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

DOMAIN : SCIENCE AND TECHNOLOGY
FILIERE : ELECTRONICS
OPTION : INSTRUMENTATION

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

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

BENAMOR Imad Eddine

&

BOUDJELLAL Fadi

Entitled

Access Control Using Specific Code and Biometric Identification

Diplôme de Master dans le cadre du décret ministériel 1275

Presented on: June 22nd, 2025, in front of the jury composed of :

Dr. ROUABHI Riyadh	University of Mohamed Boudiaf - M'sila	President
Dr. KHENNOUF Salah	University of Mohamed Boudiaf - M'sila	Supervisor
Dr. BAKRI Adel	University of Mohamed Boudiaf - M'sila	Co-Supervisor
Dr. OUALI Mohamed Assam	University of Mohamed Boudiaf - M'sila	Examiner
Pr. ATTALLAH Bilal	University of Mohamed Boudiaf - M'sila	INCUBATEUR Representative
	Directorate of Social Action and Solidarity.	Socio-Economic Partner

June 2025

Acknowledgements

الحمد لله رب العالمين

First for most we would like to thank our creator, our lord Allah, for making it all possible for us to complete this dissertation, for showing us that there is a light to every tunnel.

We want to thank our supervisors, Dr. Salah KHENNOUF and Dr. Adil BAKRI for everything patience, assistance and encouragement while we were completing this dissertation. Thank you for being available whenever we have a question or concern, whether in person or by message.

We would also like to express our profound gratitude to the members of the jury: Dr. ROUABHI Riyadh, Dr. OUALI Mohamed Assam and Pr. ATTALLAH Bilal for their gracious evaluation of this research and their valuable feedback.

A special acknowledgment goes to our friends and colleagues who participated in conceiving our Audio Database.

F. Boudjellal & I. Benamor

Dedication

My beloved parents, thank you so much for your loves, supports, prayers all the days and nights for making me able to get such success and honor, and everything that you give it to me until this moment. I will make you happy and proud of me

My family and those who left us but whose souls remain with us, my sisters, my brother thank you for your sacrifices and make me happy all the time. I am so lucky to being your brother, thank you for everything

I want to thank my friends and my soulmate for all the memories

Finally, I dedicate this graduation to myself and my friend IMAD for believing in our selves.

For all the difficult times we have experienced, for doing all this work together, for having no days off.

Fadi

Dedication

My beloved parents, thank you so much for your love, support, and prayers throughout the days and nights, which have enabled me to achieve this success and honor, and for everything you have given me up until this moment. I will make you both happy and proud.

My sisters and brother, thank you for your sacrifices, which have always made me happy. I am so lucky to have you as my brother. Thank you for everything.

I would like to thank my friends, especially my friend AYOUB BEERA for his touch in this work, and my life partner for all the memories.

Finally, I dedicate this graduation to myself and my friend FADI for believing in each other.

For all the difficult times we have been through, for putting in so much effort together, and for never having a day off.

Imad Eddine

Table of Contents

Chapter II : Speaker Recognition ; Methods And Algorithms

Chapter III: Experiments and Results

List of Abbreviations

ANN	Artificial Neural Network
ASI	Automatic Speaker Identification
ASR	Automatic Speaker Recognition
ASV	Automatic Speaker Verification
DCF	Decision Cost Function
DET	Detection Error Tradeoff
DTW	Dynamic Time Warping
EER	Equal Error Rate
FAR	False Acceptance Rate
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FRR	False Rejection Rate
FFT	Fast Fourier Transform
GB	Gradient Boosting
HTER	Half Total Error Rate
LPC	Linear prediction coefficients
LSF	Line Spectral Frequencies
LSP	Line Spectral Pair
MFCC	Mel Frequency Cepstral Coefficients
PLP	Perceptual Linear Prediction
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine
TN	True Negative

TP True Positive

TPR True Positive Rate

VQ Vector Quantization

List of figures

List of tables

Introduction

Introduction

With the rapid digital evolution and the increasing need for secure systems, biometric authentication has become an essential component in many sectors, such as government institutions, banks, businesses, and smart homes. Biometric systems rely on unique physiological or behavioral characteristics of each individual, making them more accurate and secure than traditional methods such as passwords or cards.

The biometric characteristics used today include fingerprints, facial features, iris scans, voice, and signatures. These characteristics are widely used in access control and surveillance systems due to their ease of use and high accuracy.

This study focuses on voice recognition technology, a behavioral biometric technology that relies on physiological characteristics and behavioral characteristics .This technology is divided into two tasks: identifying a speaker from a group and verifying the speaker's alleged identity.

The study raises a fundamental question: Can a person's voice be relied upon as a secure means of identity verification? To answer this question, technical challenges such as voice variability, environmental noise, and classification accuracy were analyzed. The research relied on studying two popular methods for extracting audio features: MFCC and PLP, along with testing several supervised learning algorithms such as SVM, Random Forest, Neural Networks, and Gradient Boosting, with the aim of identifying the most suitable algorithm for application in audio-based security systems.

The first chapter is dedicated the use of biometric systems to identify people using features such as face, voice, and fingerprints, explaining the role of sensors in converting these into digital signals, also highlights their application in security, health, and banking.

In the second chapter, we will discuss the stages of a speaker recognition system, starting with extracting audio features such as MFCC and PLP, then training models using algorithms such as SVM and Random Forest.

The last chapter will be devoted to Experiments were conducted on a speaker recognition system using MFCC and PLP features. Four classification algorithms were tested, with SVM achieving the highest performance. The results showed that PLP performed best with neural networks and random forests, while the Gradient Boosting algorithm was the least efficient. Finally, we will finish this work with a general conclusion.

Chapter I

General Information on Biometrics

Chapter I
General Information on Biometrics

I.1. Introduction

Attendance management system is a system that records the time of workers entering and leaving according to a specific time and during specific days. These modern systems are easy to use and avoid all kinds of fraud, they work to identify people through the various features that humans carry, their use takes only seconds or once a person crosses in front of them. These modern systems reduced many of the problems they were facing in the past, where they were recording with the same person in the papers, then it developed a little and they started to register in one machine with the papers, however this process takes a lot of work time. Then it evolved over time with the discovery of biometric sensors, which helped a lot and are the basis of these machines. The accuracy of the biometric systems lies in the accuracy of the biometric sensors. Now there are various of biometric attendance system which use one or many biometric sensors

I.2. Biometric system

A biometric system is a system that allows the recognition of a certain characteristic of an individual using mathematical algorithms and biometric data. There are several uses of biometric systems [1] is given in Fig. I.1. The word biometrics is derived from the Greek words bio and metric. Where bio means life and metric means to measure. Biometrics are used to identify his or her physical and behavioral characteristics of a person. This method of identification is chosen over traditional methods, including PIN numbers and passwords for its exactness and case sensitiveness. Based on the designing, this system can be used as an identification system or authentication system. These systems are divided into various types which include vein pattern, fingerprints, hand geometry, DNA, voice pattern, iris pattern, signature dynamics and face detection [2].

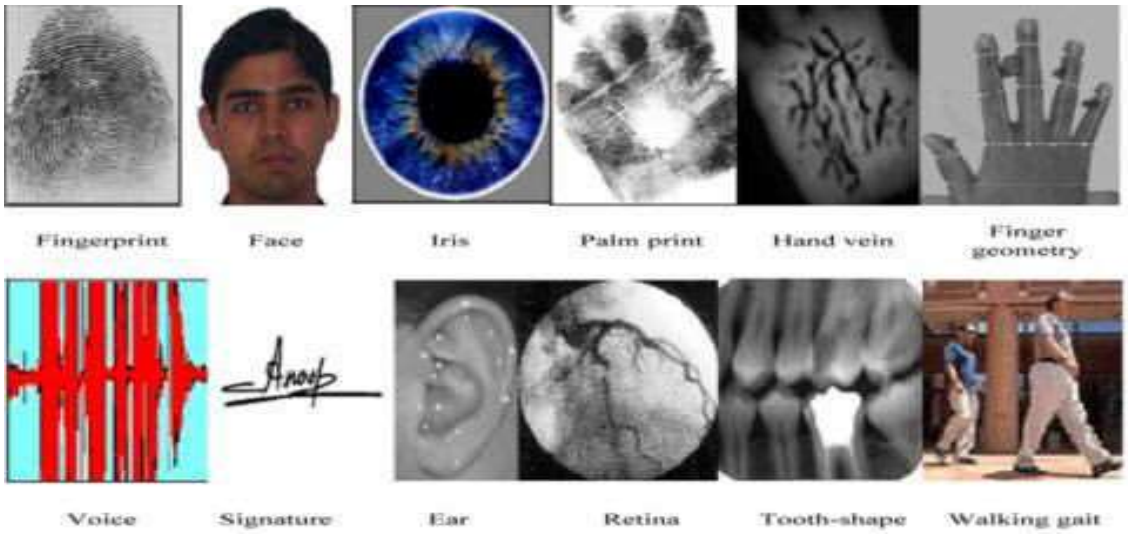


Fig. I.1. Physical and behavioral characteristics of a person [2]

I.3. Biometric Sensor [2]

A biometric sensor is a transducer that changes a biometric treat of a person into an electrical signal. Biometric treats mainly include biometric fingerprint reader, iris, face, voice, etc. Generally, the sensor reads or measures light, temperature, speed, electrical capacity and other types of energies. Different technologies can be applied to get this conversation using sophisticated combinations, networks of sensors and digital cameras. Every biometric device requires one type of sensor. The biometrics applications mainly used in a high definition camera for facial recognition or in a microphone for voice capture. Some biometrics is specially designed to scan the vein patterns under your skin. Biometric sensors are an essential feature of identity technology.

I.4 Different biometric modalities [2]

Biometric sensors or access control systems are classified into two types such as Physiological Biometrics and Behavioral Biometrics. The physiological biometrics mainly include face recognition, fingerprint, hand geometry, Iris recognition, and DNA. Whereas behavioral biometrics include keystroke, signature and voice recognition

I.4.1 physiological biometrics [2]

Physiological biometrics employ physical, structural, and relatively static attributes of a person such as their fingerprints, the pattern of their iris, contours of their face, or the geometry of veins in their hands. Some modalities are microscopic in nature but still

exhibit biological and chemical structures that can be acquired and identified e.g., DNA and odor.

Many physiological biometrics are permanent features and are resistant to change unless they are accidentally or intentionally damaged, degraded or destroyed. Examples, in the case of fingerprinted might include the partial erosion of the friction ridges on the hands of those engaged long term in handling or laying bricks, the loss or amputation of hands/fingers and, in extreme cases, invasive surgery or self-harm to remove or alter the ridge structure..

I.4.1.1 Fingerprint sensing

Fingerprint Recognition includes taking a fingerprint image of a person and records its features like arches, whorls, and loops along with the outlines of edges, minutiae, and furrows. Matching of the Fingerprint can be attained in three ways, such as minutiae, correlation, and ridge

- Minutiae based fingerprint matching stores a plane includes a set of points and the set of points are corresponding in the template and the i/p minutiae.
- Correlation-based fingerprint matching overlays two fingerprint images and the association between equivalent pixels is calculated.
- Ridge feature-based fingerprint matching is an innovative method that captures ridges, as minutiae-based fingerprint capturing of the fingerprint images is difficult in low quality.

Fig. I.2. Fingerprint sensor [2].

To capture the fingerprints, present methods employ optical sensors that use a CMOS image sensor or CCD, solid-state sensors work on the principle of transducer

technology using thermal, capacitive, piezoelectric sensors or electric field, or ultrasound sensors work on echography in which the sensor sends acoustic signals through the transmitter near the finger and captures the signals in the receiver.

The scanning of the fingerprint is very stable and reliable. It safeguards entry devices for building door locks and access of computer network are becoming more mutual. At present, a small number of banks have initiated using fingerprint readers for approval at ATMs. I.4.2.

I.4.1.2 Face sensing

A face recognition system is one type of biometric computer application that can identify or verify a person from a digital image by comparing and analyzing patterns. These biometric systems are used in security systems. Present facial recognition systems work with face prints and these systems can recognize 80 nodal points on a human face. Nodal points are nothing but endpoints used to measure variables on a person's face, which includes the length and width of the nose, cheekbone shape, and eye socket depth.

Fig. I.3. Face recognition system [2].

Face recognition systems work by capturing data for the nodal points on a digital image of a person's face and resulting data can be stored as a face print. When the conditions are favorable, these systems use face prints to identify accurately. Currently, these systems focus on smart phone applications, which include personal marketing, social networking, and image tagging purposes.

Social sites like Facebook uses software for face recognition to tag the users in photographs. This software also increases marketing personalization. For instance,

billboards have been designed with integrated software that recognizes the ethnicity, gender and estimated age of onlookers to deliver targeted marketing.

I.4.1.3 Iris sensing

Iris recognition is one type of biometric method used to identify the people based on single patterns in the region of ring-shaped surrounded the pupil of the eye.

Generally, the iris has a blue, brown, gray or green color with difficult patterns that are noticeable upon close inspection.

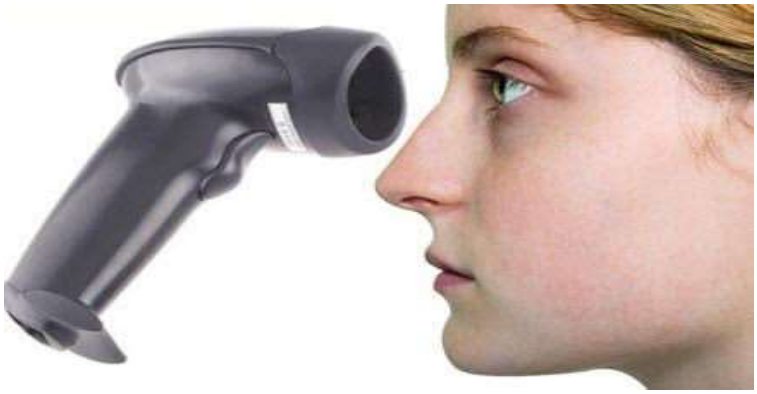


Fig. I.4. Iris recognition system [1].

I.4.2 Behavioral biometrics

Behavioral biometrics establish identity by monitoring the distinctive characteristics of movements, gestures, and motor-skills of individuals as they perform a task or series of tasks. This means human movements such as walking (gait analysis) or finger contact with a keyboard (keystroke) are captured and analyzed. Structural physiological factors will be fundamental in determining how such movements are performed by an individual e.g., proportional length of the feet, thigh and shinbones of the leg, the relative size and dexterity of hand/fingers etc.

A behavioral biometric system measures how an individual conducts an action or series of actions and then uses this data (user profile) to compare a subsequent performance of the same actions by a person to determine if he/she matches the user profile or is an imposter. The actions in question may be normal, routine functions e.g., interacting with a smartphone. It should be kept in mind that the user profile is not strictly a biometric template, which is a digital representation of a physiological feature, but a record of information, actions and settings associated with the user.

New biometric modalities are currently being developed that challenge even the broadest definition of behavioral biometrics. For example, heartbeats can be used to identify individuals, but they are an involuntary biological process and not an intentional or manipulated action performed by the individual in the same way as, for example, keystrokes. Therefore, behavioral biometrics encompasses a variety of modalities that exhibit both voluntary and involuntary repeated motions and associated rhythmic timings/ pressures of body features ranging from signatures, gait, voice, and keystrokes through to eye tracking and heartbeats.

Some applications have been developed using software and algorithms that are significantly different to standard biometric recognition systems. They are designed primarily to group people rather than identify them individually. These systems may be used to estimate factors such as age, gender or race or evaluate the mood, emotional state, or attentiveness of individuals and/or groups.

The highly variable accuracy rates of these types of applications have caused widespread concern and much public debate regarding fundamental human rights, privacy, and the acceptable limits of such technologies in civil society (Refer to the Biometrics Institute Good Practice Framework B.1.1, C.1.1/2/3/4 and B.5.1) and the Biometrics Institute Three Laws of Biometrics – formulate policy then processes and then the technology).

I.4.2.1 Voice sensing

Voice recognition technology is used to produce speech patterns by combining behavioral and physiological factors that can be captured by processing speech technology. The most important properties used for speech authentication are nasal tone, fundamental frequency, inflection, cadence. Voice recognition can be separated into different categories based on the kind of authentication domain, such as a fixed text method, in the text dependent method, the text-independent method, and conversational technique.

Fig. I.5. Voice recognition system [1].

I.4.2.2 Signature sensing

Signature recognition is one type of biometric method used to analyze and measure the physical activity of signing like the pressure applied, stroke order and speed. Some biometrics are used to compare visual images of signatures. Signature recognition can be operated in two different ways, such as static and dynamic.

Fig. I.6. Signature recognition system [1].

In static mode, consumers write their signature on paper, digitize it through a camera or an optical scanner. This system identifies the signature examining its shape.

In dynamic mode, consumers write their digitized signature in a tablet, which obtains the signature in real-time. Another option is gaining by means of stylus operated PDAs. Some features also operate with smart-phones with a capacitive screen, where consumers can sign using a pen or a finger. This type of recognition is also known as “on-line”.

I.5. architecture of a biometric system [1]

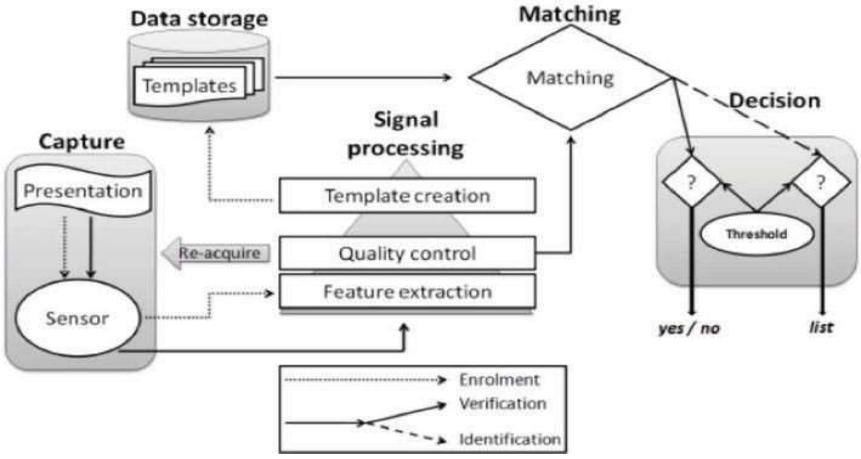


Fig. I.7. presents the architecture of a biometric system [1].

- The capture module that represents the entry point of the biometric system and consists in acquiring the biometric data in order to extract a digital representation.
- The module of signal processing makes it possible to optimize the processing time and the digital representation acquired in the enrollment phase in order to optimize the processing time of the verification phase and the identification.
- The storage module that contains the biometric templates of the system enrollees.
- The matching module that compares the data extracted by the extraction module with the data of the registered models and determines the degree of similarity between the two biometric data.
- The decision module that determines whether the similarity index returns through the matching module is sufficient to make a decision about the identity of an individual.

I.6. Difference between biometric authentication and identification [3]

Sometime it is very difficult to understand the difference between Authentication and Identification, two words and actions they perform. Authentication and identification are closely connected with verification and authorization.

Biometric Identification is the automatic identification of living individuals by using their physiological and behavioral characteristics, negative identification can only be accomplished through biometric identification, if a pin or password is lost or forgotten it can be changed and reissued but a biometric identification cannot.

For example, there is a database where all the photos of users are collected. Suddenly, somebody comes to you and greets you. You want to know who it is, and put the picture of this person to the system. The system is looking for the match. When the match is found the system represents the full information about this person.

Verification means verifying person's identity. A guy comes to you and tells his name, Bill. You take the picture of Bill and put it to the system. The system finds the Bill's file and tries to match the pictures. If the result is positive the system indicates that this guy is really Bill, is negative it indicates that it is not Bill.

Authentication is the same as verification, its task to verify if the user is actually who he claims to be. Authorization means whether the user has a right to access to the system. In practice it looks like the same as if you come to the cinema. You should buy a ticket, because you know that a person who checks the tickets will not allow you to see the movie without a ticket. There is no identification or verification process.

I.7. Areas of application

I.7.1 Security and Access Control

Biometrics are widely used to secure facilities, whether government institutions, private companies, or even homes. These systems rely on fingerprints, facial recognition, or iris recognition to control individual access to certain premises. For example, in many modern offices, the use of cards or keys is being replaced by fingerprint systems to ensure that only authorized individuals can enter.

This technology is effective because it links identity to the individual, rather than to something that can be lost or stolen, such as an access card.

I.7.2 Banking and financial services sector

Biometrics contribute to enhancing security in the banking sector, whether through ATMs that require a fingerprint instead of a card, or through mobile applications that rely on fingerprints or facial recognition to log in and make financial transfers. Some banking institutions have also begun adopting voice recognition technology to verify a customer's identity when contacting customer service, reducing the risk of fraud or identity theft. This type of technology enhances customer confidence and improves the user experience.

I.7.3 Heath sector

In the healthcare field, biometrics helps identify patients with high accuracy, reducing medical errors resulting from name similarities or record tampering. For example, a fingerprint can be used to link a patient to their electronic medical record, ensuring they receive the correct care. Some hospitals also rely on fingerprints or iris recognition to control medical staff access to sensitive patient files, enhancing the protection of health data privacy. The technology is also used to dispense controlled medications to ensure they are not misused or dispensed to unauthorized individuals.

I.7.4 Security and Law Enforcement

In the field of policing and public safety, biometrics are used to identify suspects and wanted persons. Many law enforcement agencies maintain massive databases of fingerprints and facial images, which are used to match forensic evidence with people's identities. Facial recognition systems are also used in security cameras in public places, helping to track criminals or instantly identify wanted individuals as they pass through checkpoints.

I.7.5 Smart Homes and Consumer Technology

In smart homes, it has become possible to use biometrics to unlock doors or operate devices, such as by unlocking the door with a fingerprint or facial recognition upon approach. This technology is also used to personalize the user experience. For example, a smart TV can recognize the user's face and automatically adjust the channel list or settings based on their preferences. Similarly, smart assistant devices are used to identify the speaker using voiceprint and deliver personalized content.

I.8. advantage of biometric attendance system [4]

There are many advantages of biometric attendance system, we mention some of them:

- Avoid crowding and wasting time
- Ease of use and fast (it makes seconds to put your finger, card or other thing to attend)
- The possibility of conducting studies to know the behavior and discipline of workers

- Avoid fraud (Direct registration in the machine without the presence of a person to register you and the inability to defraud the machine)
- benefit from human biometrics (More than 10,000 workers are registered)
- economic (You can use the card throughout the year instead of using papers every week or month, or using only the face or the fingerprint)

I.9. Conclusion

The attendance recording machine is a device that records the time of the workers, attendance in a short time and is different according to the nature of the work, for example the mobile device is intended for constantly moving workers and the fixed machine is intended for workers inside an organization, also it varies according to the sensors used which are the most important part of the device, and the use of sensors varies depending on the necessity of using it, for example, if the establishment is very sensitive, sometimes several advanced sensors are used.

There are 5 types of known biometric sensors, which are iris, face, fingerprint, voice and signature. The way these devices work is simple as the sensor takes a sample from the person and an algorithm processes it and extracts its features, then identifies the person by comparison or matching with the database and then records his presence, if he is not registered, then after the process of processing and extracting the features, it is added to the database.

Chapter II

Speaker Recognition; Methods and Algorithms

Chapter II
Speaker recognition: Methods and Algorithms

II.1 Introduction

The problem of Automatic Speaker Recognition [6] is more specifically linked to the problem of identification and verification. Automatic Speaker Recognition (ASR) consists of recognizing a person's identity by analyzing their voice. And recently, ASR has been the subject of increased interest, along with biometric recognition methods, namely fingerprint and genetic analysis.

While ASR is not among the most reliable biometric techniques, it has several qualities that distinguish it from other methods, particularly in terms of ease of deployment, simplicity of audio recording, low cost of the equipment involved, and finally, ASR offers the unique advantage of being usable remotely. However, the very principle of ASR induces several challenges that must be addressed when implementing a speaker recognition system.

Indeed, the ability to identify speakers relies on the differences between the voices of various speakers. But this interlocutor variability is in competition with intra-speaker variability (change of the voice of the same speaker between two recordings, voluntary (in the case of an attempt at imposture) or not), the variability of the operating environment (noise, recording level) and the speech signal transmission channel (for example during a telephone transmission).

II.2 The basics of acoustics and phonetics

II.2.1 Voice signal analysis

Acoustic analysis of speech signals involves extracting relevant information and minimizing redundancy. Generally, a set of acoustic coefficients is calculated at regular time intervals, on signal blocks of fixed length. This set of coefficients constitutes an acoustic vector.

There are many acoustic parameterization techniques. However, they can be grouped into three main families:

- Analysis by filter banks.
- Analysis by Fourier transform
- Analysis by linear prediction.

II.2.2 Acoustic pré-traitement

In the acoustic pre-processing part, two stages can be distinguished, a pre-emphasis stage and another stage of windowing of the speech signal, as shown in Figure II.1.

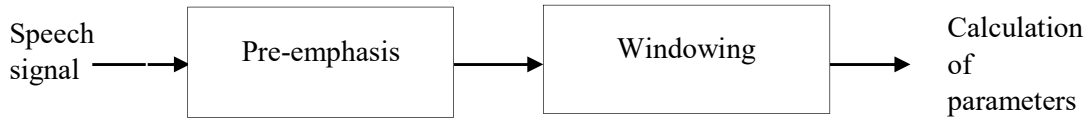


Figure II.1. Pre-processing and parameter extraction.

II.2.2.1. Pre-emphasis

The outgoing acoustic wave from the lips undergoes, due to the mismatch between the two internal and external environments, a distortion comparable to a de-emphasis of 6 dB per octave over the entire spectrum [9]. To be able to compensate for this distortion, and accentuate the high frequencies, a high-pass pre-emphasis filter of transmittance is applied:

$$H(z) = 1 - \alpha\, Z^{-1} \qquad (II.1)$$

Avec $0.9 \leq \alpha \leq 1$

II.2.2.2. Windowing

The windowing step consists of applying a sliding window of limited duration to the speech signal in order to limit the number of samples and reduce edge effects (Gibbs phenomenon). Among the different weighting windows, the most commonly used are: the rectangular window, the Hamming window, the Hanning window, and the Blackman window. In speech processing, the Hamming window is the most commonly used.

This window is given by the expression:

$$w(n) = 0.54 + 0.46cos\left(\frac{2\,\pi\, n}{N-1}\right) \qquad (II.2)$$

with $0 \leq n \leq N-1$

N: The number of samples in a window.

II.2.3. Acoustic parameters

II.2.3.1. Signal energy

Signal energy is an indicator that can help increase the performance of a recognition system. This energy corresponds to the power of the signal. It is often evaluated over several successive signal frames to highlight its variations. The formula for calculating this parameter

is: $E = \sum_{n=0}^{N-1} s^2(n)$ *(II.3)*

As an acoustic parameter, we can also use the logarithmic energy which is defined as follows:

$$E = \ln(\sum_{n=0}^{N-1} s^2(n)) \quad (II.4)$$

Where N is the number of signal samples, and s(n) are the signal samples. The energy thus obtained is sensitive to the recording level, generally it is normalized, expressed in decibels.

II.2.3.2. LPC linear prediction coefficients

The basic principle of linear prediction is that a given sample can be predicted from a linear combination of the finite samples that precede it [10]. A single set of predictor coefficients is determined by minimizing the differences between the current and predicted samples. The linear prediction technique is based on the speech production model. The function of the autoregressive model of speech production is described by:

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n) \quad (II.5)$$

Thus, each speech sample s(n) consists of a linear combination of p past speech samples. The predictor is defined as a system whose output is:

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \quad (II.6)$$

The prediction error is given by:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k s(n-k) \quad (II.7)$$

We seek to find a set of coefficients a_k so as to minimize the prediction error e(n) in a certain interval. The average of the error is given:

$$E = \sum_n e^2(n) = \sum_n \left[s(n) - \sum_{k=1}^{p} a_k s(n-k)\right]^2 \quad (II.8)$$

$$\partial E/\partial a_i = 0 \qquad (II.9)$$

pour *i*=1,….*p*

So:

$$\partial E/\partial a_i = -2\sum_n\{[s(n) - \sum_{k=1}^{p} a_k s(n-k)]s(n-i)\} = 0 \qquad (II.10)$$

This last equation leads us to write:

$$\sum_n s(n)\, s(n-i) = \sum_n \sum_{k=1}^{p} a_k\, s(n-k)s(n-i) \qquad (II.11)$$

We define : $\emptyset(i,k) = \sum_n s(n-i)s(n-k)$ *(II.12)*

So:

$$\sum_{k=1}^{p} a_k\, \emptyset(i,k) = \emptyset(i,0) \qquad (II.13)$$

i =1,….,*p*.

This set of p equations with p unknowns can be solved efficiently for the unknown prediction coefficients {a_k }. We assume that the speech segment is zero outside the interval 0 < n < La-1, where La is the length of the LPC analysis window. This is equivalent to multiplying the input speech signal by a finite window.

e(n) is nonzero only on the interval 0 < n < La +p-1.

So

$$\emptyset(i,k) = \sum_{n=0}^{L_a+p-1} s(n-i)s(n-k) \qquad (II.14)$$

i=1,….p

k=1,….p

We pose *m*= (*n-i*),

$$\emptyset(i,k) = \sum_{n=0}^{L_a-1-(i-k)} s(m)s(m+i-k) \qquad (II.15)$$

From where, $\emptyset(i) = R(i)$ et $\emptyset(i,k)$=*R* (*i-k*).

So $\sum_{k=1}^{p} a_k R(|i-k|) = R(i)$ *(II.16)*

we obtain:

$$\begin{pmatrix} R(0) & R(1) & R(2) & \dots & R(p-1) \\ R(1) & R(0) & R(1) & \dots & R(p-2) \\ R(2) & R(1) & R(0) & \dots & R(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(p-1) & R(p-2) & R(p-3) & \dots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(P) \end{pmatrix} \quad (II.17)$$

The p × p matrix of autocorrelation values is a symmetric Toeplitz matrix. This property can be exploited to obtain an efficient algorithm for solving the system of equations. The most efficient solution is an iterative method known as the Levinson Durbin algorithm [10].

$$E_0 = R_0 \quad (II.18)$$

$$K_i = -\left[R_i + \sum_{j=1}^{i-1} a_j^{(i-1)} R_{i-j}\right] / E_{i-1} \; \forall \; 1 \le i \le p \quad (II.19)$$

$$a_i^{(i)} = K_i \quad (II.20)$$

$$a_i^{(i)} = a_j^{(i-1)} + K_i a_{i-j}^{(i-1)} \forall \; 1 \le j \le i-1 \quad (II.21)$$

$$E_i = (1 - K_i^2) E_{i-1} \quad (II.22)$$

$$a_j = a_j^{(p)} \forall \; 1 \le j \le p \quad (II.23)$$

H(z) can be put in the form like the two equations (II.1) and (II.2).

II.2.3.3 The LSP and LSF coefficients

Line Spectral Frequencies (LSF) parameters, also called Line Spectral Pair (LSP) parameters, are related to the position of the poles on the frequency axis and have the property of being arranged in ascending order. This arrangement of parameters allows perceptual criteria to be taken into account and offers an efficient coding property with better stability control. Line spectral frequencies were first introduced by Itakura as an alternative to linear prediction coefficients [11,12]. LSPs are the solutions of the following two equations:

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \quad (II.24)$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \quad (II.25)$$

with

$$A(z) = \frac{1}{2}[P(z) + Q(z)] \quad (II.26)$$

It has been shown that if H(z) is stable, where A(z) is minimal phase, then the roots of P(z) and Q(z) lie on the unit circle and are intertwined with each other, the LSPs are in ascending order. The roots appear as conjugate pairs and therefore there are p LSPs positioned between 0 and π. It has been shown that if the LSPs, denoted by w_i, are in ascending order and unique, then the corresponding inverse filter A(z) is minimal phase, which guarantees stability.

$$0 = w_0^{(Q)} < w_1^{(P)} < w_2^{(Q)} < \cdots < w_p^{(Q)} < w_{p+1}^{(P)} = \pi$$

The LSF pattern corresponds to the spectrum of the LP filter. LSFs cluster around spectral peaks (see Figure I.7). A change in any LSF can only alter the spectrum in the region surrounding that LSF. LSFs can be calculated by several methods; Soong and Juang calculate LSFs by applying a discrete cosine transformation of the polynomial coefficients.

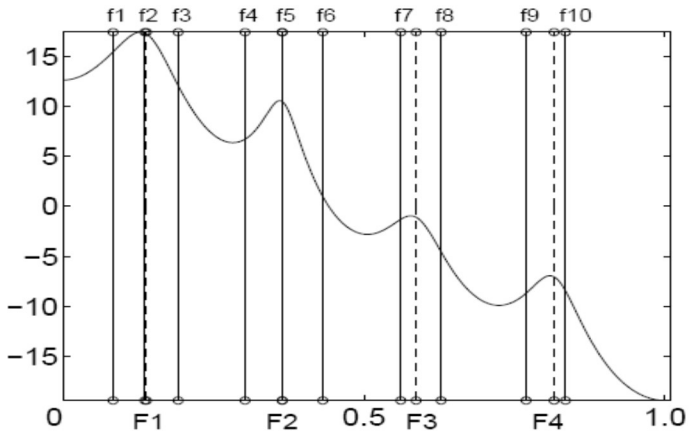


Fig II.2. LP spectrum with LSF positions.

II.2.3.4. MFCC coefficients (Mel Frequency Cepstral Coefficients)

The MFCC coefficients are extensions of the cepstral coefficients by the passage from the linear frequency scale to a non-linear frequency scale called the Mel scale [6]. The Mel-scale frequency is defined by:

$$B(f) = 2595\, log\left(1 + \frac{f}{700}\right) \qquad (II.27)$$

Where f is the frequency in Hz, B(f) is the Mel-scale frequency of f.

The advantage of the Mel scale is that it is quite close to scales derived from studies on sound perception and critical bandwidths of the ear [14].

The MFCC parameters are calculated as follows (Figure II.3):

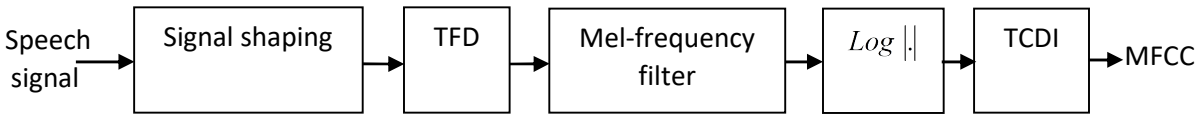


Figure II.3 Calculation of MFCC coefficients.

Let s(n) be a discrete signal with $0 \leq n \leq N-1$, N is the number of samples in an analysis window, F_e is the sampling frequency, the short-term Discrete Fourier Transform S(k) is obtained with the formula:

$$S(k) = \sum_{n=0}^{N-1} s(n)\exp\left(\frac{-j2\pi n}{N}\right).\ 0 \leq k \leq N-1$$

The signal spectrum is multiplied with triangular filters (see Figure II.4) whose bandwidths are equivalent in the mel-frequency domain. The boundary points B[m] of the mel-frequency filters are calculated as follows:

$$B[m] = B[f_l] + m\frac{B(f_h)-B(f_l)}{M+1}\ .\ 0 \leq m \leq M+1 \qquad (II.28)$$

Where M is the number of filters, f_h is the highest frequency and f_l is the lowest frequency for signal processing.

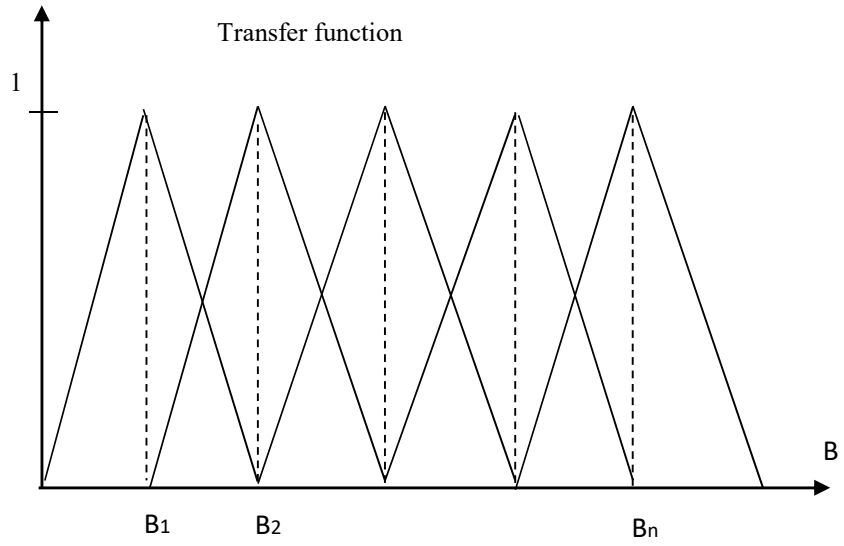


Fig II.4 Triangular bandpass filters in Mel-Frequency (B(f)).

In the frequency domain, the corresponding discrete points f[m] are calculated by

$$f[m] = \frac{N}{F_s}B^{-1}\left(B[f_l] + m\frac{B(f_h)-B(f_l)}{M+1}\right) \qquad (II.29)$$

Where B^{-1} is the mel-frequency to frequency transform..

$$B^{-1}(m) = 700\left(10^{\frac{m}{2595}} - 1\right) \qquad (II.30)$$

The coefficient Hm[k] of each filter is determined by the following system:

$$H_m[k] = \begin{cases} 0 & k \leq f[m-1] \\ \frac{k - f[m-1]}{f[m] - f[m-1]} & f[m-1] \leq f[m] \\ \frac{f[m+1] - k}{f[m+1] - f[m]} & f[m] \leq k \leq f[m+1] \\ 0 & k \geq f[m+1] \end{cases} \qquad (II.31)$$

For a smooth and stable spectrum at the output of the filters, the logarithm of the amplitude spectrum is calculated:

$$E[m] = \log[\sum_{k=0}^{N-1} S[k]^2 H_m[k]] \quad 0 \leq m \leq M \qquad (II.32)$$

The mel-frequency cepstral coefficients (MFCC) will be obtained by an Inverse Discrete Cosine Transform (IDCT), which allows for the production of weakly correlated coefficients from the filter output coefficients:

$$c[n] = \sum_{m=0}^{M-1} \mathrm{E[m]} \cos\left[\frac{\pi n}{M}\left(m + \frac{1}{2}\right)\right] \ , 0 \leq n \leq M \qquad (II.33)$$

A dozen MFCC coefficients are generally considered sufficient for speech recognition experiments [15].

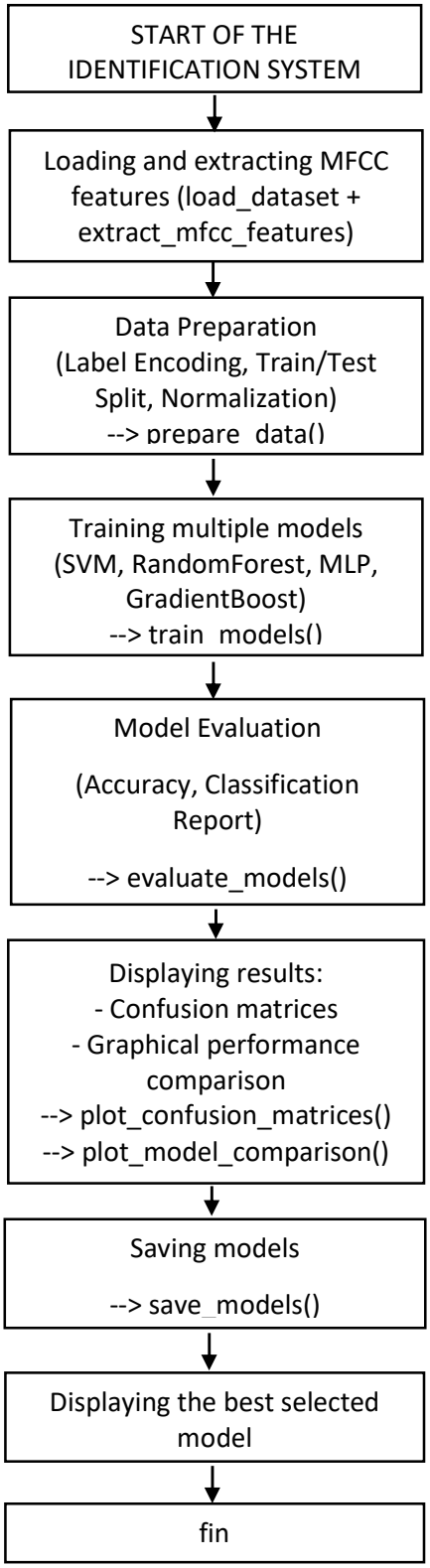


Figure II.5 Organigramme of MFCC

II.2.3.5. PLP (Perceptual Linear Prediction) parameters

Experimental studies led to the concept of a critical band: signals whose frequency falls within a critical band influence the perception of signals located in the same band, but do not influence signals outside this band.

A critical band can be considered a band-pass filter, whose frequency response roughly corresponds to the tuning curve of an auditory nerve fiber.

The LP method uniformly identifies the spectrum across all frequencies in the audible band. However, this property is far from being verified for the human ear, as it has been established that it is more sensitive to frequencies located in the middle of the spectrum analysis band.

Thus, it is possible that some important spectral details of the spectrum are not captured by LP analysis, or that they become significantly important without being physiologically taken into account by the ear.

PLP analysis [15] solves this problem. It allows for the estimation of the parameters of the all-pole autoregressive filter, which best models the auditory spectrum.

The process of calculating PLP coefficients can be described in the following figure:

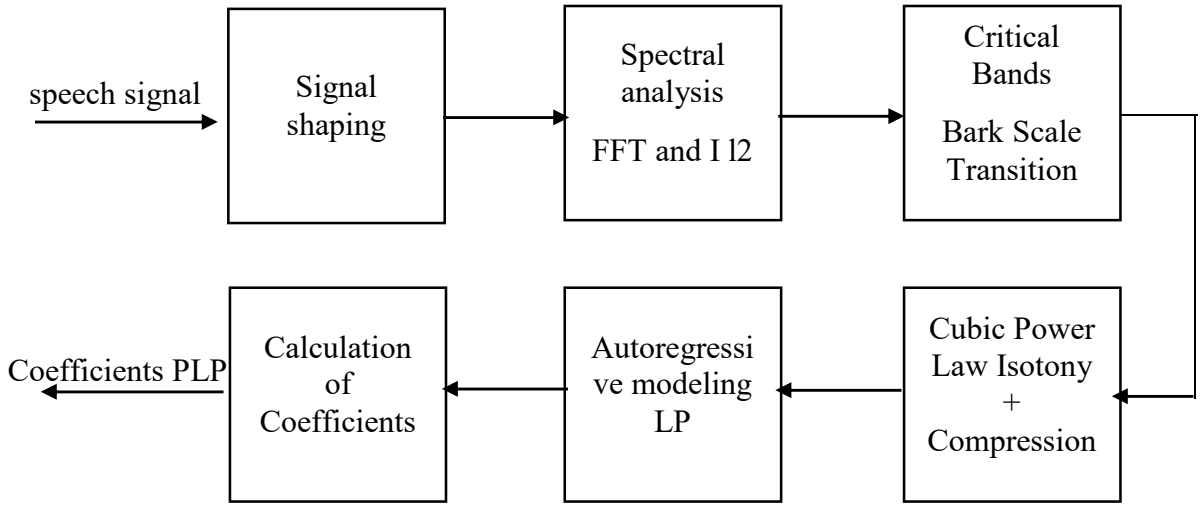


Fig.II.6 The process of calculating PLP coefficients

After shaping the speech signal, the power spectrum P(ω) is calculated.

Then, a transition from the usual frequency scale to the Bark scale is performed using the following relationship:

$$\Omega(\omega) = 6\ln\left(\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi}\right)^2 + 1\right)^{0.5}\right) \qquad (II.34)$$

ω represents the angular frequency expressed in rd/s and ω the Bark frequency.

This conversion to the Bark scale allows us to roughly approximate what we know about the shape of auditory filters. It is approximately constant along the Bark scale. The power spectrum in the Bark scale is convolved with the power spectrum of the critical band curve using the following equation:

$$\psi(\Omega) = \begin{cases} 0 & \Omega < -1.3 \\ 10^{2.5(\Omega+0.5)} & -1.3 \leq \Omega \leq -0.5 \\ 1 & -0.5 \leq \Omega \leq 2.5 \\ 10^{2.5(\Omega-0.5)} & 0.5 \leq \Omega \leq 2.5 \\ 0 & \Omega > 2.5 \end{cases} \quad \text{pour} \qquad (II.35)$$

This masking curve is an approximation of the asymmetric Schroeder masking curve.

We then attempt to approximate the sensitivity of the human ear at different frequencies using a transfer function E(Ω). The power spectrum is multiplied by this transfer function.

$$E(\Omega) = E(\omega) \cdot \Theta(\Omega) \qquad (II..36)$$

$$E(\omega) = [(\omega^2 + 56.8^{*}10)\omega^4]/[(\omega^2 + 6.3^{*}10^6) * (\omega^2 + 0.38 * 10^9)] \qquad (II.37)$$

$$\Theta(\Omega_t) = \sum_{\Omega=-1.3}^{\Omega=2.3} P(\Omega - \Omega_t)\psi(\Omega) \qquad (II.38)$$

The non-linearity between the intensity of a sound and its strength of perception by the ear is then approximated by a power law:

$$\Theta(\Omega) = E(\Omega)^{0.33} \qquad (II.39)$$

The final step consists of a classical autoregressive modeling of the all-pole auditory model spectrum, calculating the autoregressive filter coefficients.

PLP analysis is very similar to MFCC analysis. The difference is that PLP analysis uses the Bark scale instead of the Mel scale and an all-pole autoregressive model instead of the discrete cosine transform (DCT) to calculate the coefficients.

This PLP method was subsequently improved to withstand certain noise conditions. Thus, RASTA-PLP analysis [14] was developed, RASTA being the acronym for RelAtive SpecTrAl.

The PLP method, whose algorithm is based on short-term speech spectra, has difficulty withstanding the constraints that may be imposed on it by the frequency response of a communication channel.

To mitigate the effects of linear spectral distortion, Hermansky proposed modifying the PLP algorithm by replacing the short-term spectrum with an estimated spectrum where each frequency channel is modified by passing it through a filter. This modification is the basis of the RASTA-PLP method.

When implemented in the logarithmic spectral domain, this filtering (RASTA) allows for the removal of constant spectral components, thus eliminating the convolution effects of the communication channel.

II.3 The main steps of automatic speaker recognition

II.3.1 Feature extraction

A signal is an observed measurement of a physical phenomenon. It typically describes an observation of a higher-level physical phenomenon, correlated with lower-level measurement concepts such as time or space. In this context, a speech signal is an observed measurement, made with respect to the passage of time.

To simplify the processing of continuous signals, the infinite set of possible values can be reduced to a finite set through sampling. This action is called discretization, and the redefined signal is called a discrete signal. The latter is capable of transforming the finite set of points into a higher-level measurement. Speech is a time-varying signal that conveys multiple layers of information, including words, speaker identity, acoustic characteristics, language, and emotions. The information contained in speech can be observed in both the time and frequency domains.

The most widely used visualization tool is the spectrogram, in which the frequency spectra of consecutive short-term speech segments are displayed as an image. In the image, the horizontal and vertical dimensions represent time and frequency, respectively, and the intensity of each point in the image indicates the amplitude of a particular frequency at a given time (see Figure II.8).

Spectrograms are very useful for visualizing speech signals, but they are not designed for speech and speaker recognition. There are two reasons for this. First, the frequency dimension is still too high. For a 1024-point Fast Fourier Transform (FFT), the frequency dimension is 512, which is far too large for statistical modeling. Second, the frequency components after the FFT are highly correlated with each other, which makes it difficult to use diagonal covariance matrices to model the variability of feature vectors.

To obtain a more compact representation of speech signals and to decorrelate the feature components, a cepstral representation of speech is often used. In general, a speech signal varies continuously throughout its duration due to articulatory movements. However, dividing it into short overlapping frames can result in speech segments where the signal is considered invariant. In this regard, Mel-Frequency Cepstral Coefficients (MFCC) are commonly used for speech and speaker recognition applications..

II.3.2 Speaker modeling

Using feature vectors extracted from a given speaker's training utterances, a speaker model is trained and stored in the system's database. In text-dependent ARL, the model is utterance-specific, and it includes temporal dependencies between feature vectors. Text-dependent automatic speaker verification and speech recognition share similarities in their model comparison processes. In text-independent ARL, we often model the feature distribution, i.e., the shape of the feature cloud, rather than temporal dependencies. It is important to note that in text-dependent recognition, we can temporally align the test and training utterances because they contain the same phoneme sequences.

However, in text-independent recognition, since there is little or no temporal correspondence between the test and reference frames, frame-level alignment is not possible. Therefore, segmentation of the signal into phonetic classes can be used as a pre-processing step, or speaker models can be phonetically structured. Such approaches have been proposed in [6–7]. It is also possible to use data-driven units instead of strictly linguistic phonemes as segmentation units [8].

Classical speaker models can be divided into template models and stochastic models [9]. In template models, the training and test feature vectors are directly compared with each other under the assumption that one is an imperfect replica of the other. The degree of distortion between them represents their degree of similarity. Vector quantization (VQ) [10] and dynamic time warping (DTW) [11] are representative examples of template models for text-independent and text-dependent recognition, respectively.

II.3.3 Decision and performance evaluation

The last module in the RAL system is to search for the identity of the person producing the defined sequence, in other words, it is a process of comparing an unknown voice with a set of voices from a reference population, and determining whether or not this person belongs to the database.

II.3.3.1 IAL Decision and Performance

In automatic speaker identification in closed set, the processed voice is assumed to be produced by one of the speakers in the database; the identity sought is therefore the one that corresponds to the highest score if it is a similarity measure or to the lowest score if it is a distance measure, these scores are calculated between the processed voice and each speaker in this database. However, in open set identification, the system has no knowledge about the membership of the owner of the processed voice in the database. In this case, the highest/lowest score must be compared with a predefined threshold to decide whether the speaker exists or not in the reference database.

In closed set identification, performance is measured in terms of correct identification rate IC or incorrect identification rate Ii which we obtain by the following formulations

$$I_C = \frac{\text{nombre de test correctement identifié}}{\text{nombre totale de test}} \times 100$$
$$I_i = \frac{\text{nombre de test mal identifié}}{\text{nombre totale de test}} \times 100 \qquad \textit{(II.40)}$$
$$\text{avec: } I_i + I_i = 100\%$$

Whereas in open set, it is the false rejections FR and the false acceptances FA which are measured (Mami, 2003).

II.3.3.2 VAL Decision and Performance

In an automatic speaker verification system, a decision is made to accept or reject the identity of a speaker who is likely to be the source of the voice recording. In other words, it is a process of comparing the voice of a specific person with the sample of their own voice. This method verifies whether the voice of the person proclaiming a known identity matches the reference voice of that person. The performance of a VAL system is evaluated by the following criteria:

II.3.3.2.1 False FAR acceptances and false FRR rejections

To evaluate the performance of a verification system, two types of errors can be made. First, the customer is rejected while the claimed identity is his. In this case, we speak of "False Rejection" (FR). The second error occurs when an imposter is accepted while the claimed identity is not his. This case is called "False Acceptance" (FA) (Figure II.8). A biometric system is evaluated by the rates of FR and FA (in English, False Rejection

Rate, (FRR) and False Acceptance Rate (FAR)). A good verification system minimizes both error rates FAR and FRR.

$$FAR = \frac{\text{Nombre dimposteur acceptées}}{\text{nombre totale dimposteurs}} \times 100$$
$$FRR = \frac{\text{Nombre de client rejetées}}{\text{nombre totale du clients}} \times 100 \qquad (II.41)$$

II.3.3.2.2 Equal Rate Error EER

Most often, RAL system performance is expressed as Equal Error Rate (EER), which is the value where FRR is equal to FAR. The lower the EER value, the better the system. EER is used to compare RAL systems with each other.

II.3.3.2.3 DET Curve or ROC Curve

Each threshold value is associated with a pair (FAR, FRR). All the pairs obtained can be represented in the form of a ROC curve (Oglesby, 1995) or a DET curve (A. Martin, Doddington, Kamm, Ordowski, & Przybocki, 1997) which is the most commonly used representation to evaluate the relevance of the decision threshold as a function of these two error rates (see Figure II.8).

II.3.3.2.4 DCF Decision Cost Function

Since the EER does not differentiate between the two errors (FAR and FRR), which is sometimes not a realistic performance measure. A weighting is then introduced for each of these rates. A Decision Cost Function (DCF), which is a performance measure presented by the National Institute of Standards and Technology (NIST) (A. F. Martin & Greenberg, 2010)

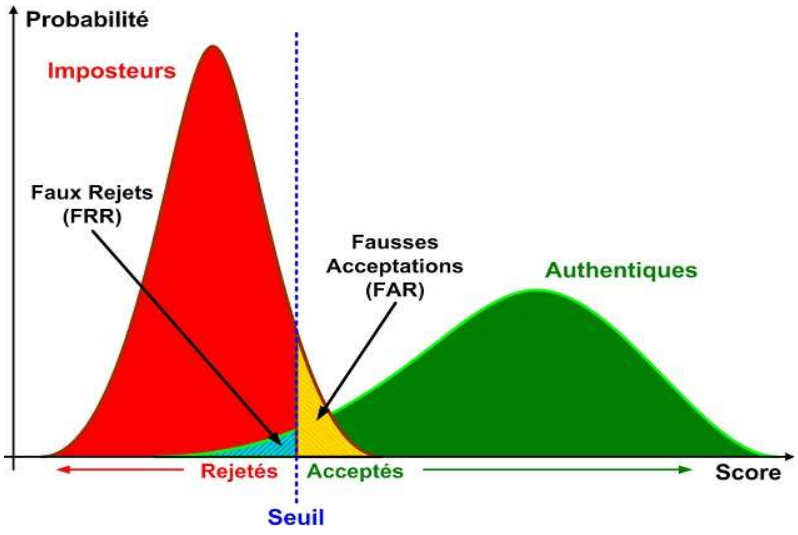


Figure II.8: Illustration of the score distributions and the decision threshold. The areas under the curves with blue and yellow colors represent, respectively, the FRR and FAR (Aloui, Nait-Ali, & Naceur, 2018).

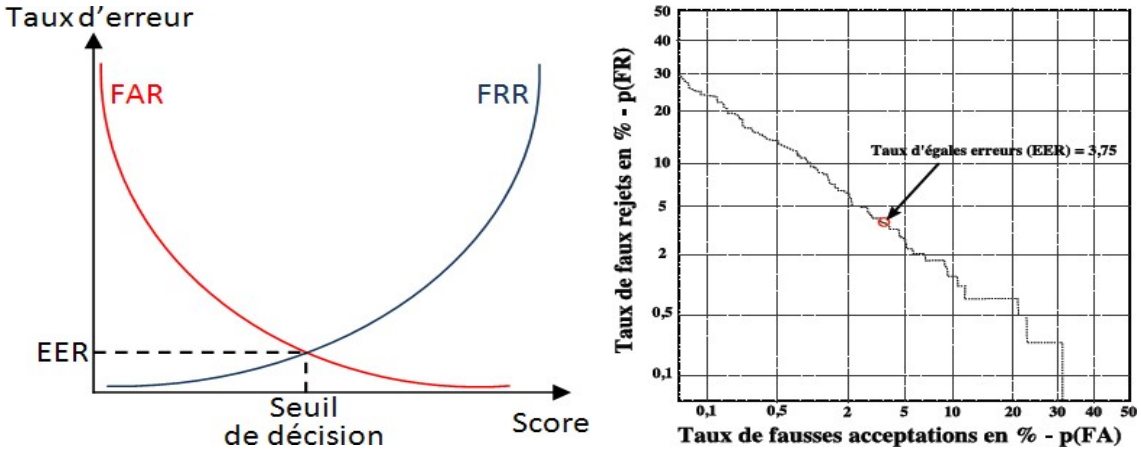


Figure II.9 Example of ROC curve (left) and DET curve (right).

With Pmiss(θ) and Pfa(θ) are the false rejection rate and false acceptance rate respectively as a function of the decision threshold θ, Ptar is the prior probability of a true speaker, Cmiss is the false rejection cost and Cfa is the false acceptance cost. Equation (1.9) is used to calculate the minimum values (MinDCF) proposed by NIST for 2008 SRE (DCF08) with Cmiss = 10, Cfa = 1 and Ptar = 0.01 (A. F. Martin & Greenberg, 2010).

II.3.3.2.5 HTER

Another measure, called HTER (Half Total Error Rate), is defined as the distribution of the average error rate for each Decision threshold (Bengio & Mariéthoz, 2004).

$$HTER = \frac{1}{2(\text{FAR}+\text{FRR})} \qquad (II.42)$$

Error rates are related to the operating point of use. The decision threshold is adjusted a priori on a population of tests. Calibrating this threshold is very important. A variation in the threshold between the calibration and operating phases moves the system away from the desired optimal operating point.

II.4 The different tasks in RAL

Automatic Speaker Identification (AIS) and Automatic Speaker Verification (ASV) [71] are the two most widespread tasks in the field of ALR. Recently, for more specific applications, other tasks have emerged such as speaker indexing, which consists of indicating when each speaker in a conversation has spoken. A related application is the detection of a speaker during a multiple conversation. In this section, we will mainly describe the two main tasks of ALR, the subject of our study: ASI and ASV.

II.4.1 Automatic Speaker Identification (ASI)

Automatic Speaker Identification (ASI) [72] consists of determining, from a set of speakers referenced in the system, the identity of the speaker present in a speech signal (test signal) [73], [74]. To do this, the system calculates similarity measures between this signal and all speaker models in the database. Two identification conditions are known: closed environment and open environment. In the case where the system must provide a set of at least one speaker, this is called closed environment identification.

But in some applications, the system may be required to provide an empty set: this is open environment identification. In a closed environment, each test access is compared to all speaker models referenced in the system. The identity of the speaker with the closest reference is output from the system.

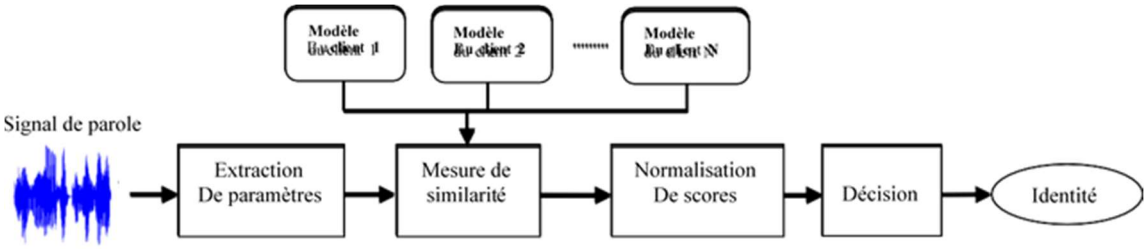


Fig.II.10 Modular diagram of an IAL system

II.4.2 Automatic Speaker Verification (ASV)

An automatic speaker verification (ASV) system must verify, from a speech signal and a claimed identity belonging to the database, whether the presented signal comes from the claimed identity or not [70], [71]. To do this, the system calculates a similarity measure between the produced test signal (claimed identity) and a particular form of the training base (real identity).

If there is a match between the claimed identity and the real identity, we can say that the speaker's identity has been verified. Otherwise, the candidate speaker of the test is an imposter. An automatic speaker verification (ASV) system must verify, from a speech signal and a claimed identity belonging to the database, whether the presented signal comes from the claimed identity or not [70], [71]. To do this, the system calculates a similarity measure between the produced test signal (claimed identity) and a particular form of the training base (real identity).

If there is a match between the claimed identity and the real identity, we can say that the speaker's identity has been verified. Otherwise, the candidate speaker of the test is an imposter.

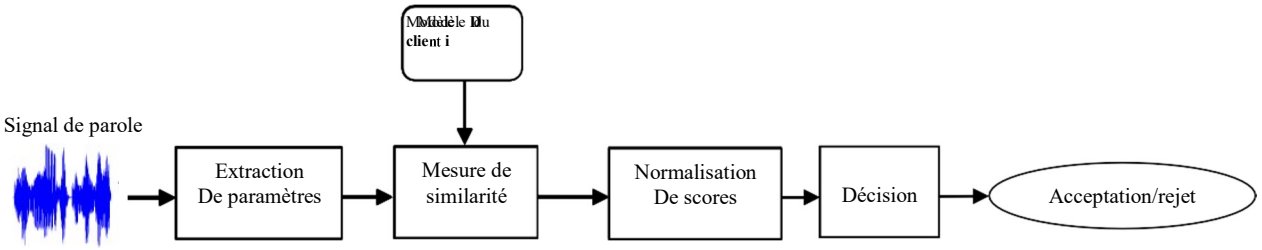


Fig.II.11 Schéma modulaire d'un système de VAL

II.5 Automatic speaker recognition modes

II.5.1 Text-dependent speaker recognition

In text-dependent ADR systems, the text (text) is imposed by the system. Text-based speaker recognition systems were developed first. In this case, at each access, the user will be prompted by the system (e.g., in the form of a synthetic voice or written text) to pronounce a basic vocabulary, which can be very large or simply contain the 10 digits that will be used to create random sequences. The advantage of this approach is that the user cannot predict the sentence he or she will be prompted to pronounce, which makes any recording unusable.

II.5.2 Speaker recognition in text-independent mode

The parameterization process consists of extracting the relevant information from the speech signal for recognition [21], [22]. The speech signal, due to its complexity (multitudes of information and redundancy), cannot be exploited directly. A simplified representation of the speech signal is therefore necessary. This representation is generally based on vectors of acoustic parameters, calculated periodically on the speech signal (see the previous chapter).

II.6 Classical methods of automatic speaker recognition

II.6.1 Methods based on Neural Network

A neural network is a series of algorithms designed to recognize patterns and relationships in data through a process that mimics the way the human brain operates. Let's break this down:

At its core, a neural network consists of neurons, which are the fundamental units akin to brain cells. These neurons receive inputs, process them, and produce an output. They are organized into distinct layers: an Input Layer that receives the data, several Hidden Layers that process this data, and an Output Layer that provides the final decision or prediction The adjustable parameters within these neurons are called weights and biases. As the network learns, these weights and biases are adjusted, determining the strength of input signals. This adjustment process is akin to the network's evolving knowledge base.

Before training starts, certain settings, known as hyperparameters, are tweaked. These determine factors like the speed of learning and the duration of training. They're akin to setting up a machine for optimal performance.

During the training phase, the network is presented with data, makes a prediction based on its current knowledge (weights and biases), and then evaluates the accuracy of its prediction. This evaluation is done using a loss function, which acts as the network's scorekeeper. After making a prediction, the loss function calculates how far off the prediction was from the actual result, and the primary goal of training becomes minimizing this "loss" or error.

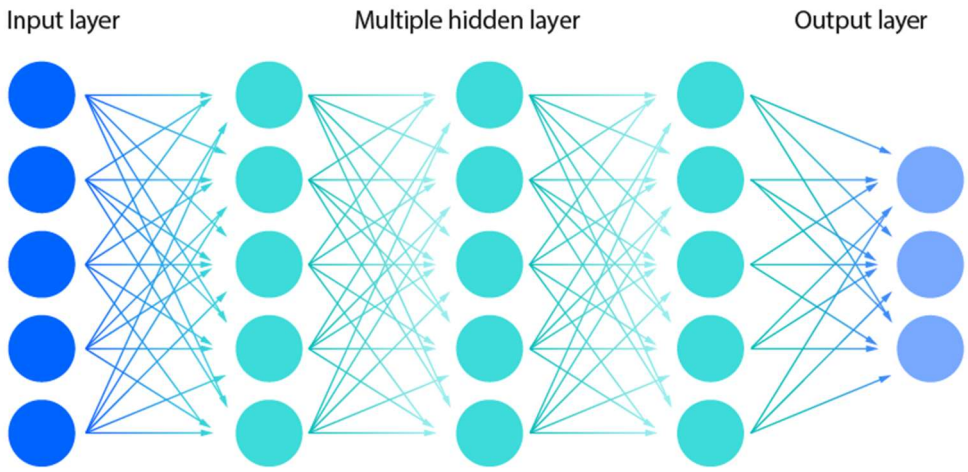


Fig.II.12 Neural Network Methods

II.6.2 Methods based on Gradient Boosting

Gradient Boosting is a ensemble learning method used for classification and regression tasks. It is a boosting algorithm which combine multiple weak learner to create a strong predictive model. It works by sequentially training models where each new model tries to correct the errors made by its predecessor.

In gradient boosting each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent. In each iteration the algorithm computes the gradient of the loss function with respect to predictions and then trains a new weak model to minimize this gradient. Predictions of the new model are then added to the ensemble (all models prediction) and the process is repeated until a stopping criterion is met.

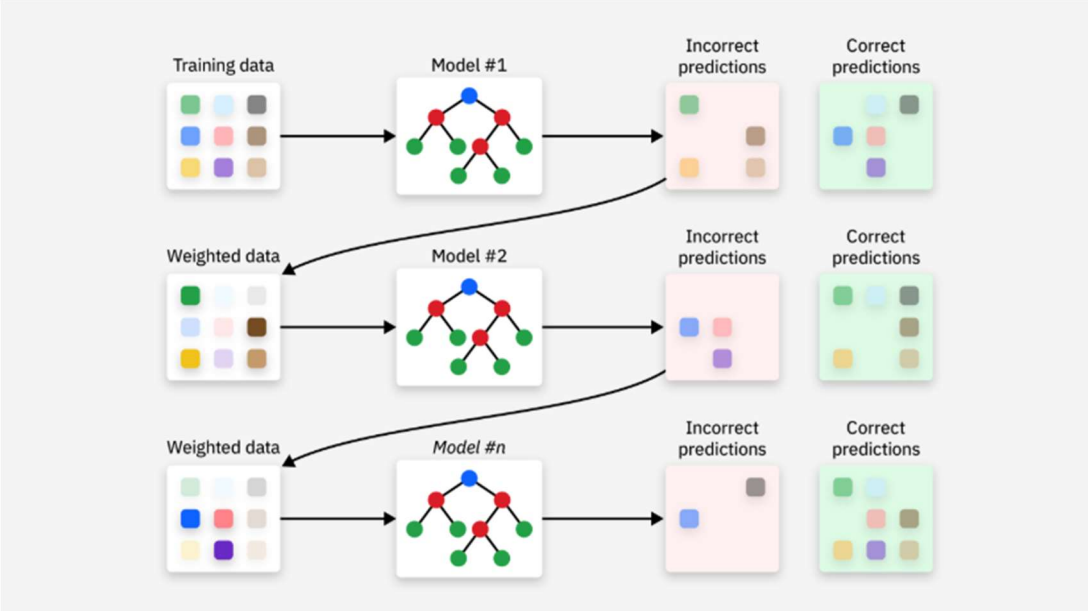


Fig II.13 Methods of Gradient Boosting

II.6.3 Support Vector Machine (SVM) Based Methods

Support vector machines (SVMs) are a set of supervised learning techniques designed to solve discrimination and regression problems [81], [82], [83]. SVMs are a generalization of linear cases. SVMs were developed in the 1990s, based on the theoretical considerations of Vladimir Vapnik [84] on the development of a statistical theory of learning: the Vapnik-Chervonenkis Theory [85]. SVMs were quickly adopted for their ability to work with large-dimensional data, the small number of hyperparameters, their theoretical guarantees, and their good results in practice. SVMs have been applied to many fields (bioinformatics, information retrieval, computer vision, finance, etc.). According to the data, the performance of support vector machines is of the same order, or even superior, to that of a neural network or a Gaussian mixture model.

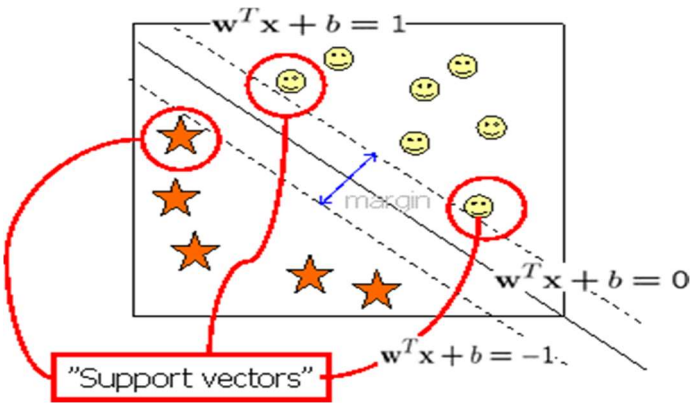


Fig.II.14 Support Vector Machines (SVM)

II.6.4 Methods based on Random Forest

Random forest is a supervised learning algorithm. The "forest" it builds is an ensemble of decision trees, usually trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result. random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

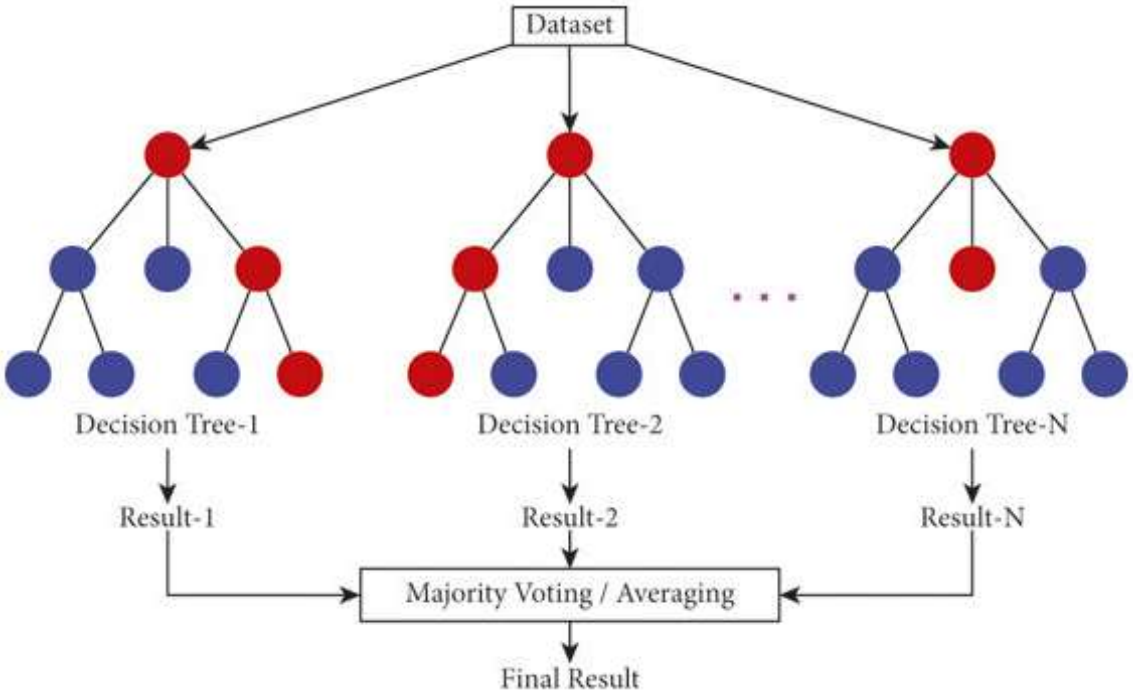


Fig II.15 Methods of Random Forest

II.7 Conclusion

In this chapter we have presented the main tasks of automatic speaker recognition (ASR), such as identification and verification, as well as some application areas of ASR. It is also highlighted in this part, the presentation of an ASR system with all its modules,

namely; the parameter extraction module, the modeling module and the decision-making module. We have focused on the two recognition modes, text-dependent and text-independent. Finally, the methods for evaluating the performance of ASR systems have been carefully described in this chapter.

Chapter III

Experiments and Results

Chapter III
Experiments and Results

III.1 Introduction

In this chapter we will present the series of speaker identification experiments carried out using our database which is composed of 24 speakers (12 male and 12 female) each of whom recorded between 7 and 10 audio files with an average length of 4 seconds. These audio files, numbered from 1 to 10 and which were recorded using voice recorder provided by windows.

These sentences were subjected to a series of experiments to verify and identify the speaker. The results obtained were then examined, discussed, and objective interpretations and conclusions were given.

III.2 The used database

The database It was recorded by 24 people, including 12 men and 12 women who are: (Imad eddine, Fadi, Salah, Adil, Mouhamed, Raouf, Mahdi, El aid, Ihab, Yahia, Amine, Achraf, Karima, Ola, Bouchra, Aya, Abir, Bassma, Nihal, Marwa, Meriam, Feryal, Roumaisa, Houyam). This database contains 240 audio recordings in (.WAV) format of 4 seconds. Each person has 10 recordings.

We use this database in feature extraction method to verify and recognize speaker identity using MFCC and PLP features and learning algorithms.

III.3 Evaluation criteria

III.3.1 Precision

Accuracy is the proportion of all positive classifications in the model that are actually positive. It is defined mathematically as follows:

$$precision = \frac{correctly\ classified\ actual\ positives}{everything\ classified\ as\ positives} = \frac{TP}{TP + FP}$$

In the spam classification example, accuracy measures the fraction of emails classified as spam that were actually spam.

A hypothetical perfect model would have no false positives and therefore have an accuracy of 1.0. In an imbalanced dataset where the number of actual positive results is very, very small (e.g., one to two examples in total), accuracy is less relevant and less useful as a metric.

Precision improves as false positives decrease, while recall improves as false negatives decrease. However, as discussed in the previous section, increasing the classification threshold tends to decrease the number of false positives and increase the number of false negatives, while decreasing the threshold has the opposite effects. Therefore, precision and recall often exhibit an inverse relationship, where improving one degrades the other.

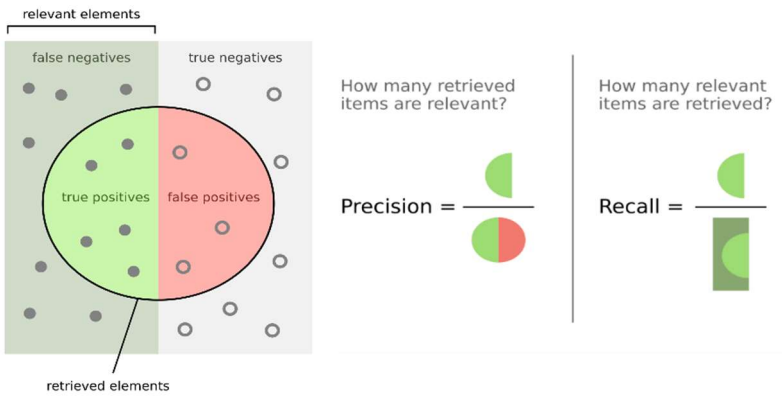
III.3.2 recall

Mathematically, recall is defined as follows:

$$recall = \frac{correctly\ classified\ actual\ positives}{all\ actual\ positives} = \frac{TP}{TP + FN}$$

False negatives are actual positives that were misclassified as negative. This is why they appear in the denominator. In the spam classification example, recall measures the fraction of spam emails correctly classified as spam.

In an imbalanced dataset where the number of true positives is very low, recall is a more relevant metric than precision because it measures the model's ability to correctly identify all positive instances. For applications such as disease prediction, correctly identifying positive cases is critical. A false negative typically has more serious consequences than a false positive. For a concrete example comparing recall and precision metrics, see the notes in the definition of recall.



FigIII.1 precision and recall

III.3.3 accuracy

Accuracy is the proportion of all correct classifications, whether positive or negative. It is mathematically defined as:

$$accuracy = \frac{correct\ classifications}{total\ classifications} = \frac{TP + TN}{TP + TN + FP + FN}$$

In the spam classification example, accuracy measures the fraction of all emails correctly classified.

A perfect model would have no false positives or false negatives, and would therefore have an accuracy of 1.0, or 100%.

Since it incorporates all four results of the confusion matrix (VP, FP, TN, FN), for a balanced dataset with a similar number of examples in both classes, accuracy can serve as a rough measure of model quality. Therefore, it is often the default evaluation metric used for generic or unspecified models performing generic or unspecified tasks.

However, when the dataset is unbalanced or one type of error (FN or FP) is more costly than the other, which is the case in most real-world applications, it is better to optimize for one of the other metrics. For highly imbalanced datasets, where a class appears very rarely, say 1% of the time, a model that predicts a negative value 100% of the time would achieve an accuracy score of 99%, although it would be useless.

III.3.4 F1-score

F1 score is a machine learning evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model.

The accuracy metric computes how many times a model made a correct prediction across the entire dataset. This can be a reliable metric only if the dataset is class-balanced; that is, each class of the dataset has the same number of samples.

Nevertheless, real-world datasets are heavily class-imbalanced, often making this metric unviable. For example, if a binary class dataset has 90 and 10 samples in class-1 and class-2, respectively, a model that only predicts "class-1," regardless of the sample, will still be 90% accurate. Accuracy computes how many times a model made a correct prediction across the entire dataset.

III.4 Experimental work

In this chapter, we will present and analyze the results of a speaker identification and verification system that relies on two common features in the field of audio signal processing: Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). We have used four well-known classifiers and evaluated the accuracy of each of them. The used classifiers are: Support Vector Machine (SVM), Random Decision Forest, Gradient Boosting and Neural Networks.

Each algorithm was trained on a registered database. Performance was measured using standard evaluation indicators such as accuracy, Precision, recall, F1-score.

III.4.1 Experiment-1: Speaker Identification using MFCC

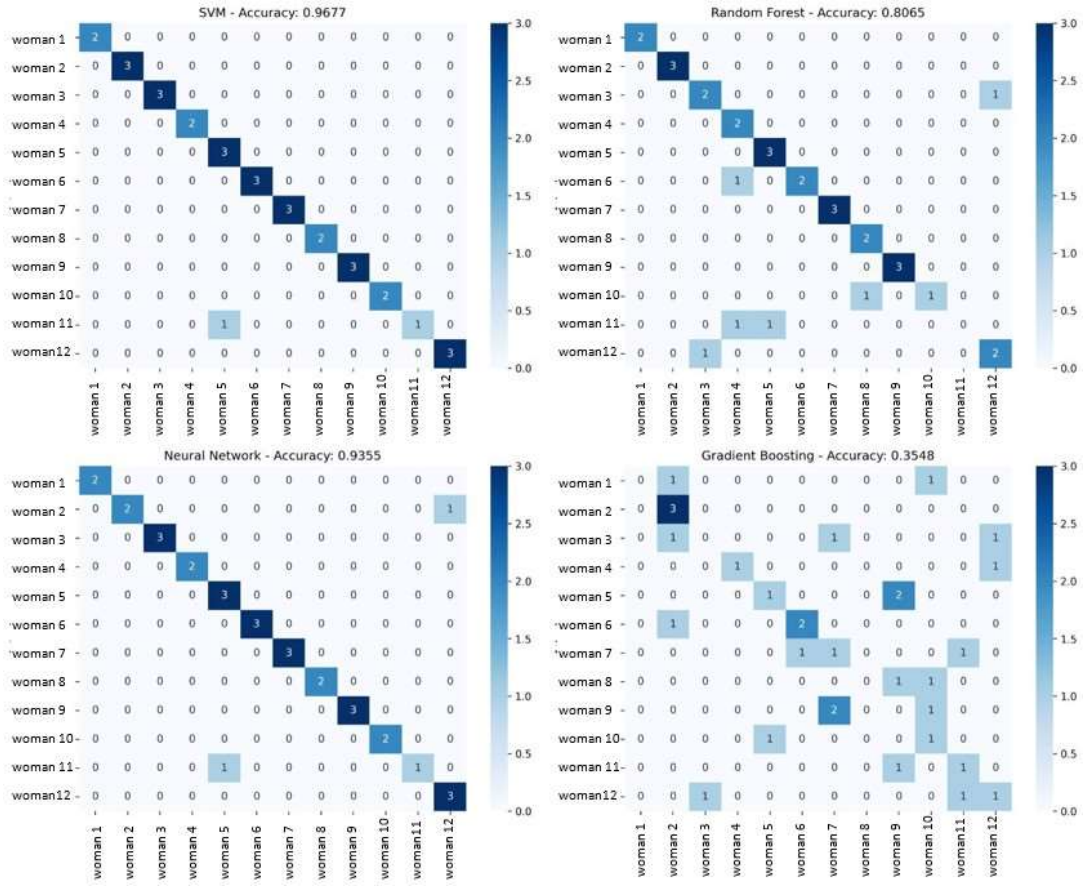
The program was evaluated through a series of experiments using our database. For each experiment, 70% of the data was used for training and 30% for testing. which is usually enough to train and test a model. The results were as follows:

III.4.1.1 women

The database contains 103 audio clips issued by 12 different speakers. In a speaker, there are 10 audio clips, making the database applicable for speaker recognition. Due to its small size, data augmentation and other things such as transfer learning can be used to improve results

Table III.1. Comparative Performance of Classifiers (MFCC Features, Women Speakers)

	Precision	Recall	F1-Score	Support
SVM	0.97	0.94	0.95	31
Random Forest	0.72	0.71	0.70	31
Neural Network	0.92	0.87	0.89	31
Gradient Boosting	0.32	0.29	0.30	31



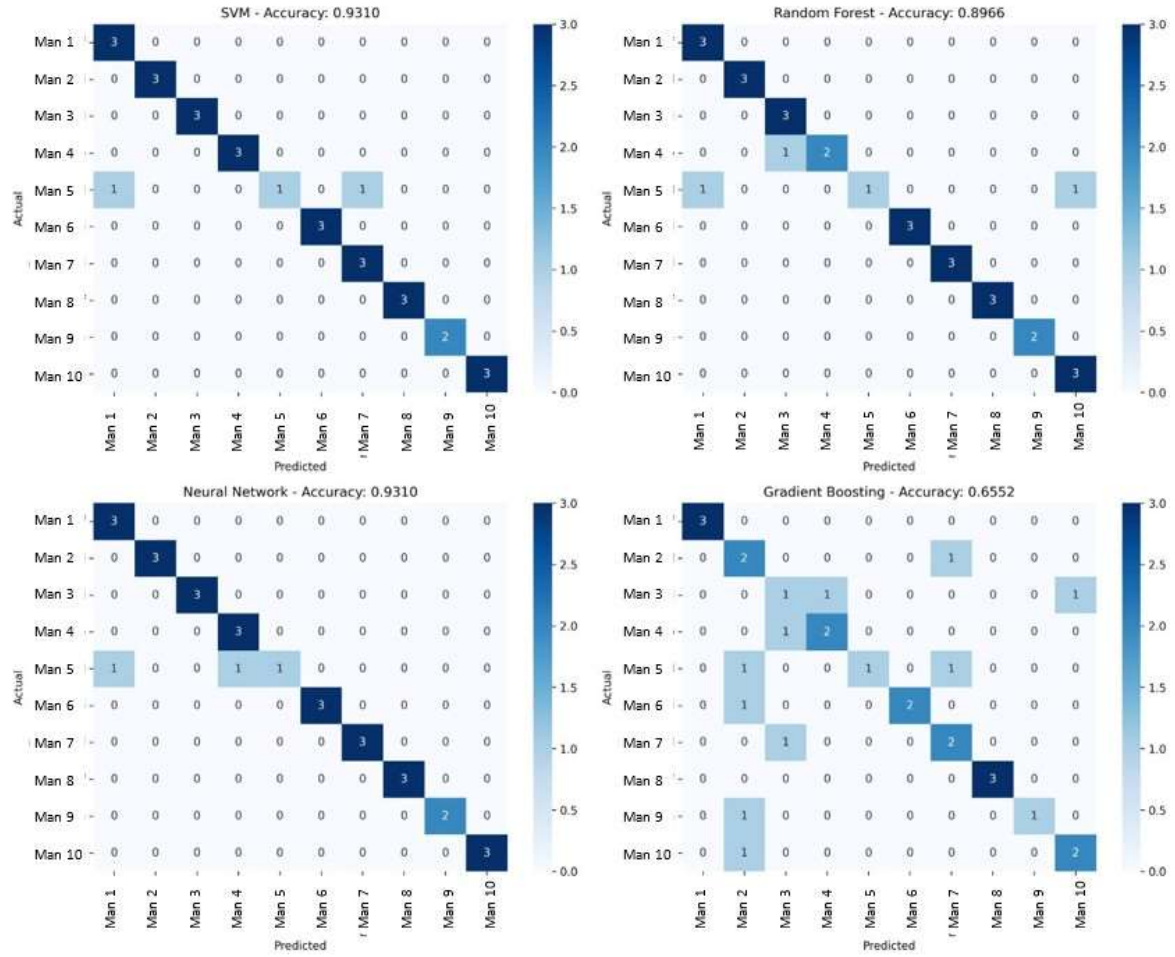
FigIII.2 Confusion matrices (MFCC-Women)

III.4.1.2 Men

The database contains 96 audio clips issued by 10 different speakers. In a speaker, there are 10 audio clips, making the database applicable for speaker recognition. Due to its small size, data augmentation and other things such as transfer learning can be used to improve results

Table III.2. Comparative Performance of Classifiers (MFCC Features, Men Speakers)

	Precision	**Recall**	**F1-Score**	**Support**
SVM	0.95	0.93	0.92	29
Random Forest	0.92	0.90	0.88	29
Neural Network	0.95	0.93	0.92	29
Gradient Boosting	0.75	0.65	0.66	29



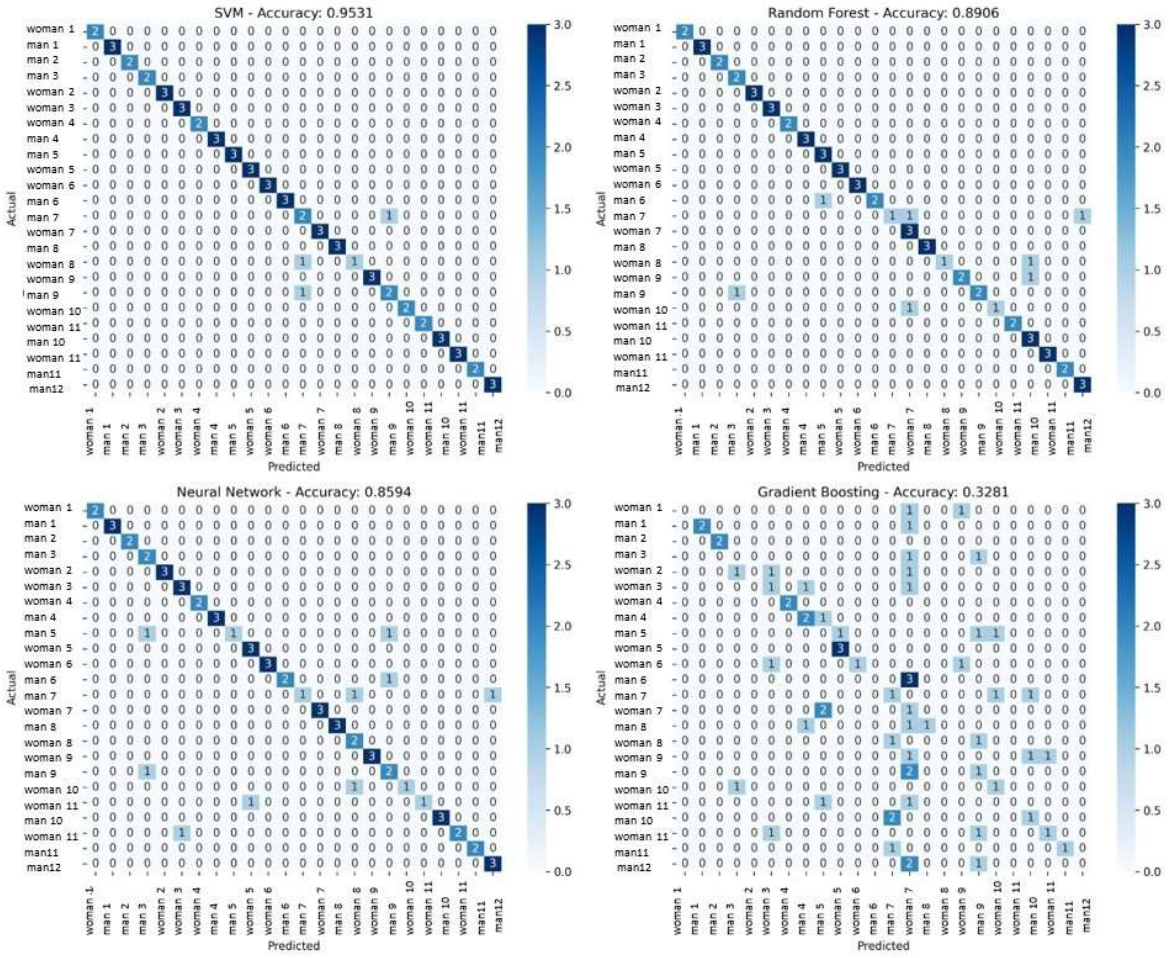
FigIII.3 Confusion matrices (MFCC-Men)

III.4.1.3 Mixed

The database contains 213 audio clips issued by 24 different speakers. In a speaker, there are 10 audio clips, making the database applicable for speaker recognition. Due to its small size, data augmentation and other things such as transfer learning can be used to improve results

Table III.3. Comparative Performance of Classifiers (MFCC Features, Mixed Speakers)

	Precision	Recall	F1-Score	Support
SVM	0.96	0.95	0.95	64
Random Forest	0.92	0.89	0.88	64
Neural Network	0.91	0.85	0.85	64
Gradient Boosting	0.37	0.32	0.32	64



FigIII.4 Confusion matrices (MFCC-Mixed)

III.4.2 Discussion the result of the MFCC

The SVM really stood out on female voices, scoring first across all metrics. This made us realize that SVM is a really good method for a speaker classification using MFCC features. The Neural Network did quite well in second place, so it can be worked on and further experimented with, maybe try out different architectures or different training paradigms. Random Forest is adequate, but Gradient Boosting is a kind of bad that we might not want to think about using on this kind of data.

For male voices, the SVM and NN models performed superbly and practically at the same level: both did great in their respective fields. Then came the Random Forest model, which was good at profiling male speakers, especially better than females for the model, it almost seemed that male voice features are easier to pick. The Gradient Boosting was slightly better in the male task rather than the female, but it remained the worst out of the four.

The single SVM was definitely the top candidate for the combined dataset of male and female speakers, proving its versatility with the here-occurring different speaker traits. Random Forest and Neural Network gave slightly poorer results than in the other experiments, still acceptable to a certain extent. However, Gradient Boosting continued to perform exceedingly poorly, adding weight to the idea that the algorithm simply does not fit well with speaker classification using MFCCs.

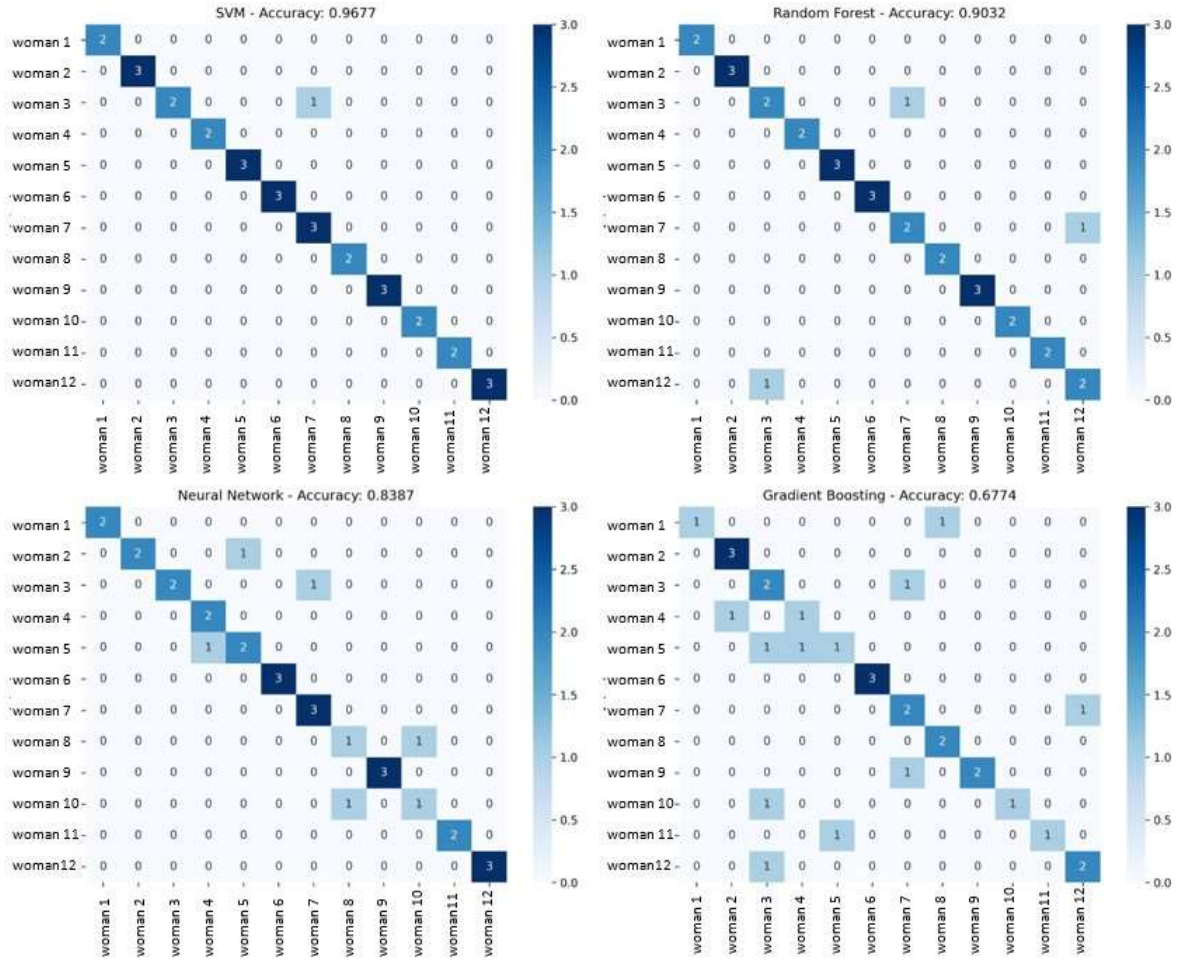
III.4.3 Experiment-2: Speaker Identification using PLP

III.4.3.1 Women

The database contains 103 audio clips issued by 12 different speakers. In a speaker, there are 10 audio clips, making the database applicable for speaker recognition. Due to its small size, data augmentation and other things such as transfer learning can be used to improve results

Table III.4. Comparative Performance of Classifiers (PLP Features, Women Speakers)

	Precision	**Recall**	**F1-score**	**Support**
SVM	0.97	0.97	0.97	31
Random Forest	0.95	0.95	0.95	31
Neural Network	0.87	0.91	0.88	31
Gradient Boosting	0.70	0.75	0.72	31



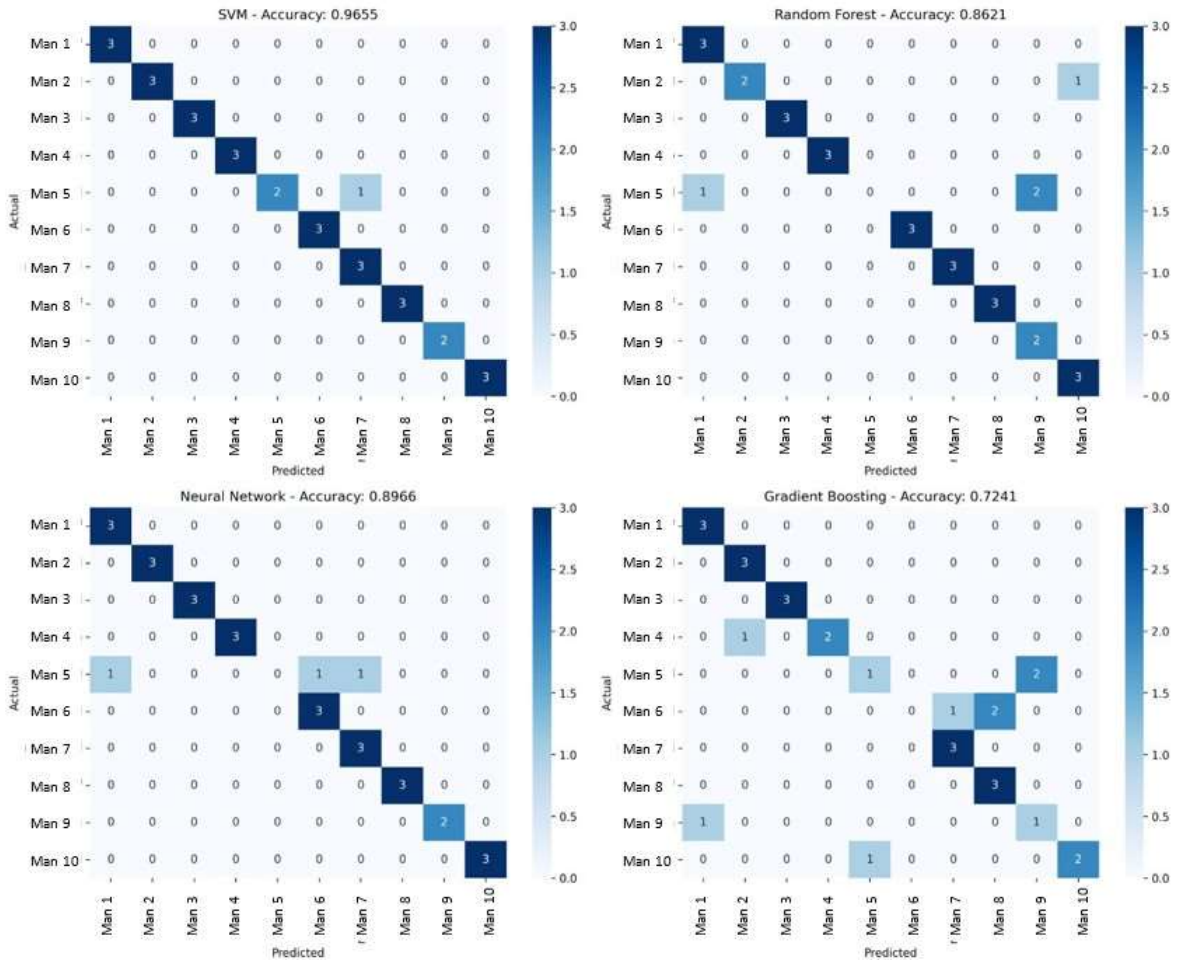
FigIII.5 Confusion matrices (PLP-Women)

III.4.3.2 Men

The database contains 96 audio clips issued by 10 different speakers. In a speaker, there are 10 audio clips, making the database applicable for speaker recognition. Due to its small size, data augmentation and other things such as transfer learning can be used to improve results

Table III.5. Comparative Performance of Classifiers (PLP Features, Men Speakers)

	Precision	**Recall**	**F1-score**	**Support**
SVM	0.97	0.96	0.96	29
Neural Network	0.82	0.91	0.85	29
Random Forest	0.80	0.87	0.81	29
Gradient Boosting	0.63	0.81	0.77	29



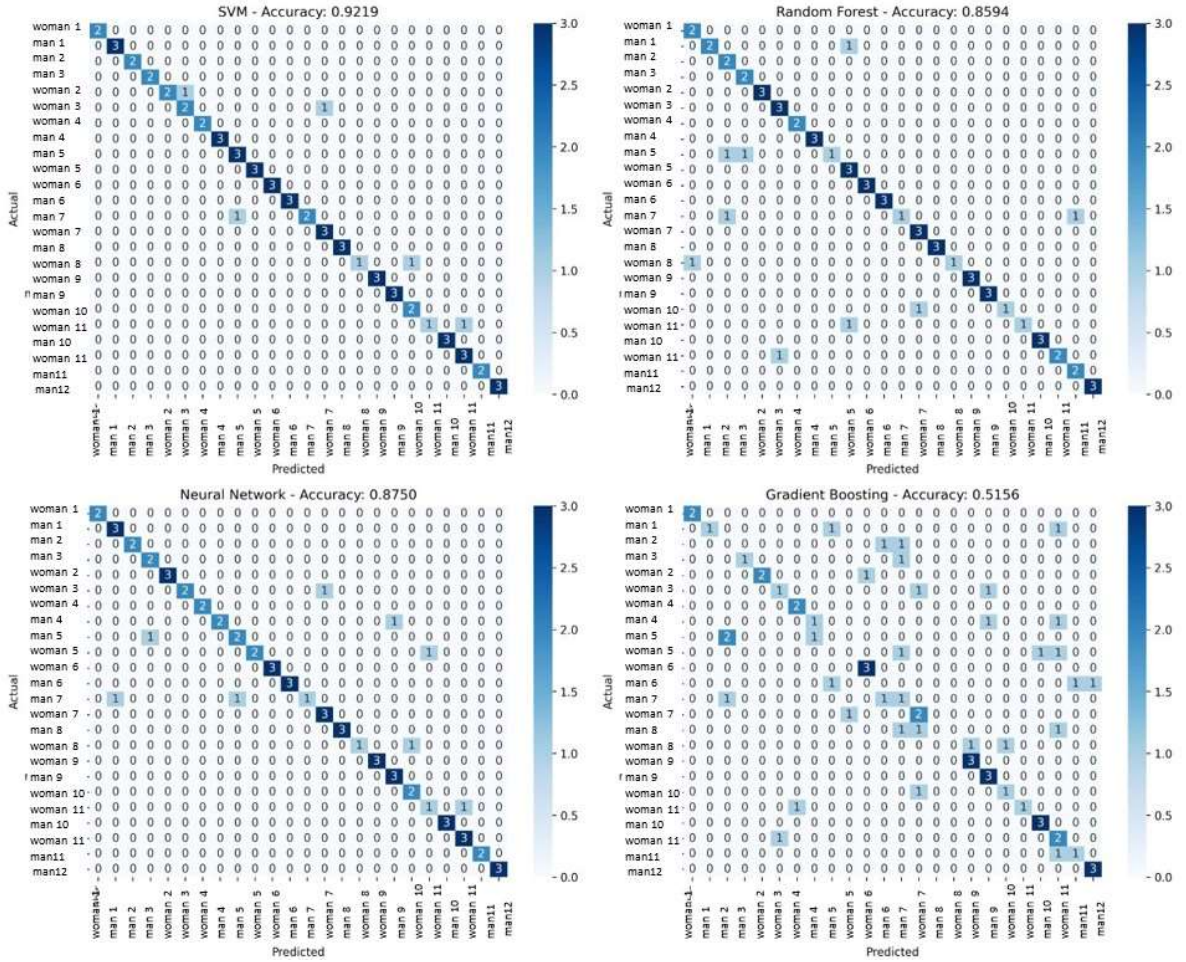
FigIII.6 Confusion matrices (PLP-Men)

III.4.3.3 Mixed

The database contains 213 audio clips issued by 24 different speakers. In a speaker, there are 10 audio clips, making the database applicable for speaker recognition. Due to its small size, data augmentation and other things such as transfer learning can be used to improve results

Table III.6. Comparative Performance of Classifiers (PLP Features, Mixed Speakers)

	Precision	**Recall**	**F1-Score**	**Support**
SVM	0.93	0..93	0.93	64
Random Forest	0.88	0.88	0.88	64
Neural Network	0.90	0.90	0.90	64
Gradient Boosting	0.62	0.65	0.60	64



FigIII.7 Confusion matrices (PLP-Mixed)

III.4.4 Discussion the result of the PLP

Considering the fact that Support Vector Machine (SVM) obtained maximum performance in all metrics is further proof that the classifier can very well separate female speakers given PLP features. The Random Forest classifier also did very well and was only a bit behind the SVM. The Neural Network gave an acceptable result, especially in recall, meaning that the Network found a good proportion of female samples, but the precision slightly decreased. Gradient Boosting yielded the worst results but still performed a little better than the experiments with MFCCs, which means it was a little better suited to PLP features.

Once more, the SVM model performed very well for male speakers, indicating that it generalizes across genders. It is important to point out that the Neural Network had higher recall and F1-score than Random Forest, identifying a very high percentage of male speakers, albeit at a moderate precision. Gradient Boosting appear to have improved recall

in comparison to the other cases but still performed worse than all of them in terms of overall balance and precision, which might indicate a tendency to generate false positives.

SVM continued to outperform all the other models, validating the choice of this model as a classifier of speakers in real-world scenarios. Achieving results similar to Random Forest, the Neural Network was in second place, a narrow margin behind the SVM. Gradient Boosting showed the worst results, even though recall was improved compared to the female only population case. However, the instability and low F1-score of the model show that it is not advisable to classify speakers with PLP features using this model.

III.4.5 Comparison between MFCC and PLP

The experiment provides clear evidence of the differences between Mel-Frequency Cepstral coefficients (MFCC) and Perceptual Linear Prediction (PLP) as far as speaker identification performance is concerned. When MFCC is given to SVM classifiers, it becomes impossible for PLP to achieve any better performance as the F1-score achieved by the MFCC-SVM system is 0.95 and its precision on the mixed dataset 95%. However, PLP features work more consistently with other classifiers. For example, for female speakers, Random Forest attained F1-score of 0.95 based on PLP as against 0.70 based on MFCC.

However, Neural Network also achieved better results with PLP, scoring F1-scores of 0.88 (female) and 0.90 (mixed), in contrast to 0.89 and 0.85 with MFCC. Despite the Gradient Boosting being the lowest performer overall, it did slightly better using PLP (F1-score of 0.60 on the mixed dataset) than using MFCC (0.32). These results indicate that while PLP means a more balanced performance with Random Forest and Neural Network, SVM is the most reliable and consistent model irrespective of feature type and hence should be the choice of robust speaker identification systems.

III.5 conclusion

The comparison between MFCC and PLP features suggested some significant variations in the performance of the classification models used in the speaker recognition system. The Support Vector Machine (SVM) displayed better and consistent performance with both MFCC and PLP features by obtaining the best precision, recall, and F1-scores in all datasets (female, male, and mixed).

This confirms its capability to work with high-dimensional data and find optimal decision boundaries in a non-linear fashion. Contrarily, the Random Forest algorithm showed better results with PLP features, especially in the female speaker dataset, implying that PLP has a more balanced and robust performance than MFCC. The Artificial Neural Network also performed better with PLP features, in particular for male speakers, where it obtained better recall and F1-scores.

This improvement can be due to the fact that the perceptually motivated characteristics of PLP suit the learning behavior of neural networks. On the other hand, Gradient Boosting was the worst model overall, although the results improved for the PLP features compared to MFCC. However, its results were still a lot lower when compared to other classifiers.

Based on these observations, it can be concluded that PLP features are more consistent when used with models like Random Forest and Neural Networks. However, SVM still remains the best and most consistent model with either of the features. Therefore, it is recommended to use SVM with either PLP or MFCC in speaker recognition systems, with neural networks as a strong alternative, while Gradient Boosting should generally be avoided unless further enhancements are made to the feature representation or preprocessing pipeline.

Conclusion

Conclusion

In conclusion, this research addressed a timely and interesting topic in the field of biometric security, specifically in speaker recognition using the voice as a biometric. Using two classical acoustic feature extraction methods, namely Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP), and four machine learning algorithms-Support Vector Machine, Random Forest, Neural Network, and Gradient Boosting-we performed experiments on a custom-built speech database.

Our results reveal that Support Vector Machine has always overseen the performance of all other machine learning models concerning all evaluation metrics and having precision, then recall, and F1-score; hence, this highlights the capability of the support vector machine to deal with very high-dimensional and non-linear data. Neural Networks and Random Forests had acceptable performances, mainly when coupled with PLP features, whereas Gradient Boosting presented relatively weak results and hence limiting itself as a less favorable choice for this particular arena.

As perspectives, the research confirms that the human voice represents a trustworthy biometric whenever researchers analyze and model its characteristics. System performance and practical application development have multiple exciting research directions for improvement. Researchers should expand their speaker count while incorporating diverse acoustic settings into their dataset and apply data augmentation techniques for model generalization and test various powerful deep learning architectures. The system's robustness under noisy conditions should be tested and voice recognition should be integrated with other biometric approaches to achieve better multi-modal biometric systems that offer improved accuracy and user convenience and higher security levels. The study's findings advance speaker recognition technology which enables future security solutions to integrate this technology.

References

References

[1] Souhail Guennouni, Anass Mansouri and Ali Ahaitouf (March 1st 2019). Biometric Systems and Their Applications, Visual Impairment and Blindness - What We Know and What We Have to Know, Giuseppe Lo Giudice and Angel Catalá, IntechOpen, DOI: 10.5772/intechopen.84845. Available from: https://www.intechopen.com/books/visualimpairment-and-blindness-what-we-know-and-what-we-have-to-know/biometric-systemsand-their-applications

[2] https://www.elprocus.com/different-types-biometric-sensors/

[3] Aleksandra, Babich. "Biometric Authentication. Types of biometric identifiers Bachelor's Thesis: Degree Programme in Business Information Technology. Finland: HAAGA-HELIA University of Applied Sciences. 2012, 53 p.

[4]. Sayah, Manel. "Système de pointage par empreinte digitale". Master thesis: embedded system. Biskra: Mohamed Khider University of Biskra, 2019, 98 p

[5] Steven Davis and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: IEEE transactions we acoustics, speech, and signal processing 28.4 (1980), pp. 357–366.

[6] Matthieu Hébert and Larry P Heck. "Phonetic class-based speaker verification". In: Eighth European Conference we Speech Communication and Technology . 2003.

[7] Robert Faltlhauser and Günther Ruske. "Improving speaker recognition using phonetically structured Gaussian mixture models". In: Seventh European Conference we Speech Communication and Technology . 2001.

[8] Asmaa The Hannani, Dijana Petrovska-Delacrétaz, and Gerard Chollet. "Linear and non-linear merger of ALISP-based and GMM systems for text-independent speaker verification". In: ODYSSEY04-The Speaker and Language Recognition Workshop . 2004.

[9] Joseph P Campbell. "Speaker recognition: A tutorial". In: Proceedings of the IEEE 85.9 (1997), pp. 1437–1462.

[10] Frank K Soong et al. "Report: A vector quantization approach to speaker recognition". In: AT&T technical journal 66.2 (1987), pp. 14–26.

[11] Sadaoki Furui. "Comparison of speaker recognition methods using statistical features and dynamic features". In: IEEE Transactions we Acoustics, Speech, and Signal Processing 29.3 (1981), pp. 342–350.

[12] Amélioration de la Robustesse des Systèmes de Reconnaissance Automatique du Locuteur par FEDILA Meriem Université des Sciences et de la Technologie Houari Boumediene

1/Steven Davis and Paul Mermelstein. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences". In: IEEE transactions on acoustics, speech, and signal processing 28.4 (1980), pp. 357–366.

[13] Reynolds, D. (2002). An overview of automatic speaker recognition. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)(S. 4072-4075).

[14] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech communication, 17(1), 91-108.

[15] Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. Speech and Audio Processing, IEEE Transactions on, 3(1), 72-83.

[16] Atal, B. S. (1976). Automatic recognition of speakers from their voices.Proceedings of the IEEE, 64(4), 460-475.

[17] Doddington, G. R. (1985). Speaker recognition—Identifying people by their voices. Proceedings of the IEEE, 73(11), 1651-1664.

[18] https://www.datacamp.com/blog/what-are-neural-networks

[19] https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/

[20] Identification et Authentification de Locuteurs, par les Techniques de Fusion des Paramètres et des Modèles dans un Environnement Réel par asbai nassim USTHB

[21] https://builtin.com/data-science/random-forest-algorithm

ملخـــــص

تهدف هذه الدراسة إلى تطوير نظام تحقق وتحديد للمتحدثين بتقنية التعرف الذكي على المتحدثين، بالاعتماد على خوارزميات MFCC وPLP، إلى جانب نماذج التعلم الآلي مثل SVM والغابات العشوائية والشبكات العصبية. ثم اختُبر النظام على قاعدة بيانات تضم 24 متحدثًا، حيث أظهرSVM ، متبوعًا بالشبكات العصبية والغابات العشوائية، أفضل النتائج مع ميزاتPLP ، بينما أظهر Gradient Boosting نتائج ضعيفة. توصي الدراسة بزيادة قاعدة البيانات، وتطبيق تقنيات تعزيز البيانات، واختبار النماذج في سيناريوهات واقعية، ودمج الصوت مع المقاييس الحيوية الأخرى لتعزيز الأمان.

الكلمات المفتاحية : المعاملات القيصرية للترددات(م. ميل)، التنبؤ الخطي الادراكي، آلة المتجهات الداعمة. MFCC.PLP.SVM

Abstract

This study aims to develop a verification and identification system for speakers with intelligent speaker recognition, by relying on MFCC and PLP algorithms coupled with ML models like SVM, Random Forest, and Neural Networks. The system was then tested on a database of 24 speakers, where SVM, followed by Neural Networks and Random Forests, showed best results with PLP features, while Gradient Boosting showed poor results. The study recommends increasing the database, implementing data augmentation techniques, testing the models in real-life scenarios, and combining voice with other biometrics for enhanced security.

Keywords: Mel Frequency Cepstral Coefficients, Perceptual Linear Prediction. Support Vector Machine. MFCC. PLP. SVM.

Résumé

Cette étude vise à développer un système de vérification et d'identification des locuteurs avec reconnaissance intelligente, en s'appuyant sur des algorithmes MFCC et PLP couplés à des modèles d'apprentissage automatique (ML) tels que SVM, Random Forest et les réseaux de neurones. Le système a ensuite été testé sur une base de données de 24 locuteurs. SVM, suivi des réseaux de neurones et des forêts aléatoires, a montré les meilleurs résultats avec les fonctionnalités PLP, tandis que le Gradient Boosting a donné de mauvais résultats. L'étude recommande d'enrichir la base de données, de mettre en œuvre des techniques d'augmentation des données, de tester les modèles en situation réelle et de combiner la voix à d'autres données biométriques pour une sécurité renforcée.

Mots clés : Coefficients cepstraux de fréquence Mel, Prédiction linéaire perceptuelle, Machine à Vecteurs de Support, **MFCC, PLP, SVM.**