Oua,2009]. The following figure shows the principle of a SAT system.

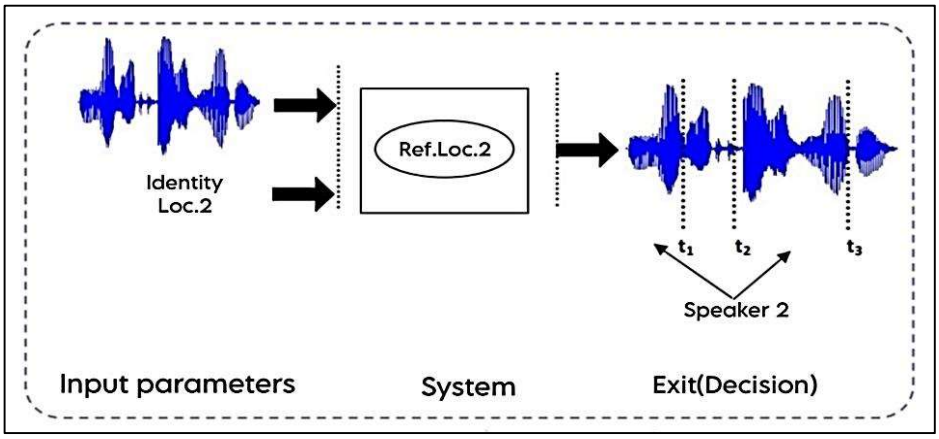


Figure 1.13: Speaker tracking system

1.4.1.5 Automatic Speaker Verification (ASV)

The ASV is used when we want to decide whether the identity declared by a speaker is compatible with the identity recorded in the database by analyzing his voice. In this type of application, the identity and the voice message are both the inputs to the ASV system. The identity, assumed to be known, designates the characteristic reference of a speaker.

A measure of similarity is calculated between this reference and the voice message, and compared with a decision threshold. If the similarity is below the threshold, the individual is accepted, otherwise, the individual is considered an imposter, and will be rejected as shown in the following figure [Oua,2009].

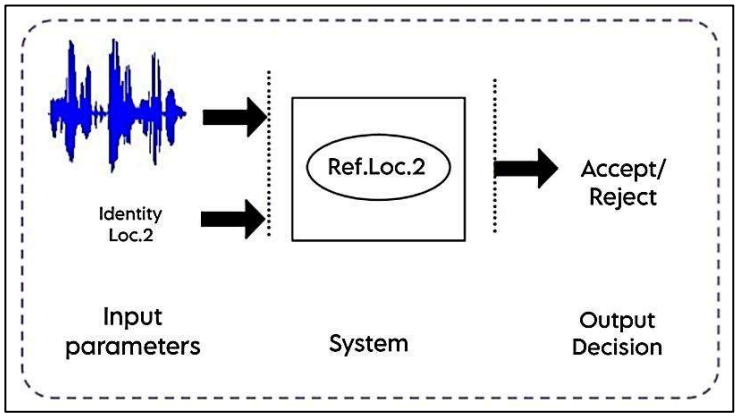


Figure 1.14: Basic principle of the automatic speaker verification task

1.4.2 General architecture of an ASR system

The general architecture of an ASR system is divided into different processing modules. Most standard ASR approaches use a similar structure:

- Parameterization module : Extracts elements for speaker discrimination;
- Creation of references : Creates a speaker reference from the previously extracted parameters.
- Test module : Compares the reference (verification task) or references (identification task) with the test signal;
- Decision module : Renders a decision based on the previous result and takes into account various elements such as the desired level of security.

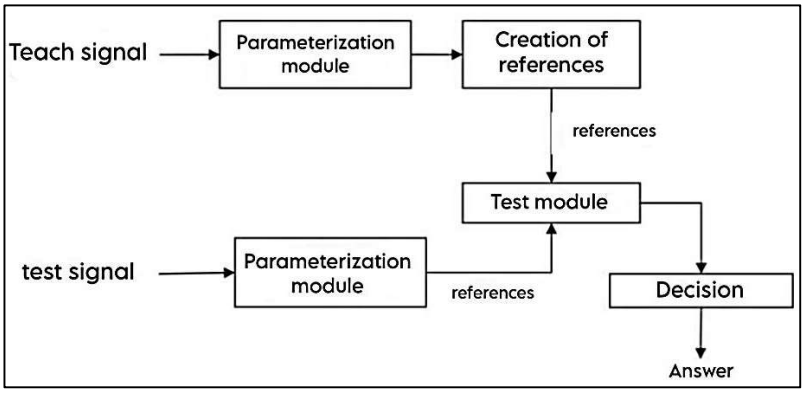


Figure 1.15: Structure of an ASR system

1.4.2.1 Parameterization module

A) Acoustic parameterization

Since the speech signal is redundant and complex, we need to apply a parameterization to exploit it. This process extracts the relevant information for each speaker. The first step in acoustic parametrization is to decompose the regularly-paced speech signal into signal frames of a length generally varying from 20 to 31.5 ms. The frames of a speech signal are processed to produce acoustic parameter vectors, which are divided into three main classes: spectral analysis, prosodic and dynamic [Say,2003].

B) Spectral analysis parameters

Spectral analysis is the most widely used analysis in ASR. The resulting parameters are generally representative of the physical characteristics of each individual's vocal tract. Multiple parameters have been used in ASR. For a quick description and comparison of different parameters. We list here the most relevant in ASR [Say,2003].

- Coefficients derived from linear prediction analysis: LPCC (Linear Predictive Cepstral Coefficients) or LPC (Linear Predictive Coefficients);
- Spectral coefficients derived from filter bank analysis: LFSC(Linear Frequency Spectral Coefficients) or MFSC (Mel Frequency Spectral Coefficients).
- Cepstral coefficients from filter bank analysis: LFCC (Linear Frequency Ceptral Coefficients) or MFCC (Mel Frequency Cepstral Coefficients).

C) Prosodic parameters

In many areas of vocal signal analysis and processing, prosodic parameters are defined as follows:

- Fundamental frequency (vocal cord vibration).
- Voice intensity (or energy).
- Duration of syllable segments.

These prosodic parameters take on particular importance in providing synthesis systems with improved intelligibility, while enabling recognition systems to perform analysis or segmentation by phonetic unit order.

The variation over time of these parameters (intonation) conveys various clues-characteristics of the individual, whether in terms of physical state (age, sex, physiology), emotional state or regional accent [Ezz,2002].

D) Dynamic parameters

The dynamic information conveyed by the speech signal is a potential source of information about the speaker's characteristic, which is still poorly exploited by ASR. The most widely used dynamic parameters are coefficients derived from instantaneous parameter vectors, known as Delta coefficients (first derivative) and Delta coefficients (second derivative). Other parameterizations have been proposed to exploit dynamic signal information, such as the use of Time Frequency Principal Components (TFPC) [Say,2003]. A brief description of these different approaches is given in [Fre,2000].

1.4.2.2 Decision in ASR system

A) Decision in an ASI system

In identification, a test signal is compared with all the known speaker references in the system, and the performance measure of an ASI system is generally given in terms of the correct identification rate. This rate is calculated by the following formula :

$$\boldsymbol{Identification\ rate} = \frac{\boldsymbol{number\ of\ correct\ identification}}{\boldsymbol{nombre\ of\ tests}} \qquad (1.4)$$

The identification errors in an open set are :

- Misrecognition rate (number of times the system confuses one database speaker with another).
- False rejection rate (number of times the system rejects a user even though the user is supposed to be known to the system).
- False acceptance rate (number of times the system accepts a user who is foreign to the database).

In a closed set, the calculated error is represented by the misrecognition rate.

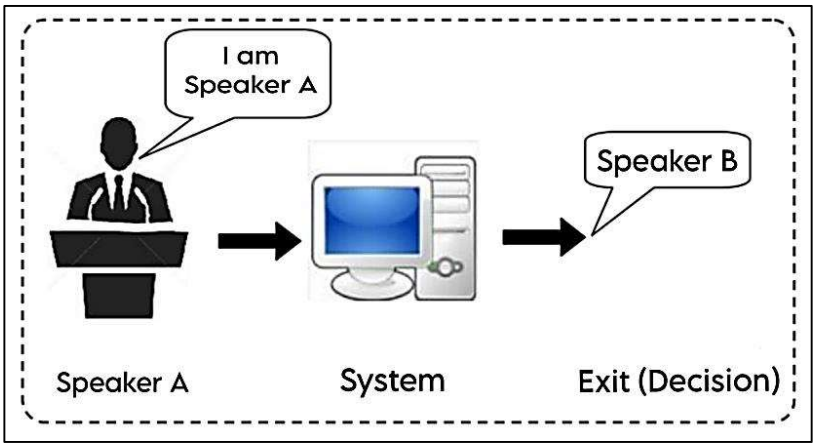


Figure 1.16 :Misrecognition error

B) Decision in ASV systems

The decision process in verification involves comparing the similarity measure between a composite test signal and a client model with a decision threshold. If the measure is greater than the threshold, the speaker is considered a client and accepted. The most important idea is to measure the performance of a ASV system, which has been proposed in the literature[Say,2003].

Two types of error characterize ASV systems: the false acceptance error (acceptance of an impostor), and the false rejection error (rejection of a customer). These errors lead to false acceptance error rates $\boldsymbol{\rho(FA)}$ and false rejection error rates $\boldsymbol{\rho(FR)}$ calculated for a given decision threshold respectively :

$$\boldsymbol{\rho(FA)} = \frac{\boldsymbol{number\ of\ tests\ that\ led\ to\ a\ false\ acceptance}}{\boldsymbol{number\ of\ imposter\ tests}} \qquad (1.5)$$

and ;

$$\boldsymbol{\rho(FR)} = \frac{\boldsymbol{number\ of\ tests\ that\ led\ to\ a\ false\ rejection}}{\boldsymbol{number\ of\ customer\ tests}} \qquad (1.6)$$

The various verification errors made by the system are illustrated in the figure below:

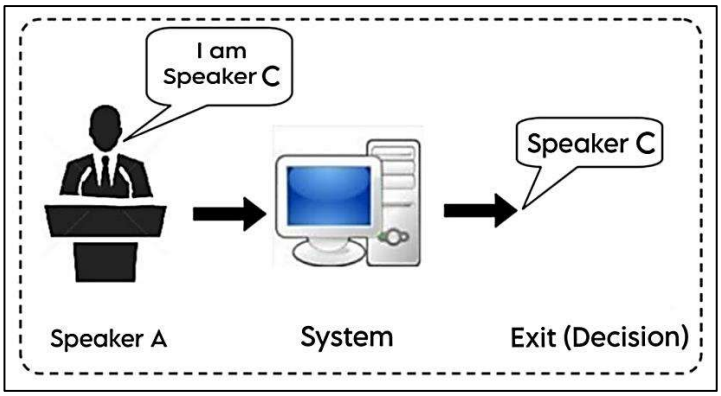


Figure 1.17: False Acceptance Error

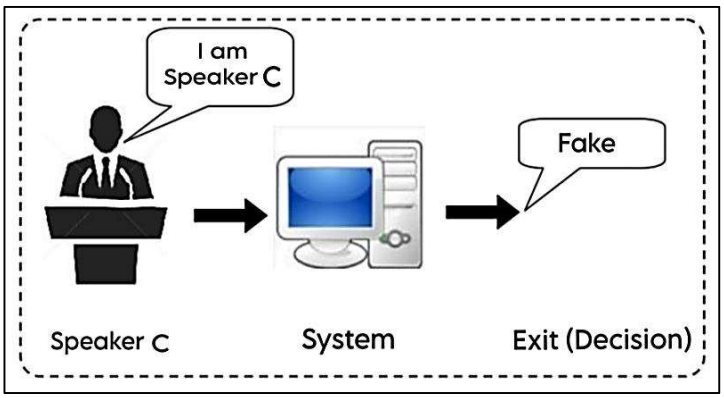


Figure 1.18: The False Rejection Error

Conclusion

The speech signal contains several types of information, such as frequency, energy and cadence. Speech recognition systems are therefore left with the task of extracting the necessary information. In this chapter, we have studied the biometrics and the phenomenon of human speech production and perception.

Also, we exposed Automatic Speech Recognition (ASR) starting with its state of the art and going through its different tasks as well as the general structure of an ASR system. In this dissertation we have proposed the use of second-order static measurements with different distances to improve the performance of our system. The proposed methods for improving speaker verification will be detailed in the next chapter.

Chapter-2

Proposed Methods for Speaker Verification

Chapter-2
Proposed Methods for Speaker Verification

Introduction

In this chapter, we are going to present the proposed methods for the speaker verification. First, we are going to define the method used for features extraction, which is the Mel Frequency Cepstral Coefficients (MFCC), and then explain the different algorithms that have been implemented to perform the speaker verification task which are; Fast Fourier Transform (FFT)

2.1 Proposed method for speaker verification

2.1.1 Features extraction using the MFCCs

The Mel Frequency Cepstrum Coefficients (MFCC) method is based on the fact that human hearing is limited to frequencies up to 1kHz. The foundation of MFCC is the difference in frequencies that the human ear can detect. The MEL scale (see figure 1.8), which is based on how observers perceive pitches at evenly spaced intervals, is used to convey the signal. This scale makes use of a filter with linear spacing for frequencies under 1kHz and logarithmic spacing for frequencies over 1kHz .

There are tones of various frequencies that make up the voice signal. A subjective pitch on the "Mel" scale is calculated for each tone that has an actual Frequency (***f***), measured in Hertz. Below 1000 Hz, the Mel-frequency scale has a linear frequency spacing, while beyond 1000 Hz, it has a logarithmic spacing. For comparison, 1000 mels is the pitch of a tone at 1kHz, 40 dB over the perceptual hearing threshold. Because of this, we can use the formula shown below to get the mels for a given frequency, ***f*** in Hz[Jrj,2000] :

$$mel(f)= 2595*\log10(1+f/700) \quad (2.1)$$

Applying a filter bank with a filter for each desired mel frequency component is one method of replicating the subjective spectrum. The filter bank has a triangle bandpass frequency response, and a fixed mel-frequency interval controls both the spacing and the bandwidth.

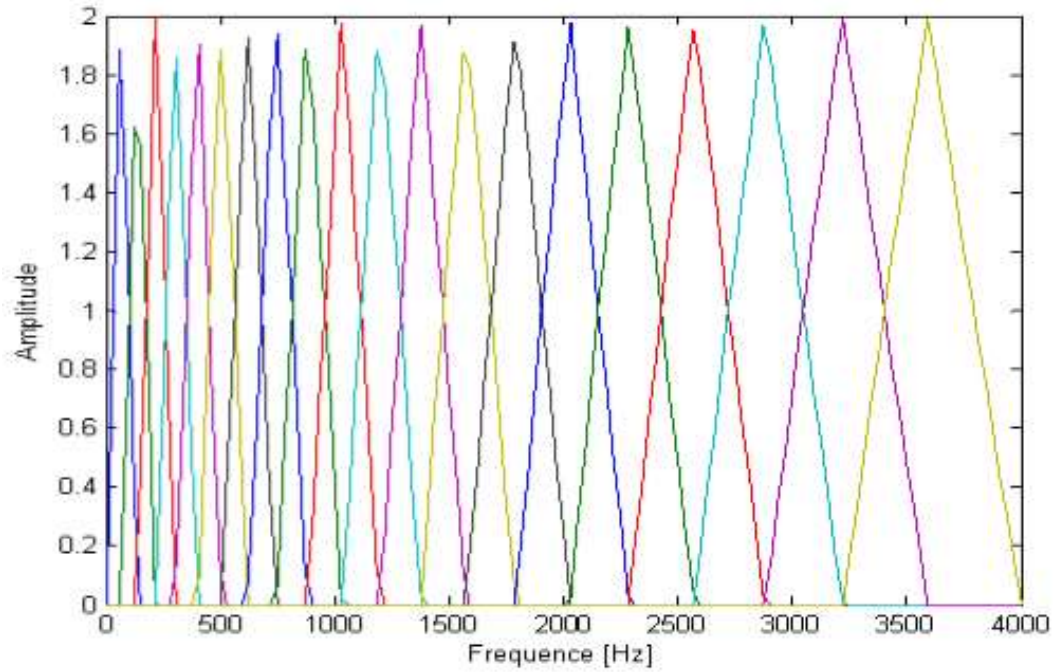


Figure 2.1 : Mel Filter Bank from a speaker saying his name

In contrast to the linearly spaced frequency bands obtained directly from the FFT or Discrete Cosine Transform (DCT), the frequency bands in the Mel-frequency cepstrum are located logarithmically on the mel scale, which more closely approximates the response of the human auditory system. Better data processing may be possible as a result, for instance in audio compression, but not like a sonogram. Since MFCCs lack an outer ear model, they are unable to appropriately reflect perceived loudness. The following is a typical derivation of MFCCs:

❶ Perform the Fourier Transform on a windowed signal extract,

❷ Using triangular overlapping windows, map the log amplitudes of the spectrum acquired in step 1 onto the Mel scale.

❸ Transform the list of Mel log-amplitudes using the discrete cosine transform as if it were a signal.

❹ The amplitudes of the resultant spectrum are the MFCCs.

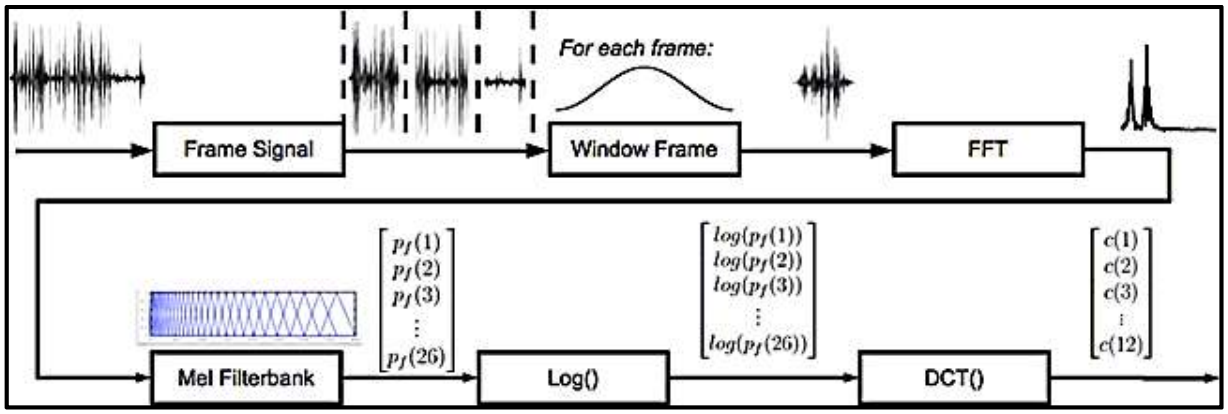


Figure 2.2: Mel frequency cepstrum coefficient (MFCC) calculation.

2.1.2 Calculating procedure of the MFCCs

The MFCCs are calculated by first performing a Fourier Transform on a signal to convert it from the time domain to the frequency domain. The frequency domain representation is then filtered with a Mel filterbank, which is a set of overlapping frequency bands.

The output of the filterbank is then converted to the logarithmic domain, and the DCT is applied. The result is a set of coefficients that represent the spectral envelope of the signal.

The DCT is used to reduce the dimensionality of the data, since it is not necessary to keep all of the coefficients. The resulting coefficients are the MFCCs, which can be used for speech recognition and audio analysis.

2.1.3 Applications of the MFCCs

The MFCCs are used in several speech recognition systems, as they capture the spectral envelope of a sound. They are also used in audio analysis to identify the characteristics of a sound. For example, they can be used to identify the type of instrument being played, or to identify the speaker of a voice. The MFCCs are also used in music analysis, to identify patterns in music. In addition, they can be used to detect the key and tempo of a piece of music, as well as to identify similar pieces of music. They can also be used to create automatic music generation systems.

2.1.4 Advantages of MFCC

The MFCCs are a useful feature for speech recognition and audio analysis, as they capture the spectral envelope of a sound. They are also relatively easy to calculate, and require minimal storage space. Furthermore, they are robust to noise and distortion, which makes them suitable for use in noisy environments. These features are relatively insensitive to shifts in frequency and amplitude, which makes them suitable for use in speech and audio applications. This is because they capture the spectral envelope of a sound, rather than the exact frequencies of a sound.

2.1.5 Acquisition and restitution of digital sound

2.1.5.1 Acquisition of digital sound

The acquisition of a digital sound is done in three stages: sampling (or taking a sample), blocking (or holding the sample during the analog-digital conversion) and finally the analog-digital conversion (see figure 2.3).

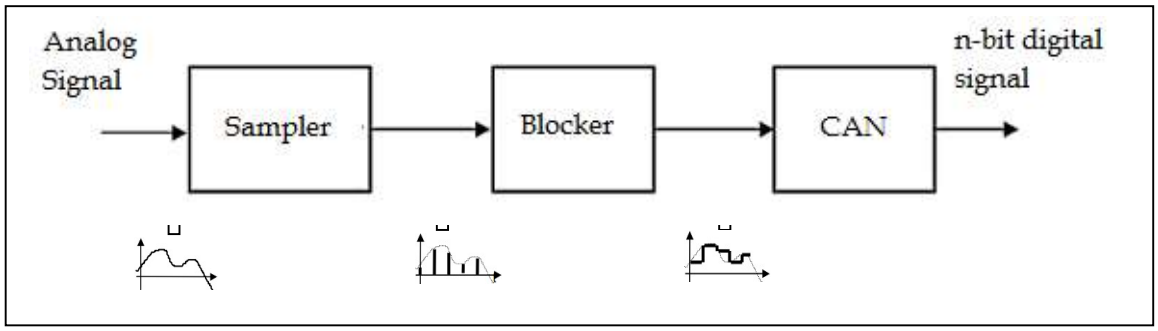


Figure 2.3: Speech signal acquisition chain

Capturing a speech signal consists of quantifying and sampling the voltage levels present at the terminals of an analog microphone. We therefore go from a signal comprising an infinity of values (continuous function) to a signal whose values are finite and spaced out in time.

The result is therefore a vector comprising the different voltage levels taken by the sound signal over time. This is not done without a loss of information from the original signal, and we then encounter so-called 'aliasing' phenomena when the interval between two level measurements is not adapted to the recorded sound.

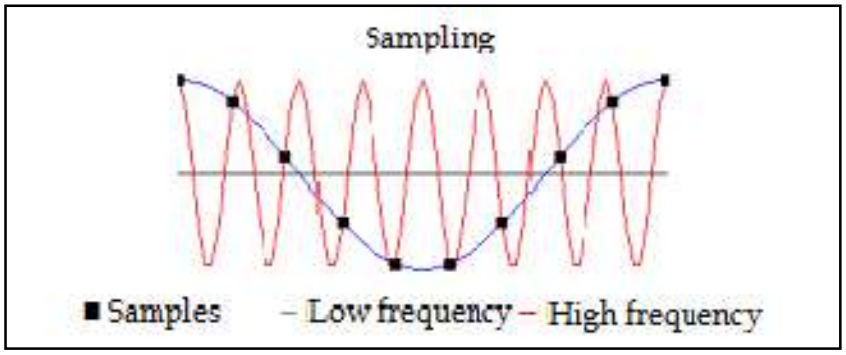


Figure 2.4: Aliasing phenomenon

Figure 2.5 above illustrates the phenomenon of aliasing. In this case, the sampling frequency is not large enough for the signal to be sampled. Voltage level measurements are not taken frequently enough to reconstruct the true signal (in red). This results in the blue signal which has a lower frequency than the red signal. Hence the name 'aliasing', because one frequency is impersonating another.

This loss of information can be minimized and thus avoid the aliasing phenomenon if we consider the frequencies we want to record. Thus, the Shannon-Nyquist theorem specifies that to record a signal in such a way as to be able to reproduce it faithfully, the sampling frequency must be at least twice as large as the largest frequency contained in the signal. This is expressed by the equation:

$$F_S \geq 2 * F_{max} \quad (2.2)$$

With :

$\boldsymbol{F_S}$: Signal sampling frequency, and

$\boldsymbol{F_{max}}$: Maximum frequency contained in the signal to be captured.

There will always be a loss of information, but we will have kept enough to guess the shape of the signal only from the values obtained. The human ear is able to hear sounds with frequencies between approximately 20 and 20,000 Hz. The sampling frequency for sound recordings is therefore most often fixed at 44100 Hz. This allows the digital sound to be reproduced without the human ear being able to distinguish any difference.

In order to avoid aliasing, it is necessary to use an anti-aliasing filter that will have the role of eliminating all the frequencies higher than the sampling frequency.

2.1.5.2Restitution of a digital sound

The restitution of a digital sound ***yn*** amounts to converting the values which constitute it into analog voltages ***y(t)*** to send it to a loudspeaker for example, with the same frequency used for the acquisition. It can be done in three ways; By using a hold circuit, by using a cardinal sine compensating filter or by oversampling.

In the general case, the reproduction of a digital sound is done in three stages; the first is the digital-analog conversion (which outputs the samples y ***(nTe)***), the second is ensured by a blocker (which maintains the sample ***y (nTe)*** at the output until the arrival of the 'sample y ***((n+1)Te)*** and the last stage is a low-pass filter (to smooth the signal in steps of stairs thus obtained) (see figure 2.5).

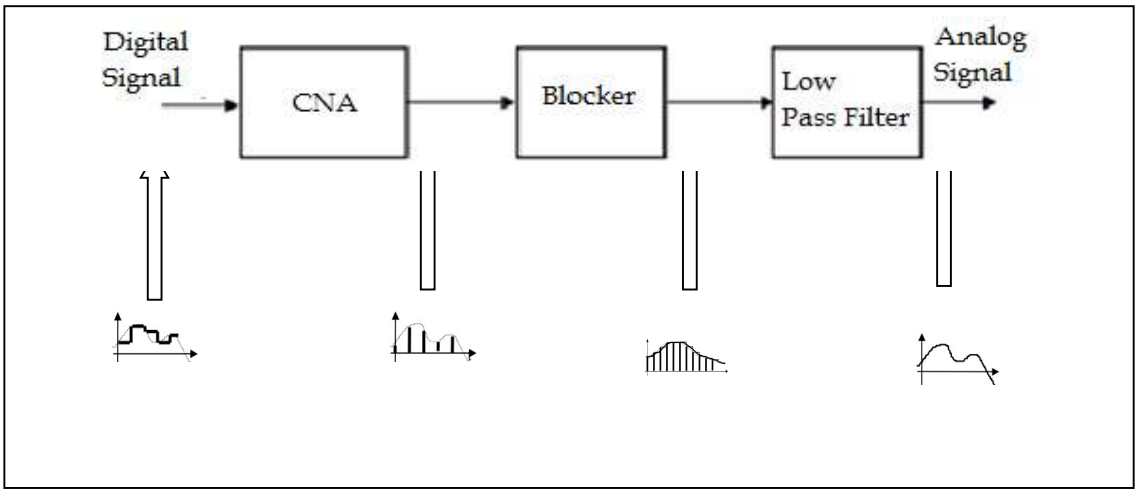


Figure 2.5: General structure of a digital sound reproduction system

2.1.6 Fast Fourier Transform

2.1.6.1What is Fast Fourier Transform?

FFT is a computational algorithm that converts a time-domain signal into its frequency-domain representation. It differs from the standard Fourier Transform by using a divide-and-conquer approach to reduce the number of computations required. This makes it much faster and more efficient than the standard method.

By breaking down the signal into smaller sub-signals and applying the Fourier Transform to each one separately, FFT can quickly compute the frequency spectrum of the entire signal. This makes it an essential tool for analyzing and manipulating signals in real-time applications.

2.1.6.2The Mathematics Behind Fast Fourier Transform

The Fast Fourier Transform (FFT) is based on the mathematical principle that any periodic function can be represented as a sum of sine and cosine functions of different frequencies. By decomposing the signal into these sinusoidal components, FFT can determine the amplitude and phase of each frequency component in the signal.

The algorithm FFT uses a technique called the Cooley-Tukey algorithm to recursively break down the signal into smaller sub-signals and compute their frequency spectra. This process is repeated until the signal can no longer be divided, resulting in the final frequency-domain representation of the signal.

2.1.6.3Applications of Fast Fourier Transform

The FFT has numerous applications in signal processing, including audio and image compression, filtering, and analysis. In audio processing, FFT is used to analyse sound waves and extract features such as pitch and timbre.

It is also used in image processing to identify patterns and structures in images. In addition, FFT is used in telecommunications to encode and decode signals for transmission over long distances. It is also used in geophysics to analyse seismic data and detect underground structures.

2.1.6.4Limitations of Fast Fourier Transform

Despite its efficiency and versatility, FFT has some limitations. One of the main limitations is its inability to handle non-stationary signals, which are signals that change over time. FFT assumes that the signal is stationary, meaning that it remains constant over time.

To overcome this limitation, alternative methods such as Wavelet Transform and Short-Time Fourier Transform have been developed. These methods allow us to analyse non-stationary signals by breaking them down into smaller sub-signals and computing their frequency spectra over time.

2.1.6.5Describe FFT analysis

One of the most popular methods for doing signal analysis across several application fields is FFT analysis. Signals are converted from the time domain to the frequency domain using FFT. Fast Fourier Transform is referred to as FFT.

Compared to looking at time domain data, FFT analysis allows for a more deeper investigation of a variety of signal properties. The signal characteristics are characterized by distinct frequency components in the frequency domain, but in the time domain, they are expressed by a single waveform that contains the total of all characteristics.

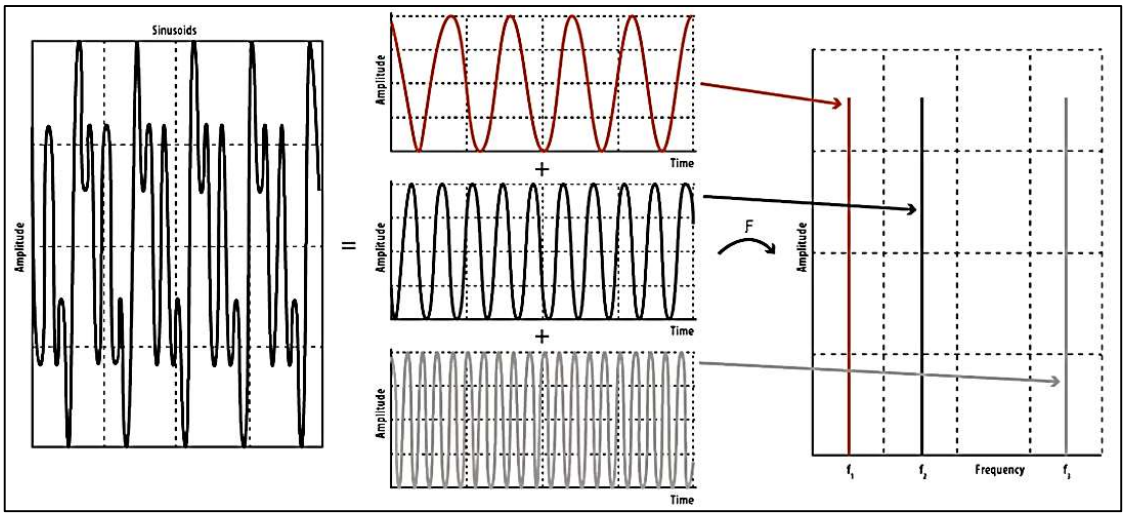


Figure 2.6: FFT analysis.

Conclusion

In this chapter, we have detailed the processes of the Mel Frequency Cepstral Coefficients, the Fast Fourier Transform. We have also, presented the different algorithms that have been implemented to perform a task Verification the speaker. In

the following chapter, we are going to explain the conducted experiments and debate the obtained results.

Chapter-3

Conducted Experiments and Obtained Results

Chapter-3
Conducted Experiments and Obtained Results

Introduction

In this chapter, we will present the experimental protocols that allowed us to achieve the obtained results. We will start by presenting the database that we have conceived and used in all experiments of the speaker verification task to define how we built the models of each speaker in the database. Before defining the calculation method used to count the error and good recognition rates, we will present the principals of the different algorithms proposed for speaker verification. In this work, we have proposed algorithm to verify the speaker identity: the FFT (Fast Fourier Transform). The MFCC (Mel Frequency Cepstral Coefficients) is also proposed as a feature extraction tool.

3.1 Conceived Database

The speech input is captured at a sample rate greater than 10000 Hz, to reduce the impacts of aliasing during the analog-to-digital conversion. These sampled signals can record all frequencies up to 5 kHz, which encompasses the majority of human-generated sound energy. All the Automatic Recognition System have two different phases. The first one is called training phase (Enrolment session) whereas the second one is referred to as testing phase (operation session).

In the training phase, each speaker has to record 7 samples of their speech or you can read audio recordings of people we already have.so that the system can build a reference model that speaker. In addition, the Automatic Speaker Verification system require a threshold value to be computed for each speaker. Throughout the testing phase, a new sample is recorded for an unknown speaker and matched with the saved reference models to make a final decision.

3.2 The proposed algorithms for speaker verification

3.2.1 The Mel Frequency Cepstral Coefficients (MFCC) Approach

Usually, the voice input is captured at a sample rate greater than 10000 Hz. To reduce the impacts of aliasing during the analog-to-digital conversion, this sampling frequency was used. These sampled signals can record all frequencies up to 5 kHz, which encompasses the majority of human-generated sound energy. The MFCC processor's primary function is to simulate human ears' activity.

Additionally, it is demonstrated that MFCCs are less vulnerable to the aforementioned fluctuations than the speech waveforms themselves [Khe, 2018].

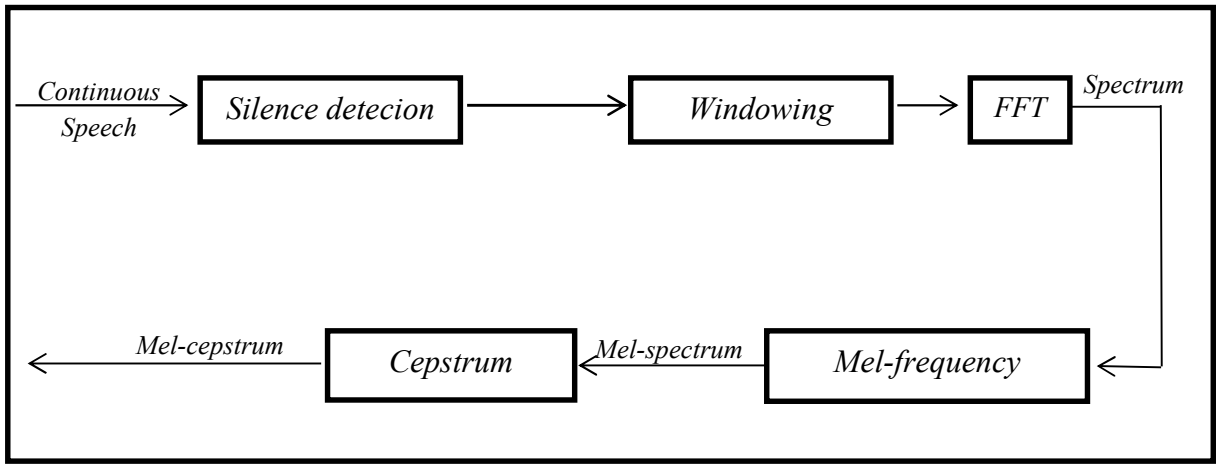


Figure 3.1: Diagram of the MFCC Process

The voice signal was first saved as a 10,000-sample vector. Our investigation revealed that the real spoken voice after removing the static components reached roughly 2500 samples. As a result, we performed the silence detection to isolate the actual spoken speech using a straightforward threshold approach.

Clearly, we aimed to develop a voice-based biometric system that could identify individual words. Our research found that in 2500 samples, practically all isolated words were pronounced. However, we did so after running the voice signal via an MFCC processor, He applied overlapping windows, each with around 250 samples, to extend this to the temporal domain. As a result, each window only has roughly 250 spectrum values when we transfer it to the frequency domain. This suggested that

since the Mel scale is linear up to 1000 Hz, the conversion to the Mel scale would be unnecessary. So, we got rid of the obstruction that was causing Mel to warp.

3.2.2 The Fast Fourier Transform (FFT) Approach

The Fast Fourier Transform (FFT) provides the standard representation of a speech signal in the frequency domain. While Short Fourier Transform (SFT) is able to hold time frequency changes. The drawback of FFT that it is not appropriate for the signals whose frequencies are time varying hence in case of FFT is assumes that the signals are stationary in nature [REY, 1995].

3.2.2.1 Speech signal transformation and processing

The FFT allows working in frequency domain and therefore using the frequency spectrum of the speech signal as a substitute of waveform. Frequency domain provides more information about the speech signal and hence can be more efficient to distinguish between speakers.

For speaker recognition, some techniques use the vice signal acquire directly by the sampling phase and some techniques use transformed form of the speech signal [SIN, 2014].

When a speech signal is represented in time domain, then it gives a little information regarding the speech signal properties. Hence, suitable transformation of speech signal is an essential problem. For this purpose, generally Fourier or Wavelet transform are used [SAM, 2011].

Speech signal processing start by converting the raw speech into a sequence of acoustic feature vectors carrying features of the signal. This is known as pre-processing (i.e. feature extraction is completed here and is called front-end processing).

The Mel Frequency Cepstral Coefficients (MFCC), which are derived from the FFT power spectrum, is the most usable acoustic vectors and MFCC features. The acoustic features are based on the spectral information, which is derived from a minor window segment of a speech signal.

The different Plots have been used to depict the various stages of the simulation. The word "Assalam alaykum" is used as the input continuous voice signal sample for this project.

3.3 Examples of Plots used in the different stages of the simulation

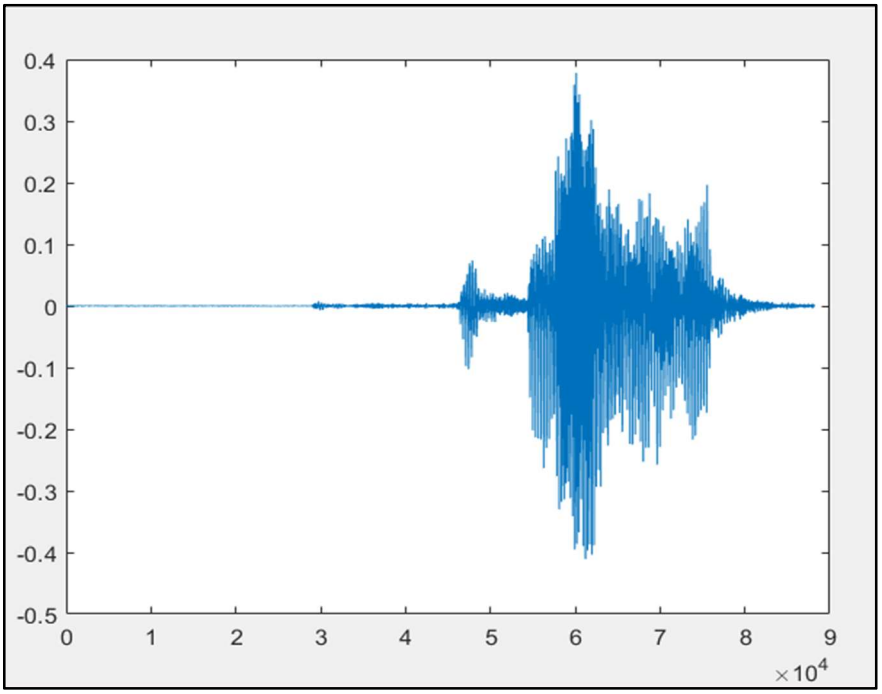
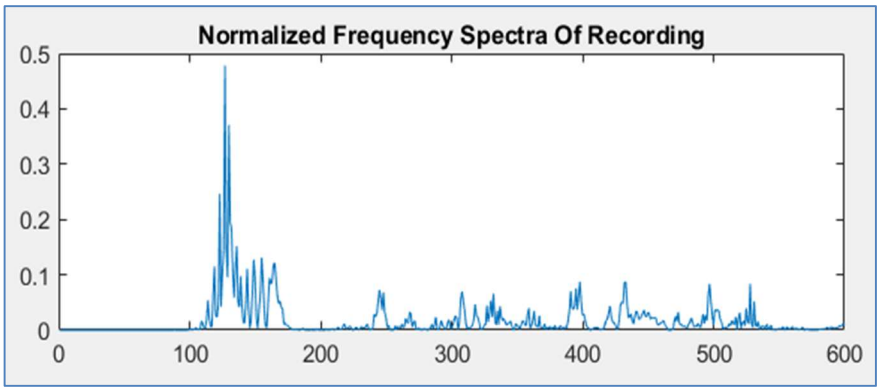
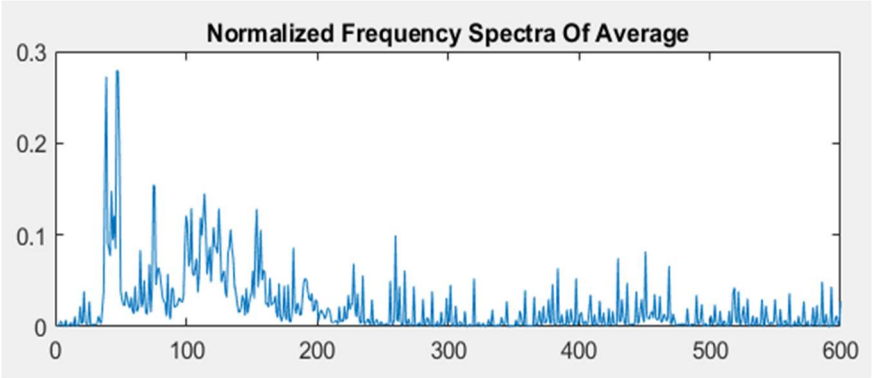


Figure 3.2: A speech signal of a man's voice saying "Assalam alaykum"



(a) Spectra of the Recording (Test)



(b) Spectra of the Average (Training)

Figure 3.3: Normalized frequency spectra of a man's voice saying "Assalam alaykum"

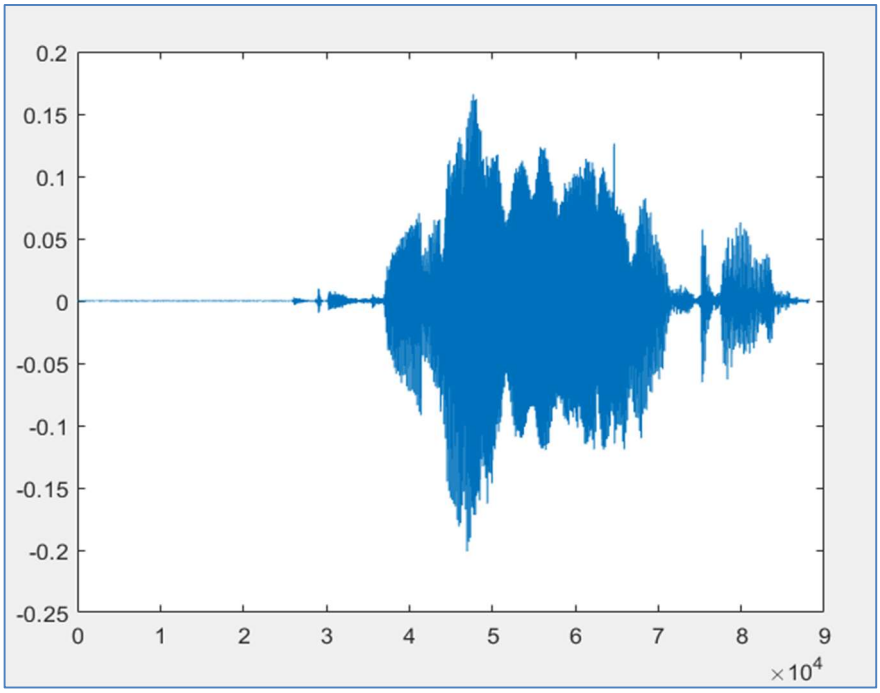
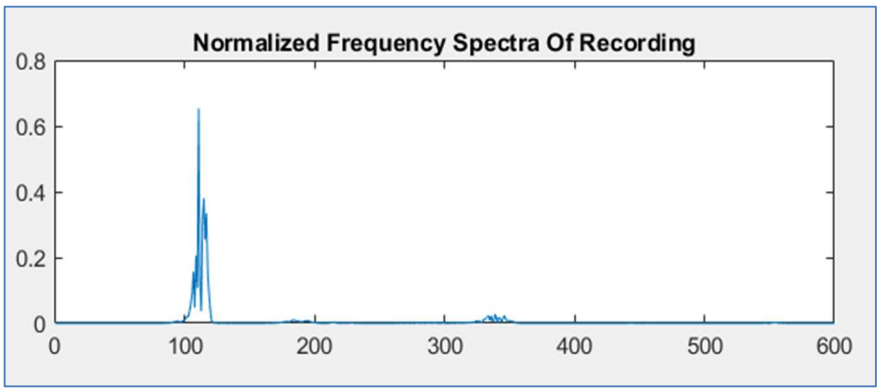
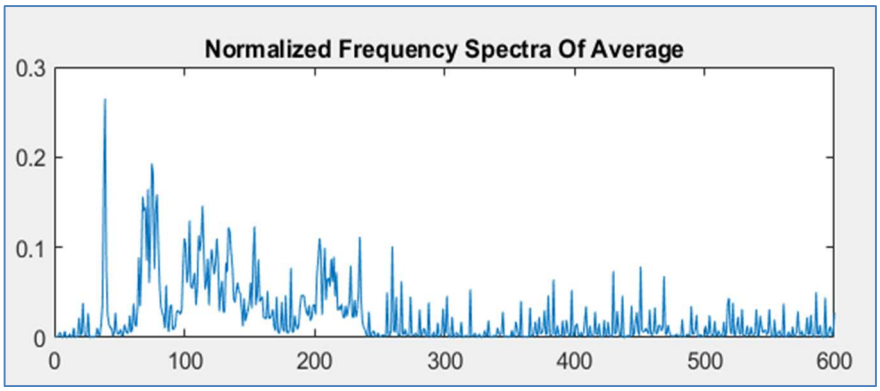


Figure 3.4: A speech signal of a woman's voice saying "Assalam alaykum"



(a) Spectra of the Recording (Test)



(b) Spectra of the Average (Training)

Figure 3.5: Normalized frequency spectra of a man's voice saying "Assalam alaykum"

3.4 Types of errors and performances measures

Two types of errors can be observed in ***ASV*** systems: false acceptance errors and false rejection errors. The first type occurs when the system accepts a speaker who is an impostor, and the second type occurs when the system denies access to a client. Evaluating the quality of an ***ASV*** system depends on several factors. Nevertheless, it is the performance measures (in terms of error rate) that will determine its quality. The most used performance measures are; the FAR rate and the FRR rate.

3.4.1 False Acceptance Rate (FAR)

This rate represents the percentage of people who are not supposed to be recognized by the system but who are accepted. FAR is given by the formula:

$$FAR = \frac{Number\ of\ False\ Acceptances}{Number\ of\ imposters'\ tests} \text{ x } 100$$

3.4.2 False Rejection Rate (FRR)

This rate represents the percentage of people who were supposed to be recognized by the system but who are rejected. FRR is given by the formula:

$$FRR = \frac{Number\ of\ False\ Rejections}{Number\ of\ customers'\ tests} \text{ x } 100$$

3.4.3 Comparison :

A false accept is worse than a false rej In this experiment, MFCC will be used to verify whether a voice belongs to the database or not. After recording the voices or selecting pre-existing voices, a threshold is set for each voice. To verify, we enter a voice or select a pre-recorded voice. After comparing it with the threshold for each person in the database, the result is shown whether it belongs to the database or not. ect. Most organizations would prefer to reject authentic subjects over accepting impostors. FARs (Type II errors) are worse than FRRs (Type I errors). Two is greater than one, which will help you remember that FAR is Type II, which are worse than Type I errors (FRRs).

3.5 Experimental results and interpretations

3.5.1 Experiment 1: Using MFCC for registration and verification

In this experiment, MFCC will be used to verify whether a voice belongs to the database or not. After recording the voices or selecting pre-existing voices, a threshold is set for each voice. To verify, we enter a voice or select a pre-recorded voice. After comparing it with the threshold for each person in the database, the result is shown whether it belongs to the database or not.

Table 3.1: Threshold results for each speaker in the database

	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5
Threshold	180.3514	568.5889	67.8376	18.2846	333.0944
Distance	12.6063	15.958	17.6542	13.5182	19.6583

In the table, we notice that the threshold for each speaker is different, as is the distance between the speakers.

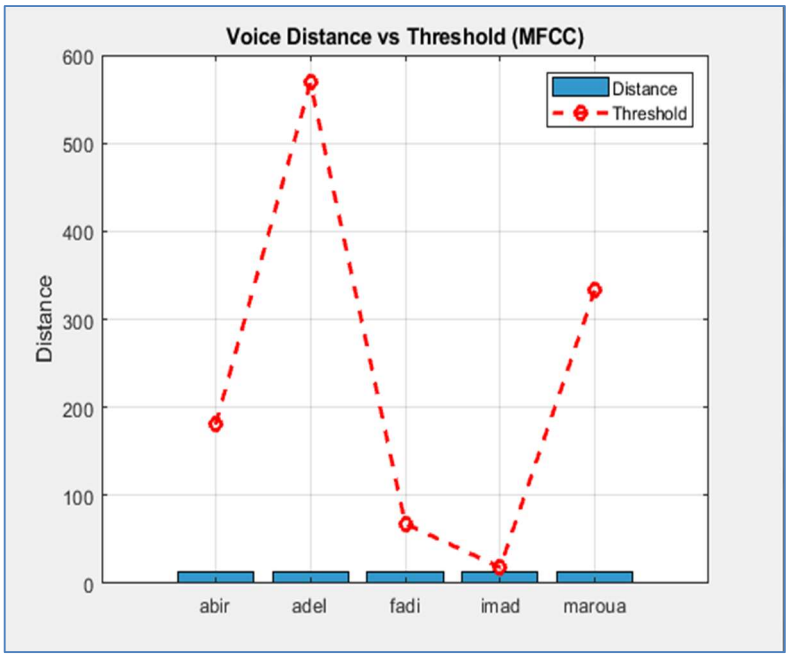


Figure 3.6: Figure of the speaker acceptance process "imad"

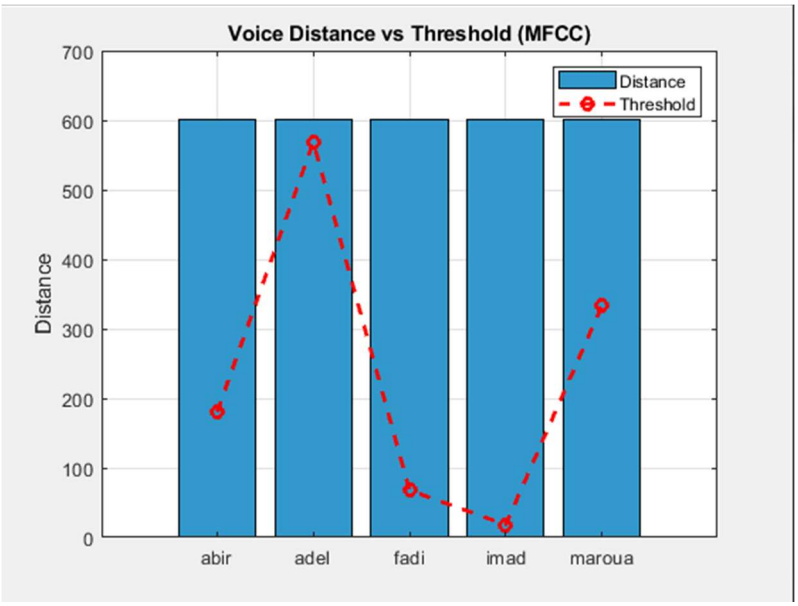


Figure 3.7: Figure of the process of rejecting a non-existent speaker

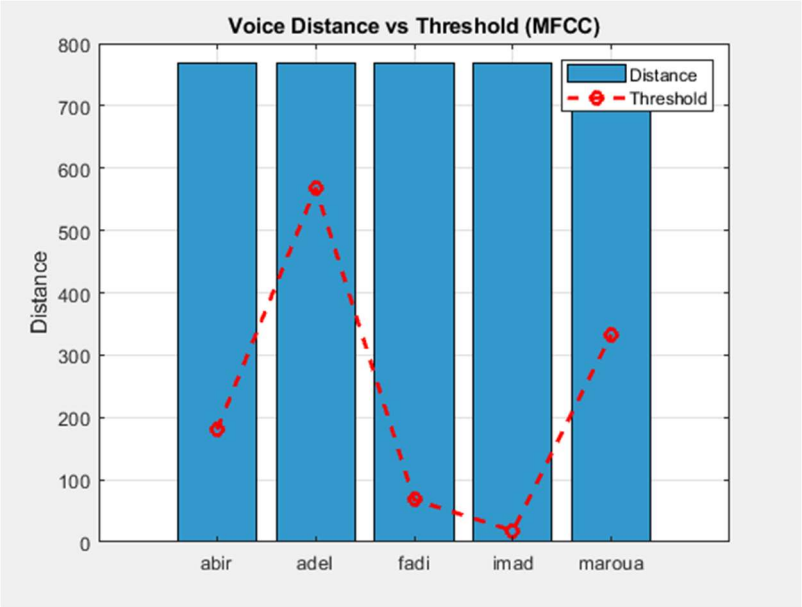


Figure 3.8: Figure of the process of rejecting an existing speaker

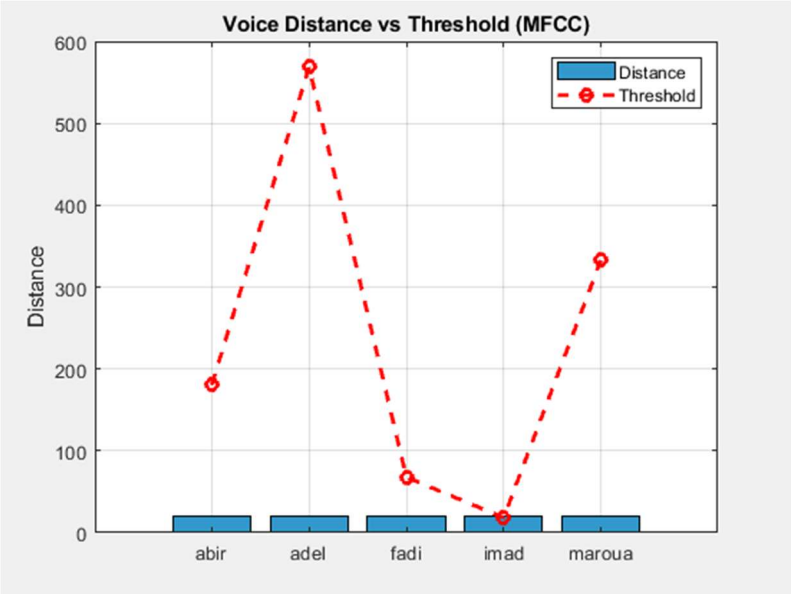


Figure 3.9: Figure of the process of accepting an absent speaker

Table 3.2 : The obtained results for performance evaluation

	Number of registrations
True Positives (TP)	31
False Negatives (FN)	4
True Negatives (TN)	6
True Negatives (TN)	8

In this table, we note that the number of accepted records was 31 out of 35 records in the database, and the number of rejected records was 4. As for the non-existent records, 6 out of 14 records were accepted, and 8 records not in the database were rejected.

Table 3.3 : The obtained results for Accuracy , FAR and FRR

	Score
Accuracy	75.51%
FAR	57.14 %
FRR	11.43%

For this table we found that the accuracy rate in vote verification was 75.51%, with 57.14 % recorded in False Acceptance Rate (FAR) and 11.43% in False Rejection Rate (FRR).

3.5.2 Experiment 2: Using FFT for registration and verification

In this experiment, the Fast Fourier Transform (FFT) will be used to verify the presence of a voice in the database. After recording voices or selecting pre-existing voices, a threshold is set for each speaker. To verify, we enter a voice or select a pre-recorded voice. After comparing it with the threshold value for each speaker in the database, the result displays whether the voice belongs to the database or not.

Table 3.4: Threshold results for each speaker in the database

	Speaker 1	Speaker 2	Speaker 3	Speaker 4	Speaker 5
Threshold	0 .6862	0.9221	0.7971	0.8125	0.7555
Distance	0.6595	0.6952	0.7967	0.6280	0.6253

We notice from the table that the threshold values for the five speakers are close to each other, and the same applies to the distance between the speakers.

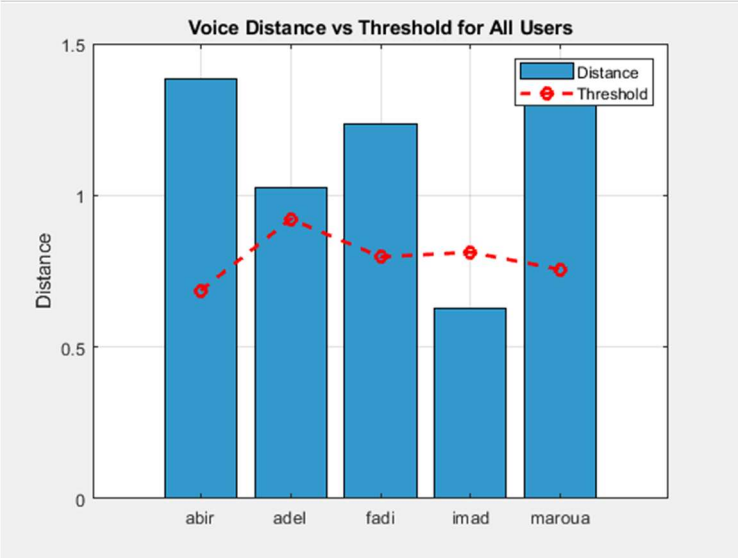


Figure 3.10: Figure of the speaker acceptance process "imad"

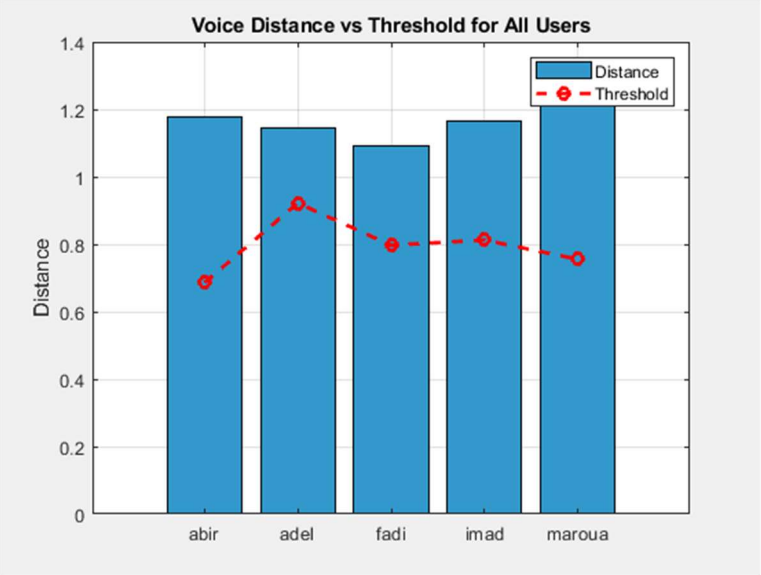


Figure 3.11: Figure of the process of rejecting a non-existent speaker

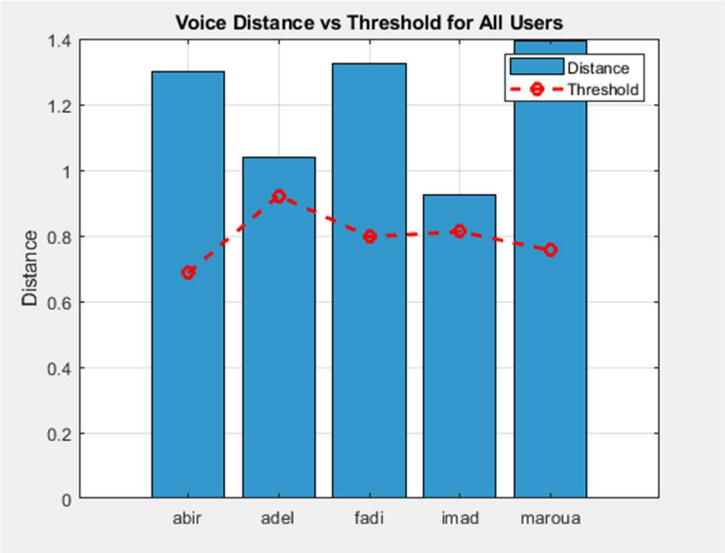


Figure 3.12: Figure of the process of rejecting an existing speaker

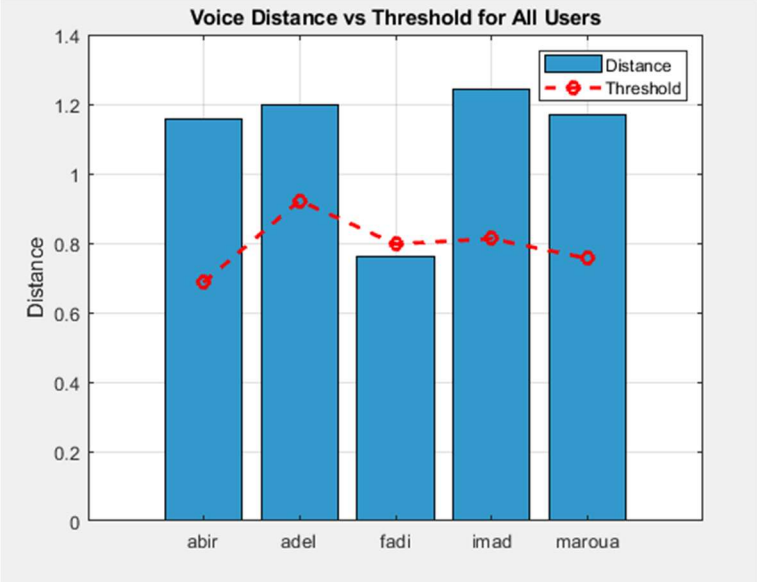


Figure 3.13: Figure of the process of accepting an absent speaker

Table 3.5 : The obtained results for performance evaluation

	Number of registrations
True Positives (TP)	23
False Negatives (FN)	12
True Negatives (TN)	13
True Negatives (TN)	1

In this table, we note that the number of accepted records was 23 out of 35 records in the database, and the number of rejected records was 12. As for the non-existent records, 13 out of 14 records were accepted, and 1 records not in the database were rejected.

Table 3.6 : The obtained results for Accuracy , FAR and FRR

	Score
Accuracy	73.47%
FAR	7.14%
FRR	34.29%

For this table we found that the accuracy rate in vote verification was 73.47%, with 7.14% recorded in False Acceptance Rate (FAR) and 34.29% in False Rejection Rate (FRR).

Table 3.7 : The obtained results for the different methodes

Approach / Scenario	MFCC	FFT
Accuracy	75.51%	73.47%
FAR	57.14%	7.14%
FRR	11.43%	34.29%

In this final table we have combined the results obtained from applying MFCC in registration and verification in the first experiment as well as the results of using FFT in the second experiment. By combining the results we find that the accuracy rate of MFCC is better than that of FFT. In FAR the rate was 57.14 % when using MFCC but when using FFT we recorded 7.14%. As for FRR, when using MFCC we got 11.43% and when using FFT the rate was34.29%.

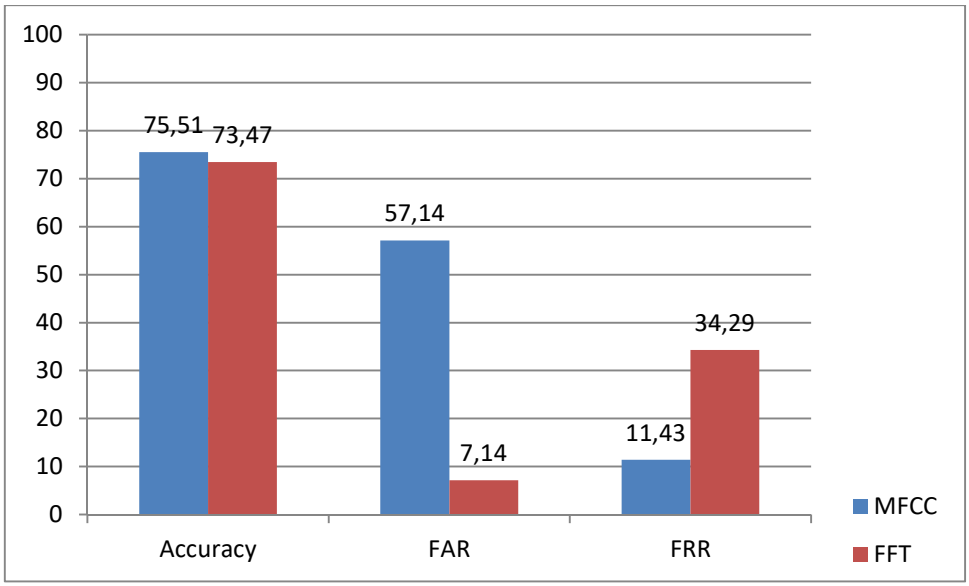


Figure 3.14: Column chart of verification results

3.6 Discussion

Through the two conducted experiments and the comparative analysis of their final outcomes, it was observed that, in terms of voice verification accuracy, the MFCC method demonstrated slightly superior performance compared to the FFT method. Concerning the False Acceptance Rate (FAR), MFCC exhibited a higher rate, indicating that FFT was more effective in correctly rejecting speakers not enrolled in the database. Conversely, regarding the False Rejection Rate (FRR), the FFT method recorded a higher rate than MFCC, meaning it rejected a greater number of genuine speakers than appropriate.

Conclusion

In this chapter, we presented and analyzed the results obtained by applying our methods to the developed database. The evaluation revealed that the MFCC technique achieved superior accuracy compared to other methods, while the FFT approach demonstrated better performance in minimizing false positive rates. Regarding false negatives, the MFCC method outperformed others in terms of rejection accuracy.

It is likely that enhancing the dataset by including more than seven individuals and increasing the number of records per individual could lead to improved results.

Conclusion

Conclusion

The primary objective of this project was to design and implement a speaker verification system capable of evaluating the identity of an unknown speaker by analyzing their speech signal. The system operates by extracting relevant features from the input speech and comparing them against reference models stored during the enrolment (training) phase for each registered speaker.

Throughout this research, it was observed that the proposed verification approaches are effective when the speaker being tested is the same as the one enrolled during the training phase. However, perfect verification accuracy was not achieved. In particular, the MFCC-based method reached accuracy rates of 75.51% and 73.47%, which can be attributed to several contributing factors.

One of the primary limitations is the quality of the acquisition equipment, as the PC's built-in microphone used in this study does not guarantee high-quality speech signal capture. Additionally, the lack of a customized decision threshold for each speaker — which can vary significantly depending on the spoken keywords or phrases — may have impacted the results. Another factor is the limited number of participating speakers and recorded samples per speaker, which constrains the robustness of the evaluation.

Despite these limitations, the developed system offers a significant advantage in that it operates in real time, making it suitable for practical deployment in various applications, including access control systems, telephone banking, biometric authentication for telephone-based transactions, information and reservation services, time and attendance monitoring, secure access to confidential information, and forensic investigations.

As future work, it is recommended to continue the development of these verification algorithms by focusing on optimizing computational efficiency and memory management, with the goal of ensuring that the application performs reliably in real-time environments."

Bibliographic References

BIBLIOGRAPHIC REFERENCES

[Ama, 2009] : AMARA KORBA Mohamed Cherif, Reconnaissance Automatique de la Parole par les HMM en Milieu Bruité : Contribution par paramétrisation acoustique robuste, Thèse de doctorat, 2009

[Cam, 1977] : CAMBELL.J.P. ''Speaker recognition: a tutorial proceeding of IEEE, Vol.85.no.9.pp 1437-1462,September 1977,

[Ezz, 2002] : EZZAIDI Hassan. "Discrimination Parole/Musique et étude de nouveaux paramètres et modèles pour un système d'identification du locuteur dans le contexte de conférences téléphoniques''. Thèse de doctorat, 10 octobre 2002, Université de Québec, Canada.

[Fre, 2000] : FREDOUILLE Corinne. "Approche statique pour la reconnaissance automatique de locuteur : Informations Dynamiques et Normalisation Bayésienne des Vraisemblances". Thèse de doctorat, 24 octobre 2000, Université d'Avignon et des Pays de Vaucluse, France.

[Jai, 1999] : JAIN.L.C. ''Intelligent Biometrics Techniques in Fingerprint and Face recognition. CRC Press, 1999.

[Jrj, 2000] : Jr., J. D., Hansen, J., and Proakis, J. Discrete-Time Processing of Speech Signals, second ed. IEEE Press, New York, 2000.

[Khe, 2018] : S. Khennouf « Fusion multi-classifieur décisionnelle en vue d'une discrimination inter-locuteur », Mémoire de Doctorat, Université des Sciences et de la Technologie Houari Boumediène (USTHB), 2018, Alger (Algérie)

[Khe, 2010] : S. Khennouf « Système automatique pour l'orientation de caméra mobile vers des cibles sonores », Mémoire de Magister, Université des Sciences et de la Technologie Houari Boumediène (USTHB), 2010, Alger (Algérie)

[Nor, 2004] : NORDQVIST Peter. "Sound classification in hearing instruments". Doctorat thesis, Royal institute of technology, 2004, Stockholm.

[Nor, 2004] : NORDQVIST Peter. "Sound classification in hearing instruments". Doctorat thesis, Royal institute of technology, 2004, Stockholm.

[Oua, 2009] : OUAMEUR.S. "Indexation automatique des documents audio en vue d'une classification par locuteurs –Application à l'archivage des émissions TV et Radio-". Thèse de doctorat, 2009, Ecole Nat. Sup. Polytechnique, Algérie

[Red, 2012] : Reda JOURANI, Reconnaissance automatique du locuteur par des GMM _a grande marge, Thèse de doctorat, 6 septembre 2012

[REY, 1995] : A. Reynolds, Douglas . "Automatic Speaker Recognition Using Gaussian Mixture Speaker Models." THE LINCOLN LABORATORY JOURNAL VOlUME 8, NUMBER 2, 1995 : 173-192. Print.

[SAM, 2011] : l. Samborski, R., & Sierra, D. (2011). Hybrid wavelet-Fourier-HMM speaker recognition. International Journal of Hybrid Information Technology, 4(4), 25-42.

[Say, 2003] : SAYOUD. H. ”Reconnaissance automatique du locuteur : Approche connexionniste”. Thèse de doctorat d’état, novembre2003, USTHB, Algérie.

[SIN, 2014] : Singh, Nilu. "A Study on Speech and Speaker Recognition Technology and its Challenges." proceedings of National Conference on Information Security Challenges, DIT, BBAU, Lucknow, INDIA. lucknow: Bharat Book Center, 2014. 34-37.

[Vel, 2006] : VELHO Filipe. ”La reconnaissance du locuteur à l’aide de la transformée en ondelettes contenue”. Mémoire présentée à l’école de technologie supérieure, Université de Québec, 2006, Canada.

[Vel, 2006] : VELHO Filipe. ”La reconnaissance du locuteur à l’aide de la transformée en ondelettes contenue”. Mémoire présentée à l’école de technologie supérieure, Université de Québec, 2006, Canada.

[Yes, 2014] : YESSAD Dalila, Reconnaissance automatique du locuteur dans les réseaux de communication V0IP, Thèse de doctorat, 24/06/2014

[Out, 2016] : https://outilsrecherche.overblog.com/pages/Notes_131_Lappareil_Phonatoire_Humain-3083095.html 03-06-2016

الملخص:

يركز هذا العمل على التحقق الأوتوماتيكي من هوية المتكلم من خلال الصوت، وهي مهمة حديثة من مهام مشروع التعرف الآلي على المتحدثين. ينقسم هذا العمل إلى مرحلتين: تتمثل المرحلة الأولى في إدخال تسجيلات صوتية لمجموعة من المتحدثين في ملف صوتي يحتفظ به وتسمى هذه المرحلة "مرحلة التدريب أو التعلم". وفي المرحلة الثانية (مرحلة الاختبار) يتم إدخال تسجيل صوتي مجهول الهوية للتعرف عليه من خلال مقارنة خصائصه مع خصائص التسجيلات السابقة وذلك باستعمال مجموعة من الخوارزميات (مثل: MFCC، FFT). مقاطع الكلام التي تم تسجيلها ومعالجتها في هذا العمل هي تسجيلات تم الحصول عليها باستعمال ميكروفون الحاسوب بدل المعدات والتجهيزات الخاصة بذلك مما يفسر ضعف النتائج المتحصل عليها.

Abstract:

This work focuses on automatic speaker verification identity via his voice, which is a recent task of the Automatic Speaker Recognition field. This work is divided into two phases: The first phase "called training or learning phase" consists of acquiring and saving audio recordings of a group of speakers into PC. In the second phase "testing phase", an anonymous speech is introduced and compared its characteristics with the characteristics of previous recordings, using a set of algorithms (such as: MFCC, FFT) to make a decision. The speech files that were recorded and processed in this work are obtained using a personal computer microphone instead of special equipment, which explains the weakness of the obtained results.

Résumé

Ce travail porte sur la vérification automatique de l'identité du locuteur via sa voix, qui est une tâche récente du domaine de la reconnaissance automatique du locuteur. Ce travail est divisé en deux phases : La première phase "dite phase d'apprentissage" consiste à acquérir et sauvegarder sur PC les enregistrements audio d'un groupe d'intervenants. Dans la deuxième phase "phase de test", un son anonyme est introduit et comparé ses caractéristiques avec les caractéristiques des enregistrements précédents, en utilisant un ensemble d'algorithmes (tels que : MFCC ,FFT) pour prendre une décision. Les fichiers vocaux qui ont été enregistrés et traités dans ce travail sont obtenus à l'aide d'un microphone d'ordinateur personnel au lieu d'un équipement spécial, ce qui explique la faiblesse des résultats obtenus.