T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Electrical and Computer Engineering

SPEAKER VOICE RECOGNITION USING A HYBRID PSO/FUZZY LOGIC SYSTEM

Nuha ALTEKREETI

Master Thesis

Supervisor

Asst.Prof.Dr.Abdullhai Abdu IBRAHIM

Istanbul, 2021

SPEAKER VOICE RECOGNITION USING A HYBRID PSO/FUZZY LOGIC SYSTEM

by

Nuha Waleed Dahham ALTEKREETI

Electrical and Computer Engineering

Information Technologies

Submitted to the Graduate School of Science and Engineering

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2021

The thesis titled “SPEAKER VOICE RECOGNITION USING A HYBRID PSO/FUZZY LOGIC SYSTEM” prepared and presented by Nuha Waleed Dahham ALTEKREETI was accepted as a Master of Science Thesis in Civil Engineering.

_____

Asst. Prof. Dr.Abdullahi Abdu IBRAHIM

Supervisor

Thesis Defense Jury Members:

Asst.Prof.Dr. Abdullahi Abdu IBRAHIM	School of Engineering , And Nature science Altınbaş University	_____
Asst. Prof.Dr.Muhammad ILYAS	School of Engineering , And Nature science Altınbaş University	_____
Asst. Prof. Dr. Tareq MOHAMMED	Computer Science Imam Jaafar AL-Sadiq University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies:

____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Nuha Waleed Dahham ALTEKREETI

DEDICATION

I would like to thank God Almighty for the strength of mind, health, strength, guidance, knowledge and skills to complete this study.

This thesis is sincerely dedicated to my mother. There are no words to describe what it means to me; There is nothing I can repay for what you did to me. I will continue to do my best to fulfill your expectations.

Finally, I dedicated this to my brothers, sisters, relatives and friends who have been encouraging me during this study..

ACKNOWLEDGEMENTS

I would like to thank my supervisor Asst.prof.Dr.Abdullahi Abdu IBRAHIM for all support during my study. It's great pleasure to express my deepest gratitude to my friends who have shared with me best moments during my study for the Master degree.

ABSTRACT

SPEAKER VOICE RECOGNITION USING A HYBRID PSO/FUZZY LOGIC SYSTEM

Nuha Waleed Dahham ALTEKREETI

M.Sc., Electrical and Computer Engineering, Altınbaş University

Supervisor: Asst. Prof. Dr. Abdullahi Abdu IBRAHIM

Date: June 2021

Pages:54

Smart grids are electric grids that are composed of multiple power sources and devices connected to each other to provide better reliability in power generation and power management, modern developments of the smart grid aim at either improving the control of power sources and loads connected to the smart grid by developing a specialized software/hardware, or by improving the communication within the parts of the smart grid and the central control. In this paper we aim at improving both sides of the smart grid system (communication and control), we propose a fuzzy logic based controller for renewable energy and fossil fuel sources in a grid and an internet of things based monitoring system which oversees the state of the smart grid, faults that occur in the grid, and how the fuzzy controller overcomes those faults, all in which provide an extra layer of support to the smart grid.

Keywords: Fuzzy logic, Internet of things IOT, Matlab Simulink

ÖZET

SPEAKER VOICE RECOGNITION USING A HYBRID PSO/FUZZY LOGIC SYSTEM

Nuha Waleed Dahham ALTEKREETI

M.Sc., Electrical and Computer Engineering, Altınbaş University

Supervisor: Asst.Prof. Dr. Abdullahi Abdu IBRAHIM

Date: June 2021

Pages:54

Akıllı şebekeler, güç üretimi ve güç yönetiminde daha iyi güvenilirlik sağlamak için birbirine bağlı birden fazla güç kaynağı ve cihazdan oluşan elektrik şebekeleridir; akıllı şebekenin modern gelişmeleri, güç kaynaklarının kontrolünü ve akıllıya bağlı yükleri kontrol etmeyi amaçlamaktadır. özel bir yazılım / donanım geliştirerek veya akıllı şebekenin bileşenleri ile merkezi kontrol arasındaki iletişimi geliştirerek. Bu makalede akıllı şebeke sisteminin (iletişim ve kontrol) her iki tarafını da geliştirmeyi hedefliyoruz, akıllı bir şebekede yenilenebilir enerji ve fosil yakıt kaynakları ve devleti denetleyen şeylere dayalı bir izleme sistemi için bulanık mantık tabanlı bir kontrolör öneriyoruz akıllı şebekeye, akıllı şebekede meydana gelen arızalara ve bulanık denetleyicinin bu hataların üstesinden nasıl geldiği, bunların hepsi de akıllı şebekeye ekstra destek katmanı sağlar.

Anahtar Kelimeler: Bulanik Mantik , Nesnelerin İnterneti , Matlab Simulink

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

1. INTRODUCTION

1.1 BACKGROUND

The importance of security and the identification of people is growing in today's society, giving rise to several methods to meet this need. One way to meet this need is using biometrics to identify a person by their unique characteristics. the biometrics measures an individual's physiological or behavioral characteristics to identify them it, for example, the characteristics of a person's face (physiological characteristic) or the way a person walks (behavioral) :

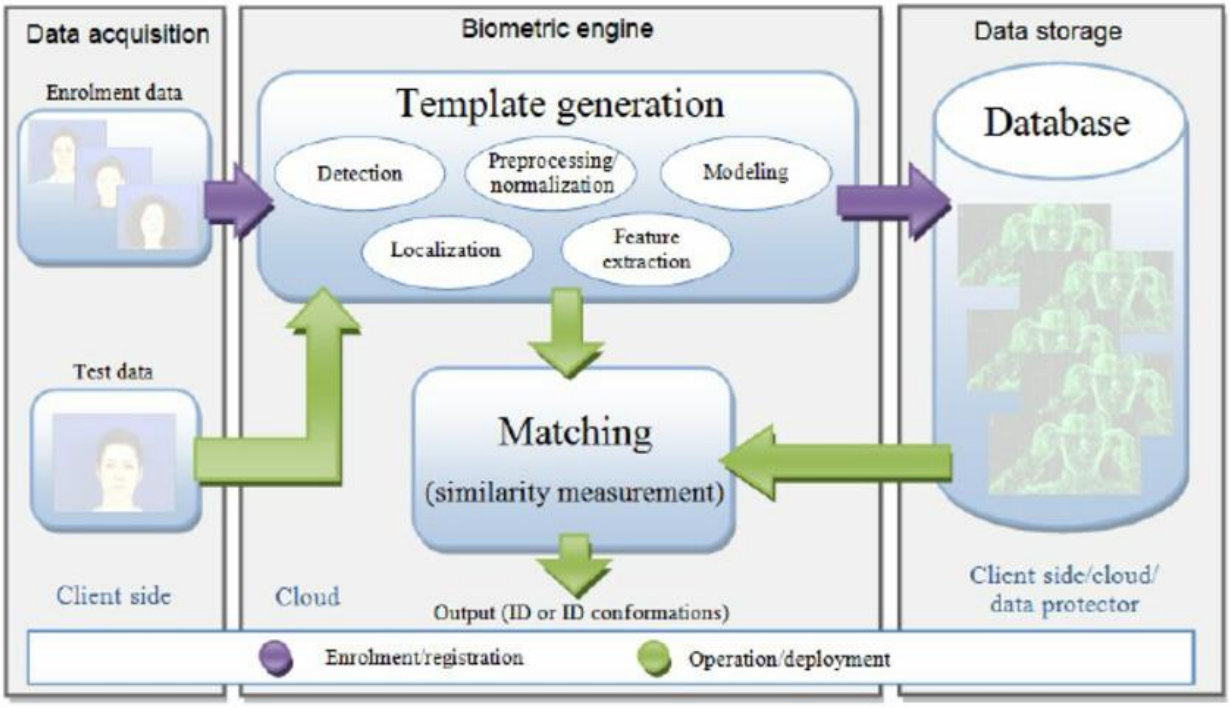

Figure 1.1: Simple biometrics recognition system [1].

1.2 MOTIVATION

There are types of biometrics that are often used for security, like digital, iris and voice. Digital and iris fit as a biometric of physiological characteristic, where it comes of something immutable. In the case of the voice, biometrics of the physiological characteristics and also of the behavior, enabling alternative methods, since speech offers a lot of flexibility and different levels to perform recognition. For example, the system can force the user to speak about a particular and different way for each attempt [1]. Another advantage of voice biometrics is that it

can be used at a distance, unlike other biometrics, an example is when you want to authenticate or identify someone via phone. Each person's voice has unique characteristics like a fingerprint, due to the formation larynx, vocal cords and the entire system responsible for the voice. In addition to physiological differences, there are also differences in the way people speak, such as accent, rhythm, intonation, pronunciation, choice of vocabulary and so on [2] Based on these characteristics it is possible to recognize an individual, that is, recognize the speaker.

1.3 PROBLEM STATEMENT

Speaker recognition is commonly used in the field of verification or identification, but it can also be used in other branches, such as classification, detection, segmentation, etc. The purpose of verification is to find out if the person speaking is really who they claim to be, if the system must be prepared for possible attempts at deception. In the case of identification the goal is to distinguish a person from a group of previously registered people [3] Currently the branch of verification is the most popular for speaker recognition due to its importance in security and access control [4] so it will be the branch used in that job:

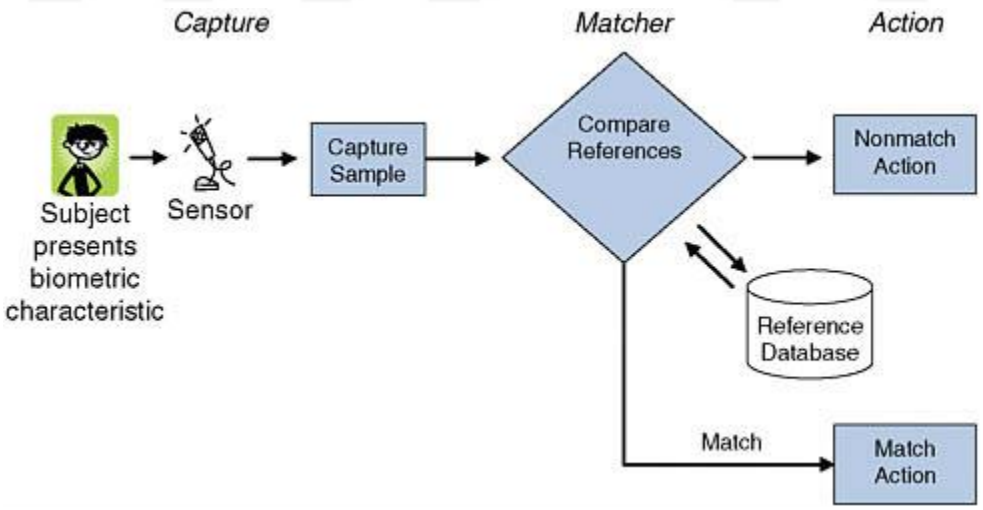

Figure 1.2: Traditional biromantics recognition system [3]

1.4 CONTRIBUTION

In this work we have designed and implemented a system whose main purpose is to detect the signals of the human voice and recognize the identity of the speaker, we implement our scheme using machine learning techniques where we extract the features of the speaker's audio file and

then classify it into labels. In our work, we implement the algorithms on four individual voice records from famous personals, the reason behind choosing these four voices is because they have different sound attributes such as frequency and pitch.

1.5 THESIS ORGANIZATION

The thesis is structured as follows: Section 2 is where we review some of the previous work and implementation on smart grids. Section 3 is where we give a background about all the components. Section 4 is where we explain our model in details and implementation of that model in details, Section 5 is where we simulate our model and record the results obtained. Section 6 is where we conclude our work and put a future scope into perspective.

2. LITERATURE REVIEW

The technique for the process of translating cartesian coordinates to the polar co-ordinate system of non-concentric circles, called the rubber layer, is defined in John Daugman et.al [9] and [10]. Automatic Speaker Recognition in recent decades has achieved significant success. Many works were suggested to increase recognition rates based on various mathematical paradigms. (Tolba et.al [11] Hidden Markov Model HMM applied it to classify the ten Arabic speakers where speech signal description (MFCCs) was picked. During text-dependent experiments the recognition rate was 100% and for text-independent experiments 80 percent. Daqrouqa et.al [14] used the entropy packet formants and wavelet as inputs to a classification neural network. A recognition rate was found for vowel-dependent 90.09 percent and for vowel-independent 82.50 percent. The GRNN General Regression Neural Network and the Back Propagation Neural BPNN Network were used as classification devices Wu et.al [14] to identify 36 Chinese speaking speakers using the EMD method of extraction by empirical mode decomposition. The first form yielded 78.16 percent, while the second yielded 88.82 percent. As a classification algorithm, the artificial neural network model NN is used Chen et.al [15] MFCCs function from the speech signal was extracted. Then the input dimension is reduced and the redundant information is removed with Kmean Linde-Buzo-Gray algorithm. Training network error was 0.015 where the amount of the hidden layer is 950. Ge et.al [16] proposed the TIMIT 8K Database for neural NN system to classify and verify text-independent speakers. The 39 (MFCCs) were generated from the talk. The system was classified by 100%. The ANN Back Propagation Classifier (BPNN) is used Wali et.al [17] seventeen MFCC features from 50 users are extracted and approximately 92 percent to 70 percent recognition exactness for 10 to 50 users is achieved. Multiclass-SVM-based speaker cluster method Apsingekar et.al [18] has been proposed to provide a functional vector of 13 MFCCs, 13 delta CMS-MFCCs and RASTA-MFCCs. NIST-2002 speech corpus experiments with 64 speaker accuracy amounted to 97 percent. SVM; GMM; and SVM are fused with a GMM for speaker and language recognition on NIST 2003 data where 38 MFCCs and 36 LPCs were used in describing the spoken signal; Campbell et.al [19] The SVM 6.46 percent EER error rate was 7.47 percent for GMM and 5.55 percent for SVM/GMM. In addition to the discriminative classifier SVM combined with the standard GMM pattern classification, the combination has also been checked on 64 speakers in the TIMIT database, using a cepstral function vector 39-dimensional MCC extracted from the speech Chakroun et.al [20]. They

reached an identification rate of 100 percent with just three pronunciations per speaker for the training task in the second set of experiments. Campbell et.al [19] built a GMM supervector support kernel for the vector machine. Experiments were conducted on the 2005 NIST speech recognition (SRE) corpus, which included extracting from the pre-fixed speech signal a 19-dimensional MFCC vector, an equivalent error rate (EER) of 5.68% and a minimum decision cost value (minDCF) of 0.0222. In the context of the three acoustic features (RASTA-MFCC, Gamma tone frequency cepstral coefficients GFCC and Mean Hilbert Envelope Coefficients MHEC in different noisy conditions), Dhineshkumar et.al [21] proposed the GMM/SVM automatic speaker identification efficiency. TIMIT phone-labeled database corpus experiments were carried out. Accuracy was 70.32% for RASTA-MFCC, 68.49% for GFCC and 73.27% for MHEC under street noise. In the context of GMM-UBM, an Arabic speaker recognition scheme was introduced Algabri et.al [22] Thirty-nine MFCCs were used for extraction of features. All tests have been carried out with speeches by 40 KSU Speech Database speakers. They obtained an EER equal to 1.98 percent at approximately 97.8 percent using mobile channel recording. Different deep characteristics are provided for text-dependent speaker verification Liu et.al [23] These profound features are both implemented in the GMM-UBM and identity vectors. In order to assess all processes, the RSR2015 database was selected. The optimal system is 0.1 percent EER. Table 1. summarizes all these works. In device detection, prediction and monitoring, ANFIS has the advantage of good applicability, capacity and performance. It was used in several different systems. Since not many research projects have used it to recognize expressions. Some of those who do this are presented, it is important to explore the use of ANFIS in speech recognition and to investigate this subject more carefully.

3. MATERIALS AND METHODS

In this chapter we review some of the essential components that we used to implement our model, we give a brief introduction to the component, and then we show how we simulate that component inside MATLAB environment.

3.1 BIOMETRICS RECOGNITION

The recognition system has text-dependent, text-independent modalities or a variation between them. In the case of text-dependent modality, the person must speak a word or a set of words that were previously registered by the speaker in the system so that it is done the acknowledgment. On the other hand, in the independent text modality, the system must identify the person who is speaking independently of what is being spoken, that is, the system learns exclusively the person's voice and not the combination of the voice and the text [5]. In this work, the independent text modality will be used, as it is the most versatile and can be used in all branches of speaker recognition [6]. With the recent popularity of machine learning and deep learning, speaker recognition applications are getting more powerful. Machine training has become faster and the number of qualified speakers has increased

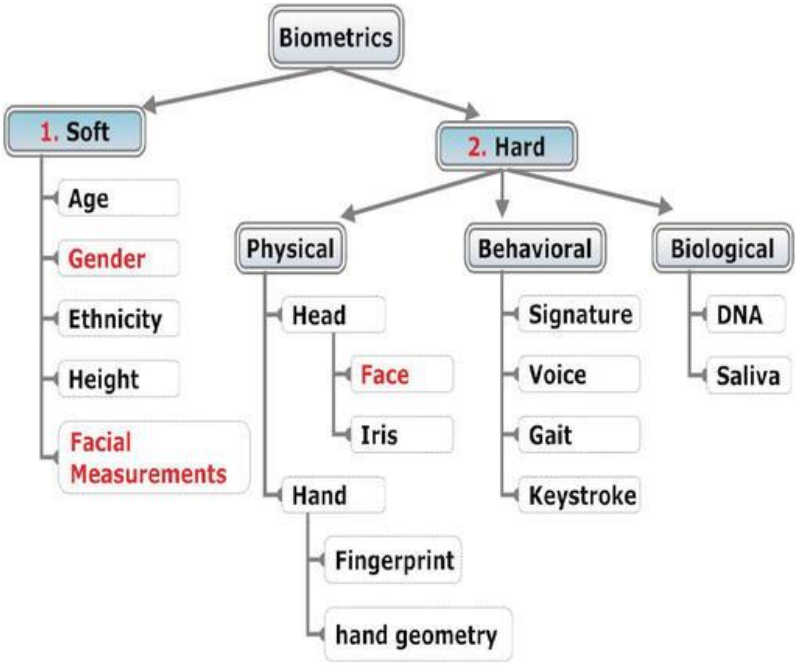

Figure 3.1 Types of biometrics

[7] This makes machine learning a very attractive prediction tool for voice recognition, but it's not the only one, there are other approaches like using Dynamic Time Warping (DTW) [8] Machine learning techniques are computational algorithms that identify patterns and models through the analysis of a data set [9] Such standards and models can be used to make predictions in various applications such as speaker recognition. In order to use machine learning algorithms, it is necessary that the information of the voice goes through pre-processing. For this, a technique is used to obtain characteristics of the voice signal. There are several techniques that can be used to extract these characteristics, as Linear predictive coding (LPC), Mel frequency cepstral coefficients (MFCC) and Linear Perceptual Coding (PLP) [10] Putting pre-processed information as input from a machine that has already finished its learning process, the output is obtained prediction that the speaker is who he claims to be or whether the speaker is the output, depending on whether the application interest is verification or identification. The MFCC technique has been shown to be quite efficient to extract the characteristics of the voice signal [11] therefore it will be the technique used in this work. Performing a search for existing systems and scientific articles related to the theme of this work, we found tools available for the development of speaker, such as Spear [12] and Alize [13] and also ready-made applications like Microsoft Azure's Speaker Recognition API and Ok Google , from Android operating system . However, these tools and applications, in some situations, are paid for, run in the cloud, or not very flexible for speaker verification implementations with modality independent of text. With regard to scientific articles, you can find several techniques that use machine learning for speaker recognition [14] each focused on a respective application. It is important to highlight that this work does not intend to propose tools for compete with those previously mentioned, or propose techniques other than those found in the literature, but rather to present an introductory to speaker recognition systems using some machine learning techniques, highlighting important aspects in the implementation process and providing code samples to assist future implementation work on this topic. For As such, we will use the Python programming language with the Scikit-learn toolbox library and one of the most used in the literature, we will give greater emphasis on the use of the Gaussian Mixture algorithm

.

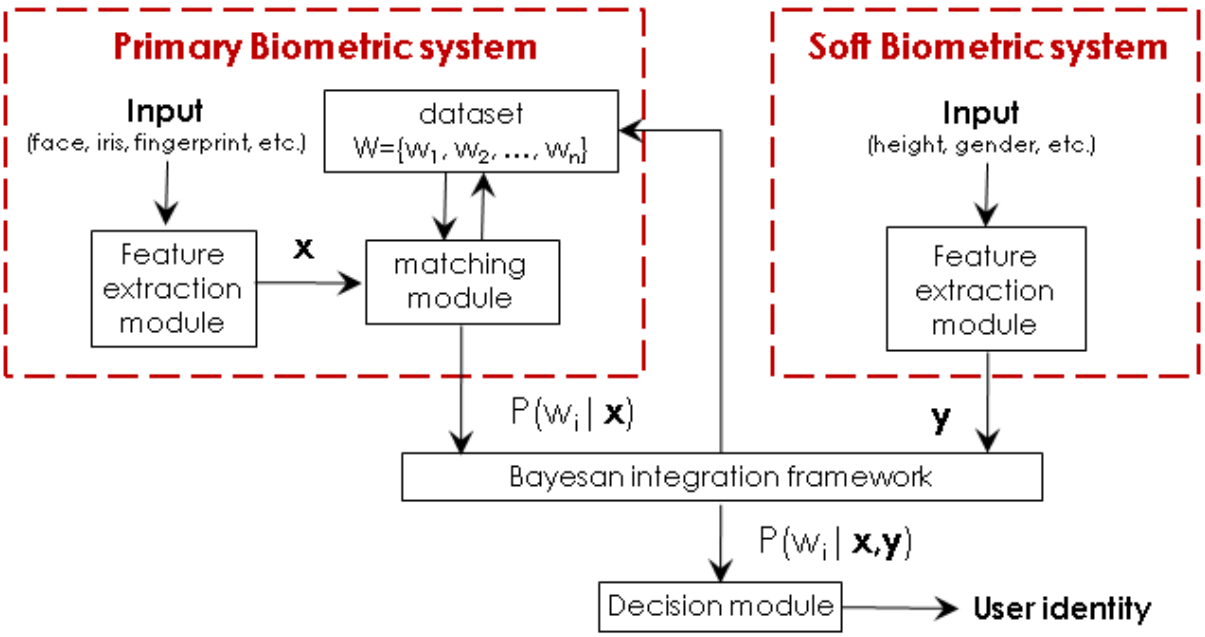

Figure 3.2 example of expert system in voice recognition [13]

3.1.1 Types of biometrics

Biometrics covers any type of measurement of physiological and / or behavioral characteristics of the human body, it is usually associated with the recognition of people, as it is reliable and convenient for use. Biometric information is personal information that is distinct from other information by permanently linked to humans and most of the time resistant to changes [22] Biometrics can be the size of a limb of the body, the way a person eats, distance between the eyes or any other physiological and / or behavioral characteristic that can be measure. Figure 1 shows some examples of behavioral, physiological and both biometrics (voice). In the security area, the most important types of biometrics are those that can distinguish one individual from all others, such as fingerprint, iris, retina, face, voice, etc. It is inherent that biometrics are more reliable and more competent than techniques that use a knowledge base and the based on token (username and password) in differentiating between authorized persons and imposters, because biometric characteristics used for this purpose are unique [24]

3.1.1.1 Digital Printing

The fingerprint is the design formed by the papillae of the fingers, it is formed by the accumulation of dead and cornified cells that constantly fade as a scale from the exposed surface [24] An example of the design formed is Figure 2. The fingerprint is probably the type of

biometrics that society in general is more accustomed to, being widely used for personal identification, access control, check-in and check-out, etc. There are many types of fingerprint scanners on the market, including optical, electronic solid state and ultrasound. These scanners started to be mass produced years ago and have low cost [25] Because it is very consolidated, and also due to the cost, printing digital is widely used. Even though it is quite advanced, digital printing has problems with fraud, because can be easily replicated using advanced latex molding techniques. Another problem is the legibility of digital, where many people who work with manual services do not have digital readable [26]

3.1.1.2 Irises

The iris of the eye carries a large amount of information, enough to distinguish a person. It's so unique that even a person's two irises have different patterns [29] shows the information contained in a small area of the iris. Iris The identification via the iris is done by positioning the eye in front of a camera, which will use light to locate the iris, isolating other parts of the eye by digital photography. After locating the iris it is analysis of segments of the iris was performed to identify the person [30]question of positioning where the person needs to stand in front of and close to the camera during a time, it is a disadvantage of use and something very invasive. The iris has other problems, some eye diseases do not allow for correct analysis and a high quality photo allows to defraud many systems [31]

3.1.1.3 Voice

Figure 1 shows that voice biometrics can be done based on behavior as well as in the physiology of a person's voice. Physiological biometrics is due to the formation of the nose, mouth, larynx, vocal cords and the rest of the anatomy responsible for the voice, making a person's voice only [32] having enough information to be able to be used in systems security as a form of authentication. The behavioral biometry of the voice is focused on the speech rate, accent, language, etc Figure 4 shows a voice signal generated by a person, that voice signal was shaped by the bodies responsible for the voice, leaving their characteristics in it, that is, information that can be used to recognize a person. Unlike many biometrics, voice biometry is not very invasive, as it is possible to recognize a person simply talking to him, having a device that is capturing audio and running that application. An example would be talking to a bank manager, where he

would have a microphone picking up the conversation. In this case, the system recognizes the manager's voice, so it can separate the manager's voice with that of the customer, allowing customer authentication simply by conversation, not requiring the client to provide personal information so that he can make the desired procedures and if an imposter appears, the system will notify the manager [33] The use of voice biometrics also has its disadvantages, as it is necessary that the person be able to speak. Other situations that can present problems is when the user has some voice problem, either because of a blocked nose because of an illness or because the person is hoarse. In this work, we will focus on voice biometrics for person recognition, usually called speaker recognition.

3.2 Speaker Recognition Concepts

Speaker recognition is a generic term used for any procedure that involves the recognition of a person's identity based on their voice [33] An important mention to make is the difference between speech recognition to speech recognition of speaker. In speaker recognition, biometric characteristics are used to recognize an individual. In speech recognition, it is desired to recognize what is being spoken in an audio, the only focus is the identification of words and numbers present in the audio, that is, the desired information in audio is different between the two recognition systems. This work is focused on the characteristics used for speaker recognition. A speaker recognition system attempts to model the characteristics of the vocal tract of a person. This can be done via a mathematical model of the physiological system of human speech production or simply a statistical model with output similar to the characteristics of human speech. Once the model is generated and associated with the individual, new instances of speech can be compared with the model of the target person, having the model of other people as a contrast to know the probability that the instance belong to the same person who generated the target model. This is the fundamental methodology for all speaker recognition applications [33]. It is worth noting that speaker recognition uses the only biometrics that can be easily tested remotely through the existing infrastructure, the telephone network. This makes very valuable and incomparable speaker recognition in many real-world applications [33]

3.2.1 Speaker Recognition Branches

Speaker recognition has many branches that are directly or indirectly related. In general, there are 6 main branches, which are verification, identification, classification, segmentation, detection and tracking. The branches can be used together. Verification is currently the most popular branch for its importance in security and access control [33] .In this work, the branch verification.

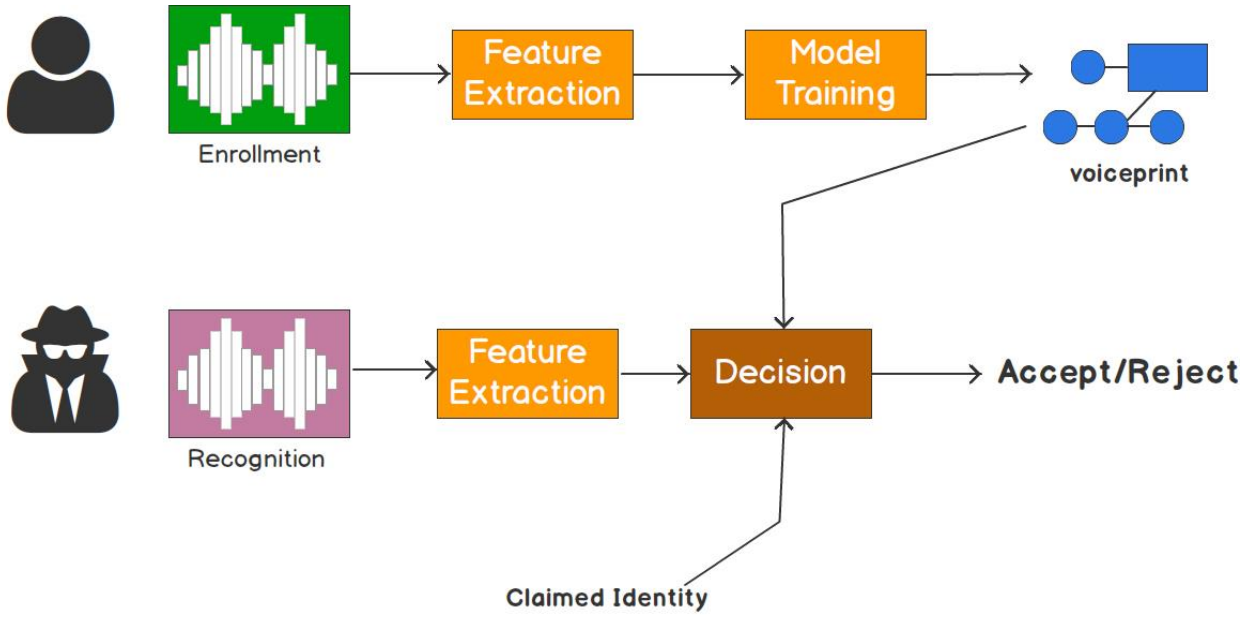

Figure 3.3 Security system based on sound biometrics

3.2.1.1 Verification

In a generic verification application, the person being verified first must identify, usually by a method that does not involve the voice. The presentation of the Identity (ID) can be done, for example, by providing the Personal Identification Number (PIN) or the user. Provided identification, the system retrieves the vocal tract model that is related to that identification so that the verification is done. That is, the person's voice will be compared with the model for confirm (authenticate) that the person really is the individual of the past identity. This model is called the target speaker model (target speaker model) [34]

3.2.1.2 Identification

There are two different types of speaker identification, the closed-set and open- set (open set). Together, the voice is compared to the registered speaker model and returns the ID of the

speaker who has the voice closest to the voice that was entered as input, that is regardless of the person who spoke, even if he is not registered, a message will always be returned ID [35] The open set identification can be seen as a combination of the closed set with the check. In this case, identification of the same speaker is made as it is done with the closed set, where it is the ID of the person closest to that voice is identified , then the voice test is performed on the verification model related to that ID. The ID is only returned if the person passes the verification test [36] The voice is used to locate the person's ID , after that the system works the same as it would if the user had provided a PIN as commented in subsection 3.2.1.1.

3.2.1.3 Classification

The classification objective is a little more vague. The idea of this type of system is to classify the speaker, which may be the age or gender of the person for example [37]

3.2.1.4 Segmentation

Segmentation is used to separate the parts contained in the audio. It could be music, speakers, noise or any other sound present in the audio [38] An application in recognition of speaker would be a conference, where it is necessary to separate the voice of everyone present in the audio. After the separation, each speaker present can be identified and verified.

3.2.1.5 Detection

It is the act of detecting one or more specific speakers in an audio. So he is fundamentally the segmentation junction with identification and / or verification [39]

3.2.1.6 Tracking

It has a subtle difference in relation to detection where the idea is to return at what point the person spoke and how much was the duration of the speech, that is, mark (locate) at what times and what period that an individual is present in the audio [40]

3.2.2 Modalities

Speaker recognition systems can be implemented using different modalities, in which they are linked to linguistic use, context and other means. However, in a practical sense these modalities

are only relevant for speaker verification [41] The following is presented succinctly the most relevant modalities for speaker verification.

3.2.2.1 Text-dependent

In this modality the person must speak words or phrases that were used in his registration in the system. The system learns and generates a model of the individual taking into account the biometry of the voice of the person and also what is being said. This system only makes sense for the verification branch [42]

3.2.2.2 Independent of Text

The text-independent speaker recognition system is the most versatile of all modalities. It is also the only viable modality that allows it to be used for all branches of speaker recognition. The system should recognize depending only on the characteristics of the tract regardless of the person who spoke, even if he is not registered, a message will always be returned vocal of the speaker and makes no assumptions about the speech text. As it is a system that does not depend

text, can easily be defrauded, as any good quality recording can system

3.2.2.3 Text-Prompted (Random Text)

It is a system that shows the speaker a random phrase and forces the speaker to speak that specific phrase at the time of the test. This modality was created to fight imposters. If the speaker cannot anticipate the requested phrase, he cannot prepare himself to be able to deceive the system. This system only makes sense for the verification branch There are two main approaches to this system. The first uses the dependent modality text to randomly generate phrases and create a text dependent template for the generated phrase. This model will take into account the speech content and the biometry of the voice. The second approach, is the use of a text-independent speaker recognition system in conjunction with a speech recognition system. In this case, it is necessary that the text present in the speech corresponds with the phrase requested by the speech recognition system and that the speaker recognition system check the voice biometry

3.2.2.4 Knowledge-Based

It is a modality that combines the independent text modality, or Text-Prompted, with speech recognition system. In this case, the user must answer some questions that he should know, such as the Individual Taxpayer Number (CPF), mother's name and others. The system of speech recognition checks whether the answer is correct and the speaker recognition system checks biometrics. This system is of greater importance for verification, but it can also be used in other branches like detection, where you can search for audios of a speaker talking about certain topics

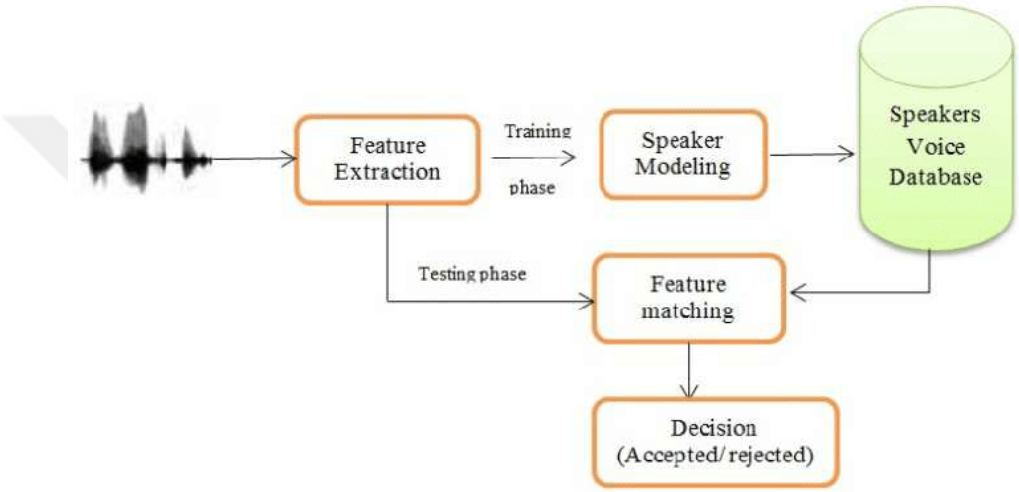

Figure 3.4 Speaker voice recognition system [43]

3.3 Extracting Features from the Voice Signal

A voice signal has a large amount of information, but it is necessary to be able to extract the important part of the information for an efficient and reliable system. In recognition of speaker the important thing is to obtain the personal information that we call biometrics, having then a focus greater in phonetics. In speech recognition on the other hand, it would be important to obtain the information allows to identify phonemes or any other information that helps to identify the content (words) of the audio. Before presenting the characteristics of the voice signal, we will introduce the basic concepts of analog-digital conversion to the voice signal

3.3.1 Voice Signal

The extraction of the characteristics is done in an existing voice signal, so a very important factor important is the audio quality. There are factors that influence audio quality, such as frequency sampling, jitter , truncation error, overlap and noise.

Below, some factors that influence the audio quality are briefly presented.

2.3.1.1 Sampling

The voice generated by a person is an analog signal, so sampling is necessary of the voice signal so that computational operations can be performed on top of the signal. When making signal sampling, the highest frequency of the signal should be considered, as it is necessary that the frequency sampling rate is at least twice as high as the highest frequency of the signal so that it does not occur signal overlap [44] Due to technical situations, analog telephony uses a sampling frequency of 8 kHz, but they are of low quality audios, because in these cases the important thing is the intelligibility and not the fidelity of the voice [45] The reason for the poor quality is the lost information, some signal frequencies of voice require a higher sampling frequency so that there is no overlap. Increasing the sample rate improves audio quality, but in increased processing, a higher transmission rate when it comes to an application in real time, plus a larger storage space.

3.3.1.2 Quantization

After sampling the signal still has infinite possibilities of amplitude and to make it digital it is necessary to discretize the signal amplitude. The technique used for this is quantization, where the signal amplitude value is replaced by a finite amount of values. The number of possible values is 2 , where b is the number of bits used in quantization is an example of quantization of a continuous analog signal over time, one can realize that the amplitude is no longer continuous and becomes discrete. It is possible to observe in this figure the levels of quantization and decision levels. The level of quantization is a value that can be represented in bits and the decision level is the voltage range to which the signal belongs. This figure shows the quantization levels of 147v and 148v and a sample value of 147.39 which will be represented by the value 147v, as it is the representation that will have the smallest quantization error (-0.39v). . By increasing the number of bits, it is possible to reduce the distance between the quantization levels and how Consequently, decrease the quantization error and improve the audio quality. In the same way as the rate sampling, the more bits, the more expensive the audio is in the transmission network, in storage and processing

3.3.1.3 Noise

Noise can be generated due to physical aspects or the environment where the audio is captured. THE noise generated by physical aspect may be due to the quality of the microphone that was used for the capture audio or the medium in which the audio signal is transmitted. The noise generated by the environment would be when the environment itself has background sound, such as close people talking or any other type of sound

3.3.2 Techniques for Extracting Characteristics from Voice Signal and MFCC

The characteristics of the voice signal can be used for both speaker recognition and for speech recognition. These characteristics make up the behavioral and physiological biometrics, but they also have characteristics that allow to recognize speech This work has focus on physiological biometric characteristics. In a voice signal there is a large amount of information strongly correlated or that is not essential for understanding the message, an example of non-essential information is silence which reduces system performance and can cause unexpected results In this way, it is It is convenient to remove excerpts of silence from the voice signal. A good part of the information present in the voice is related to the vocal characteristics of the speaker, so the job of speech recognition is to try to filter that information irrelevant signal and interpret the message being transmitted. Recognition systems speaker, on the other hand, seek to use all the information contained in the signal that makes it unique The characteristics obtained from the voice signal must be sufficient to achieve recognition good quality, but in large quantities generates the need for more processing, which can be a problem for real-time scans [46] One way to acquire characteristics of the voice signal is using a technique of characteristics short-term spectral (short-term spectral features). Short-term spectral characteristics, of frame computation between 20 and 30 milliseconds in duration. These characteristics are usually spectral envelope descriptors short period (short-term spectral envelope), which are acoustically correlated with timbre, such as the resonance properties of the vocal tract [47] The reason for the separation of audio into small frames is due to the constant variation in the voice. Separating in small frames a variation of the signal and its characteristics relatively stationary facilitating the extraction of characteristics. Next, a brief explanation of two important characteristics for speaker recognition, the resonance of the vocal tract and timbre. The resonance of the vocal tract is about the fundamental frequencies generated by the vibration of the vocal cords in a certain length of time, it depends on the characteristics of the air flow, vocal tract shape, shape and

tension of the vocal cords at the time observed by having great importance for short term spectral characteristics, since the fundamental frequency varies from person to person. The timbre is the harmonic content that is related to the location of formants and their featureshaving a direct relationship to the resonance of the vocal tract. You can see the importance of timbre in speaker recognition with analog telephony, where the frequencies above 3.4 kHz. In this case, there is no difficulty in understanding what is being said, but it is difficult to identify the person who is speaking, because a large part of the harmonics of the voice were filtered. There are several techniques for extracting physiological characteristics from the voice signal such as MFCC, LPC and PLPbut for a limited time, we will focus only on the MFCC, which is widely used in the type of application proposed In this job.

3.4 Machine Learning

On the internet you can find several interesting definitions about machine learning, among them we can mention the Google Cloud (2017), " Machine learning is functionality that helps software perform a task without explicit programming or rule "and by the authors [46] " The field of study interested in the development of computer algorithms to transform data into intelligent action is known as machine learning ". With these definitions we can say that machine learning techniques can be used to teach a computer to make future decisions based on characteristics of a given available data set. Machine learning techniques can be classified as supervised, unsupervised and by reinforcement. Supervised learning techniques are those in which the training process is performed with a data set that has one or more independent variables (inputs) and one or more dependent variables (outputs), in which both inputs and outputs are known. These techniques can be used to solve regression problems, where you want to predict data numerical problems, or in classification problems, in which it is desired to predict categories [3]. In unsupervised learning, the model generation process is performed with a set of data in which only inputs are provided. These methods can be used for problems where you want to identify useful patterns in the data and for segmenting groups with similarities, for example [47] Reinforcement learning techniques will not be covered in this work. Succinctly, these techniques work with a training process carried out from rewards in interactions with the environment. Examples of use can include game optimization, among others. In this work, the system can be seen as a classifier, as recognition is done for verification, where the output of this

system classifies whether the incoming voice matches or does not match the voice being verified. Taking this into account, following we will present some machine learning techniques that can be tested in this work.

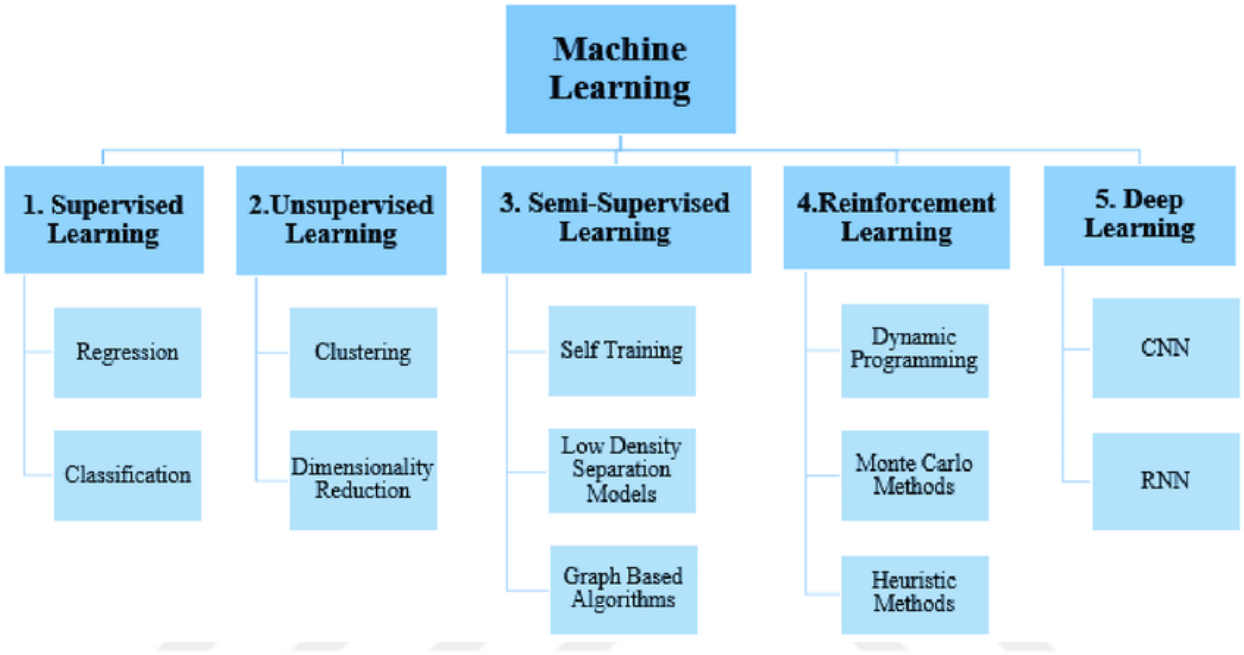

Figure 3.5 Types of machine learning [15]

3.4.1 Machine learning algorithms

3.4.1.1 Support Vector Machines

The Support Vector Machines (SVM) is a classifier that models decision limits between classes via hyperplane [49]For example, Thinking about Cartesian plan, the hyperplane separator separates the plane in two, where one side of the plane belongs to a class and the other belongs to the other class, as can be seen in Figure 9. If multiple limit solutions exist SVM addresses this problem through the concept of margin for which it is defined to be the shortest distance between decision limits and any other sample. The decision limit chosen must be that which has the maximized margin The vectors support (support vectors) are vectors corresponding to the maximum margin of the hyperplane. In Figure 9 would be the vectors X a and X b . After training the rest of the vectors can be discarded as they make no difference to the decision limit decision, then it is necessary to try to find the one that generates the least amount of generalization errors. In speaker verification, a class consists of the target speaker's training

vectors and the other class consists of the imposter speaker vectors. Using these vectors labeled by their classes, SVM finds a hyperplane separator that maximizes the separation margin between the two classes also shows the separation made when it comes to two classes that are linearly separable in the original input plane (X 1 and X 2), but this is often not the case. In these situations, The kernel functions that allow the computation of the internal products of two vectors in the feature of the kernel , which has a larger dimension space than the input space. On a larger space, it is easier to separate the two classes with a hyperplane. Someway intuitive, a linear hyperplane in the kernel characteristic space corresponds in decision limit in the original input space, an example would be the MFCC space [50].

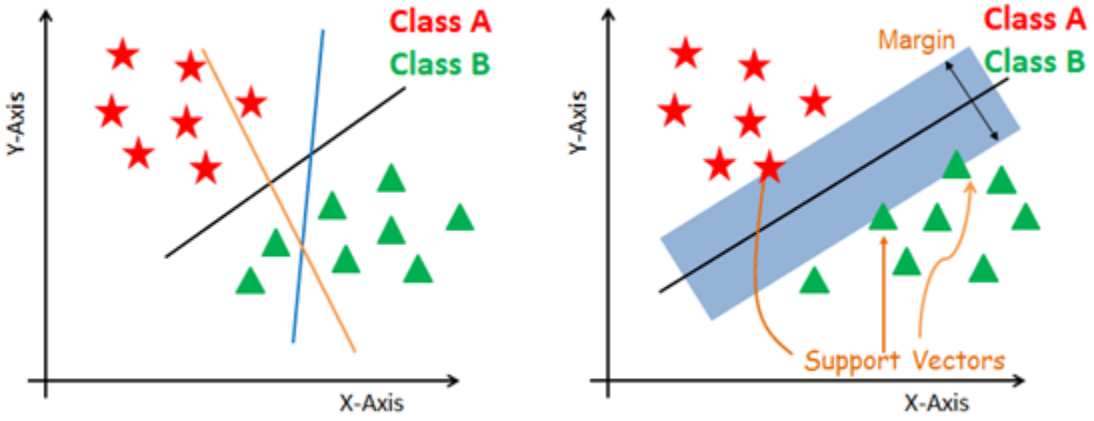

Figure 3.6 SVM classification [24]

2.4.1.2 Naive Bayes

Naive Bayes is a model in which the training data entry characteristics are considered conditionally independent to simplify the model structure. Because of this consideration, one can simply apply equation 2.2 adapted from With this equation is calculated the probability of being the real speaker given the input characteristics. P (X) is the marginal probability of the input characteristics, P (Fantee) is the marginal probability of being the real speaker and P (X | F alante) is the probability of the real speaker displaying the input characteristics. In case There are only two classes of verification, so it is only necessary to apply the equation once, since knowing the probability of being the real speaker, we also know the probability of not being the real speaker and the desired one is to locate the class that is most likely to be classified.

$$P\,(\,False\,|\,X\,)\;=\;P\,(\,X\,|\,Fante\,)\;\cdot\;P\,(\,Fante\,)\,P\,(\,X\,)\;(2.2)$$

The assumption of conditional independence in the model is clearly strong, which can lead poor representations of the conditional densities of the classes. Even if the assumption is not precisely satisfied, the model can still provide good performance in practice because the decision limits may be insensitive to some of the details in class conditional densities.

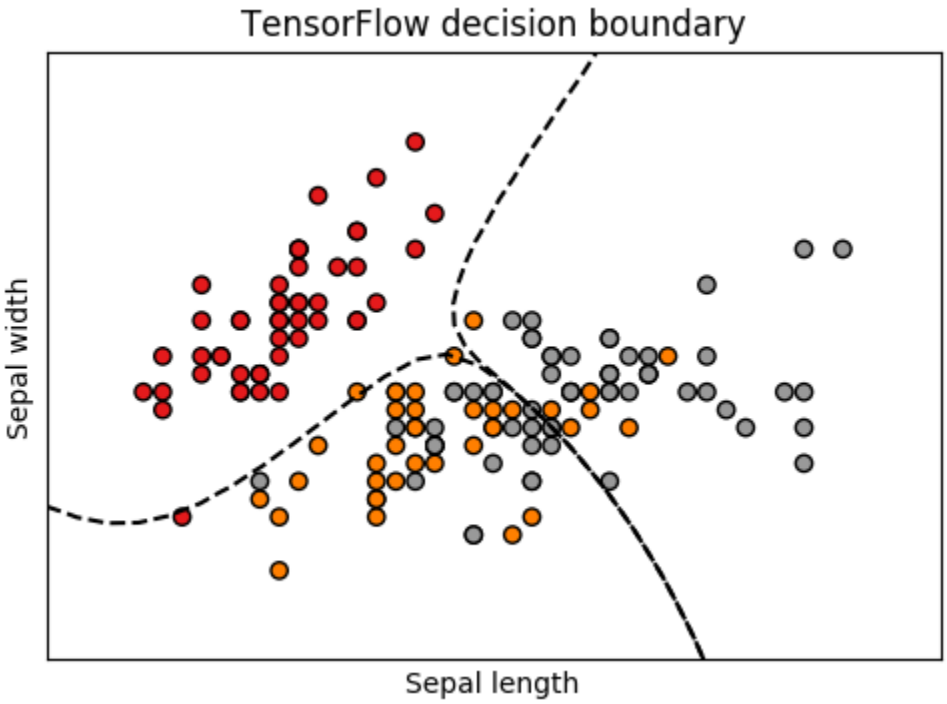

Figure 3.7 Naïve bias Classification [16]

3.4.2 Fuzzy logic

Fuzzy logic arises from the theory of fuzzy sets, the basis for the development of the linguistic approach [26], and its truth values can be intermediate between the true and false values. In this way, fuzzy logic, unlike propositional logic, is not dichotomous, it quantifies truth values and has the ability to deal with approximate causal inference [27]. Fuzzy logic is the basis for rule-based fuzzy inference systems [28]. These are essentially made up of inputs, outputs, membership functions, and inference rules. Through these components and their configuration, the estimated output values are obtained as approximate truth values for the classification, according to the input values. The design process of a fuzzy model involves the definition of the set of inputs and

outputs of the fuzzy system, as well as the position, shape and domain of its membership functions and the rules of inference. Classification with a low margin of error depends on the design of the system, which may depend on the judgment, testing, and validation of experts in the particular area of research [29].

Unlike the neural network approach, fuzzy systems can be interpretable [28], which makes them useful in problems control panel with high interpretability. This allows the description of the outputs in terms of the inputs, inference rules, and membership functions. The certainty in the description of the outputs will depend on the performance value of the fuzzy classification system. However, the interpretation of a fuzzy system can be difficult if its structure is highly complex [30]. The structural complexity of a fuzzy system depends proportionally on the number of inference rules and membership functions at inputs and outputs. The fuzzy models proposed in this article have a structure of moderate complexity to avoid obstructing their interpretability. In recent decades, the application of fuzzy sets has been ignored in the field of natural language and speech processing, and the usefulness and contribution of fuzzy logic in applications related to the processing of speech signals has been questioned [31]. However, the evolution of computing and technological contributions in recent years have favored the fuzzy logic approach for its use in various fields of study [32].

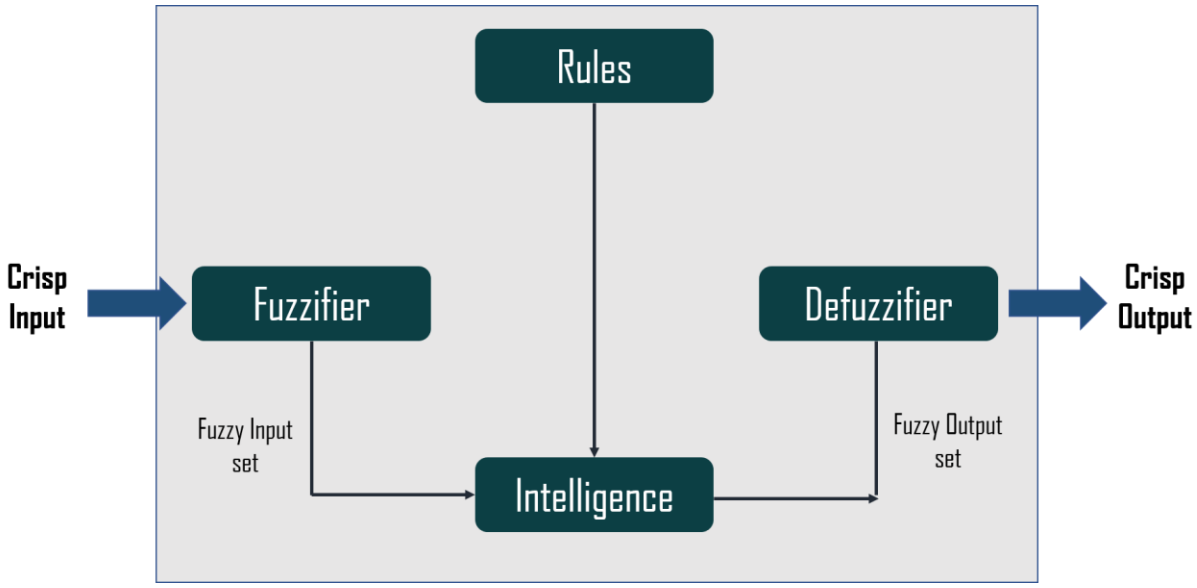

Figure 3.8 Fuzzy membership function [28]

3.4.3 PSO algorithms

The particle swarm optimization algorithm is inspired by emerging social behavior that can be observed in flocks of birds and schools of fish of some species [40].

The population is composed of particles, the position of each of them being a solution point in the space of possible solutions of a number of dimensions equal to the number of parameters of the problem. In each iteration the best position of each particle is updated (best individual position), and the best position of the swarm (best global position). With these values the velocity of each particle is calculated using equation (4).

$$\vec{v_1}(n+1) = w(n)\vec{v_1}(n) + \beta_{i(1)}\left(\vec{x_{pi}} - \vec{x_i}(n)\right) + \beta_{2(1)}\left(\vec{x_g} - \vec{x_i}(n)\right)$$

Where:

i is the index of the individual.

n is the iteration index.

vi is the velocity of the i-th individual.

xi is the position of the i-th individual.

w n ^ h is the inertia function.

x pi is the best position of the i-th individual.

xg is the best position in the swarm.

b are random values between 0 and 1.

Once the velocities have been calculated, the position of each particle is determined for the next iteration using equation (5).

$$\vec{x_i}(n+1) = \vec{x}_i(n) + \vec{v_i}(n+1)$$

Where wmax and wmin are the maximum values and Inertia minima and N is the total number of iterations. The flow chart describing the operation of the particle swarm optimization algorithm is shown.

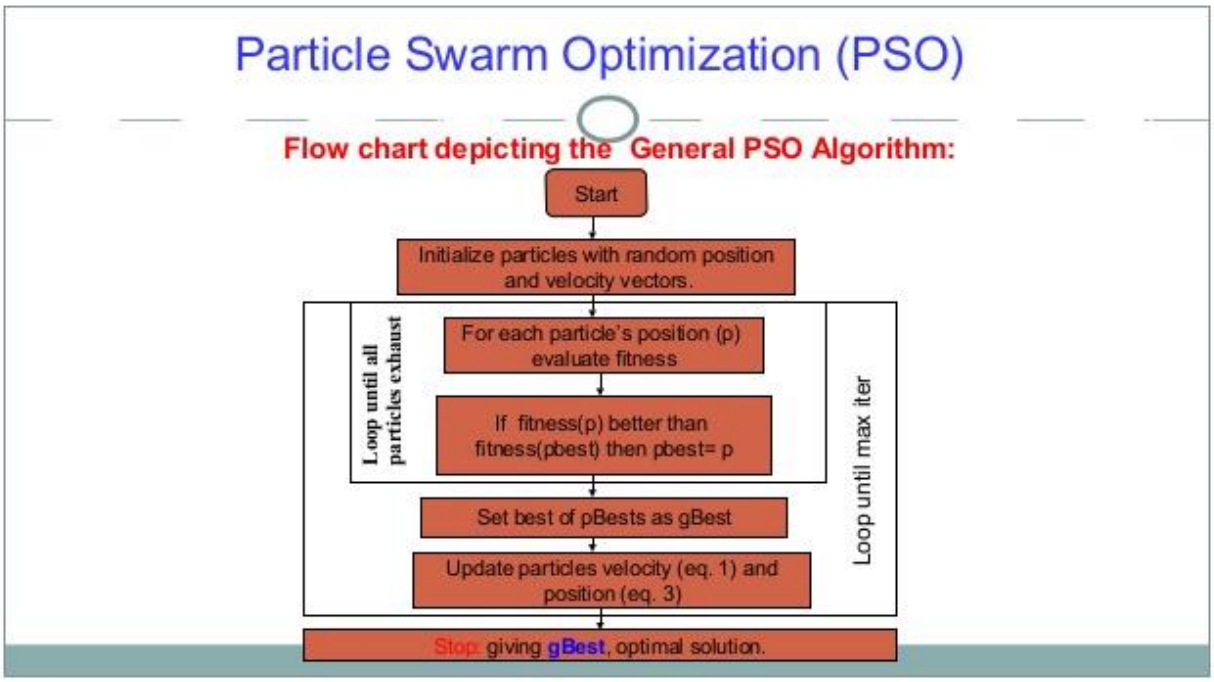

Figure 3.9 PSO workflow [28]

4. PROPOSED METHOD

The purpose of this work is to be an introductory to speaker recognition systems, showing important aspects to build a system that uses machine learning techniques to find patterns, describing available tools and providing sample code that may be used in future work. The speaker recognition system considered in this the work will be based on verification applications using the independent text mode. Speaker models will be generated using different machine learning algorithms. The PSO algorithm will be the most analyzed, as it has shown good performance in applications of speaker verification in independent text scenarios, according to the literature. We will also do a detailed explanation of the available tools that can be used for this purpose.

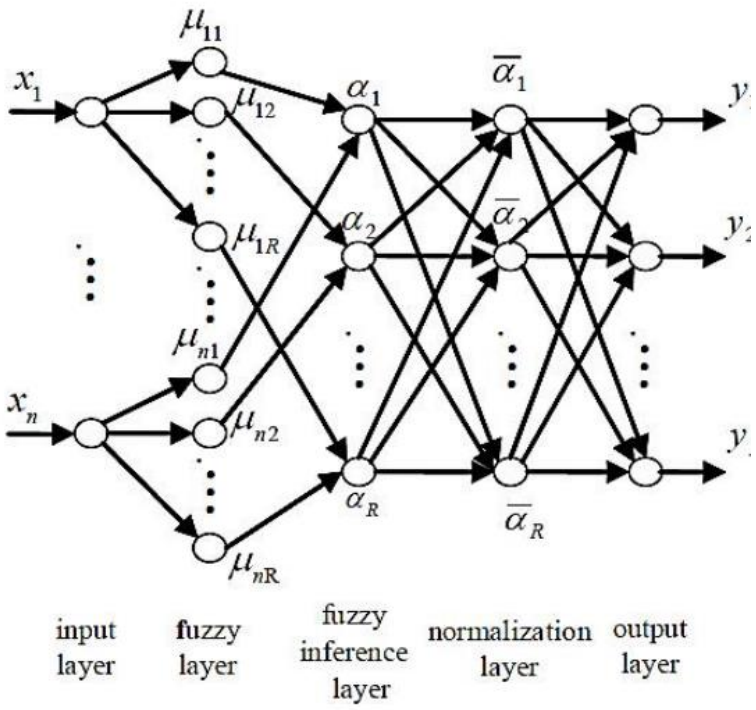

Figure 4.1 Fuzzy logic model.

At this point, the objective is to show the reader the advantages and limitations of the available tools. The tool we intend to use for the development of the system proposed in this work is the Scikit-learn toolbox, available for Python. The scikit-learn a free tool that enables the use of machine learning for non-specialists through the programming language of high level [51]

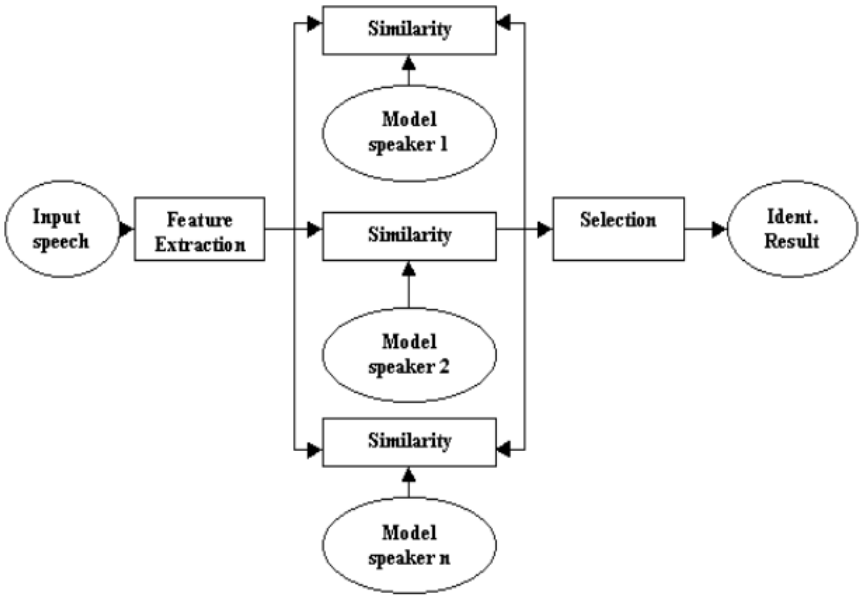

Figure 4.2 Proposed Recognition model

There are free development tools focused on speaker recognition, for example, Spear and Alize, but that do not give much freedom in adjusting the algorithm that generates the models. There are also ready-made applications like Speaker Recognition API , from Microsoft Azure that is paid and machine learning is completely abstracted, having an Interface Application Programming (API) which is used for both training and testing. Because we want a free tool that also allows greater freedom in changing machine learning parameters in development, the Scikit-learn toolbox was chosen for the development. For the other tools mentioned there will be a description and also a comparative analysis with the developed system. Initially we will model the system according to Figure 10 for the PSO and according to with Figure 11 for the other machine learning algorithms. GMM is not a classifier, therefore a modeling technique will be used that uses a background model in conjunction with the target speaker model to make the classification.

4.1 DATASET

In this work we implement our algorithm on 3 different datasets, the we evaluate the results of these datasets in the next chapter, these dataset contain multiple voices of famous people yet in

our model and to avoid confusion, we experiment with only 4 different voices each has a unique pitch and frequency, the datasets are:

Gender Recognition by Voice: This database's goal is to help systems identify whether a voice is male or female based upon acoustic properties of the voice and speech. Therefore, the dataset consists of over 3,000 recorded voice samples collected from male and female speakers.

Common Voice: This dataset contains hundreds of thousands of voice samples for voice recognition. It includes over 500 hours of speech recordings alongside speaker demographics. To build the corpus, the content came from user submitted blog posts, old movies, books, and other public speech.

VoxCeleb: VoxCeleb is a large-scale speaker identification dataset that contains over 100,000 phrases by 1,251 celebrities. Similar to the previous datasets, VoxCeleb includes a diverse range of accents, professions and age.

The datasets can be downloaded from the links below:
https://www.kaggle.com/primaryobjects/voicegender

https://www.kaggle.com/mozillaorg/common-voice

https://www.robots.ox.ac.uk/~vgg/data/voxceleb/

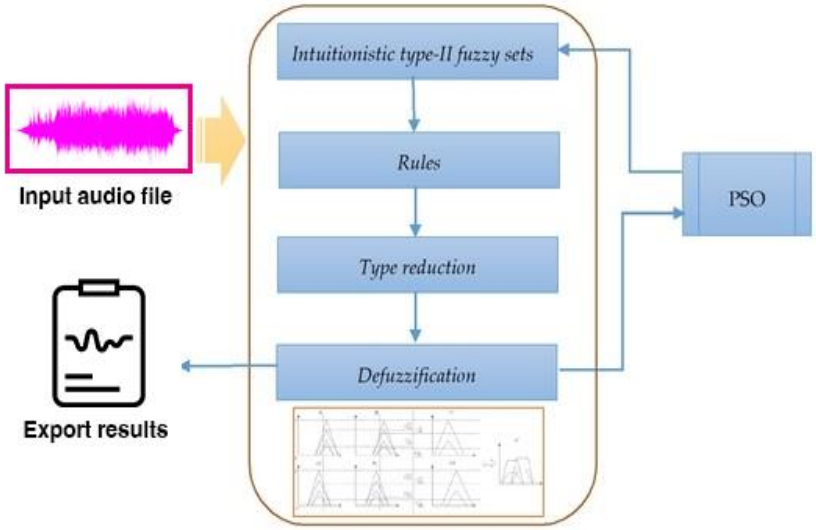

Figure 4.3 Proposed Scheme

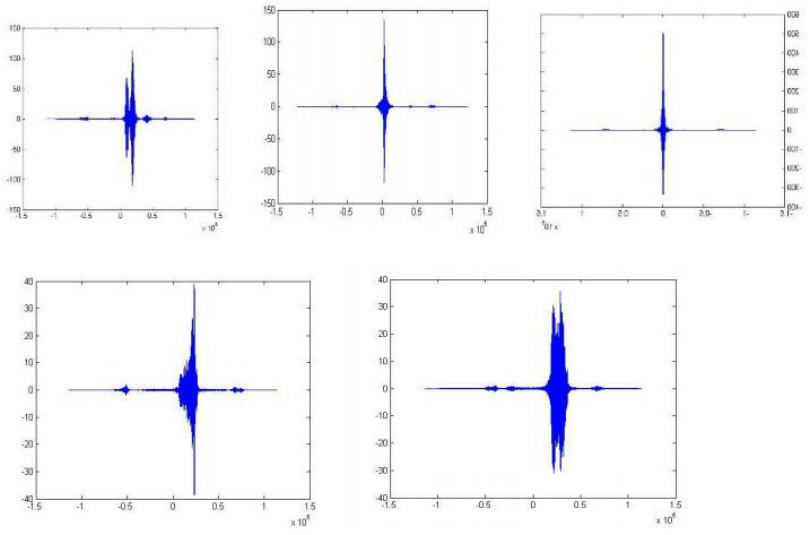

Figure 4.4 Speech features in GUI

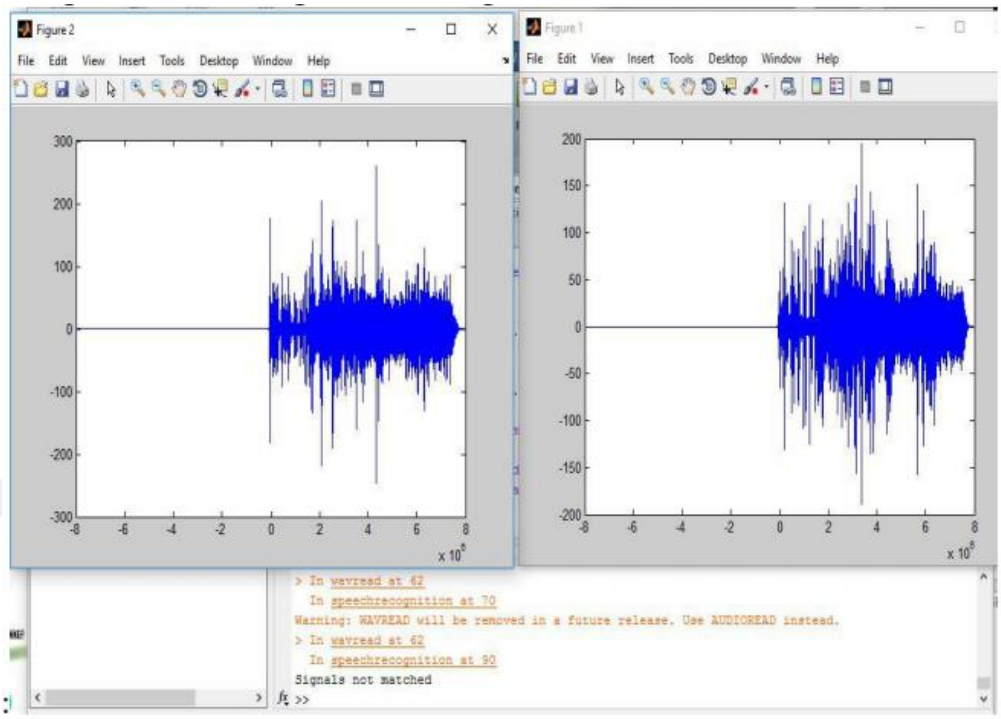

Figure 4.5 a chunk of speech in matlab Gui

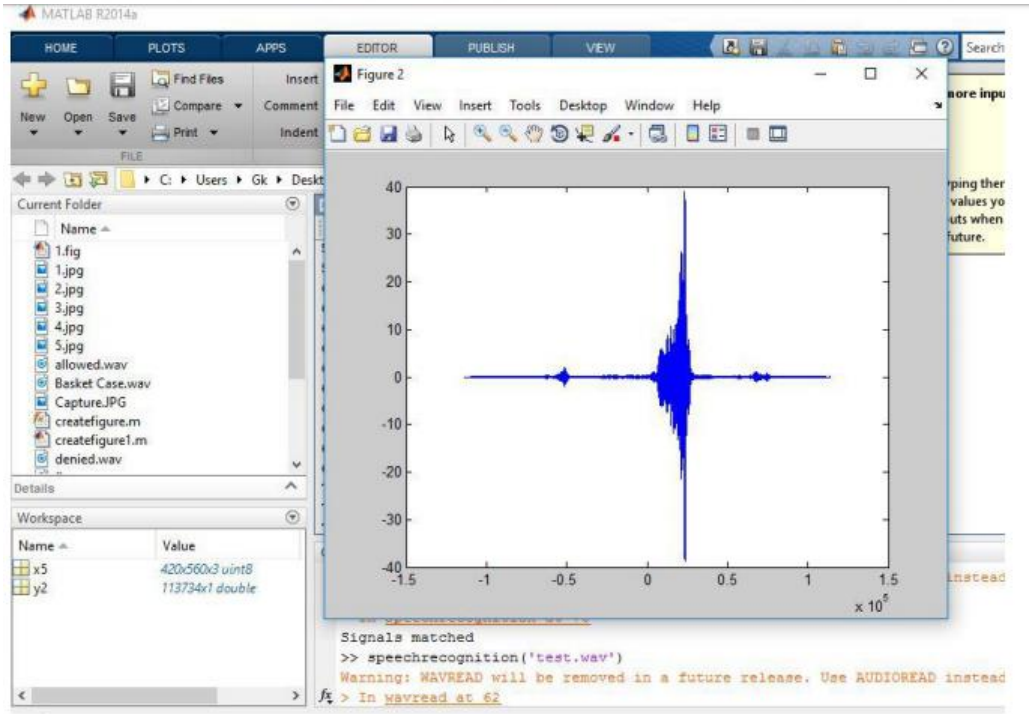

Figure 4.6 Isolating a feature in the speech chunk

The background model reports the probability of input signal belongs to an imposter and can be created for each registered speaker or can be made a universal model. The Universal Background Model (UBM) is the most prevalent approach at the time [52] where a single background template is created for all speakers. This model is created in the same way that the speaker model is created, and the audio samples used are chosen at random from the available training data. Voice samples from UBM will be distributed so that half are women and half are men, in order trying to prevent the system from working better for one of the genres. In both models, there is a voice signal as input and a Front-end processing. THE Front-end processing is about extracting the characteristics of the voice signal (MFCC) and the withdrawal of useless information that can impair speaker recognition, for example, silence [53] The database used in this work will be the English Language Speech Database for Speaker Recognition (ELSDSR), which is a database dedicated to speaker. This database has the audio recording of people reading a single reading session, as well as com has a relatively extensive amount of audio samples to learn the features.

5. SIMULATION AND RESULTS

Tables 5.1, summarize the performance results in each option for each bio-inspired algorithm, regardless of the fuzzy classification model. The results of the best optimized fuzzy classification models found by option in each algorithm are specified in tables 5.2, and the best option and model are determined based on their performance indicators. The best results per model are shown in Table 5.3 with the bioinspired algorithm and data set used to achieve each of its indicators. Table 5.4 shows the average results by number of inputs and the average results of the models with a common input.

Table 5.1: The performance results in each option for each bio-inspired algorithm

GA		**Time better in choice**			**Average of indicators per**		
		EP	EA	MSE	EP	EA	MSE
A	1	43.75%	52.94%	50.00%	62.39%	72.25%	18.07%
	2	31.25%	41.18%	37.50%	62.71%	72.50%	17.81%
	3	12.50%	0.00%	6.25%	58.38%	65.38%	20.36%
	4	12.50%	5.88%	6.25%	59.48%	68.88%	19.87%
B	1	37.50%	25.00%	37.50%	63.27%	71.13%	17.36%
	2	50.00%	50.00%	56.25%	63.90%	74.25%	17.47%
	3	12.50%	10.00%	6.25%	59.61%	64.63%	19.78%
	4	0.00%	15.00%	0.00%	59.52%	67.88	19.74%
C	1	43.75%	57.89%	50.00%	68.80%	76.13%	15.18%
	2	25.00%	26.32%	31.25%	64.69%	73.50%	16.70%
	3	12.50%	0.00%	12.50%	62.11%	68.38%	18.63%
	4	18.75%	15.79%	6.25%	63.01%	70.25%	18.25%
Total times best	1	41.67%	44.64%	45.83%	**Final Average**		
	2	35.42%	39.29%	41.67%	62.32%	70.43%	18.27%
	3	12.50%	3.57%	8.33%			
	4	10.42%	12.50%	4.17%			
Best set of data		1	1	1			

Table 5.2: The best option and model are determined based on their performance indicators

DE		Time better in choice			Average of indicators per		
		EP	EA	MSE	EP	EA	MSE
A	1	50.00%	50.00%	56.25%	65.75%	74.50%	16.43%
	2	25.00%	22.73%	18.75%	64.86%	72.63%	17.06%
	3	6.25%	4.55%	6.25%	61.93%	68.25%	18.51%
	4	18.75%	22.73%	18.75%	63.29%	71.75%	17.94%
B	1	37.50%	44.44%	50.00%	56.86%	73.38%	16.53%
	2	25.00%	27.78%	31.25%	63.70%	71.50%	17.69%
	3	0.00%	0.00%	0.00%	59.05%	66.63%	20.00%
	4	37.50%	27.78%	18.75%	63.30%	71.75%	17.91%
C	1	31.25%	50.00%	37.50%	63.43%	72.75%	17.29%
	2	37.50%	25.00%	25.00%	62.63%	72.38%	18.04%
	3	6.25%	0.00%	6.25%	59.31%	66.25%	19.89%
	4	25.00%	25.00%	31.25%	62.36%	69.38%	18.65%
Total times best	1	39.58%	48.21%	47.92%	**Final Average**		
	2	29.17%	25.00%	25.00%	62.96%	70.93%	18.00%
	3	4.17%	1.79%	4.17%			
	4	27.08%	25.00%	22.92%			
Best set of data		1	1	1			

Table 5.3: with the bioinspired algorithm and data set used to achieve each of its indicators

HS		Times better in choice			Average of indicators per		
		EP	EA	MSE	EP	EA	MSE
A	1	37.50%	25.00%	43.75%	58.33%	66.25%	19.90%
	2	43.75%	25.00%	25.00%	59.05%	67.25%	20.10%
	3	0.00%	6.25%	6.25%	55.82%	60.88%	21.81%
	4	18.75%	43.75%	25.00%	58.31%	67.00%	20.44%
B	1	37.50%	36.84%	43.75%	58.82%	69.25%	19.56%
	2	37.50%	31.58%	37.50%	59.08%	69.13%	20.01%
	3	0.00%	5.26%	0.00%	55.14%	60.00%	22.15%
	4	25.00%	26.32%	18.75%	58.60%	66.50%	20.57%
C	1	50.00%	38.89%	50.00%	60.44%	68.63%	19.03%
	2	12.50%	44.44%	18.75%	58.73%	69.63%	19.89%
	3	18.75%	11.11%	25.00%	56.84%	61.50%	21.28%
	4	18.75%	5.56%	6.25%	57.67%	64.25%	21.22%
Total times best	1	41.67%	33.96%	45.83%	**Final Average**		
	2	31.25%	33.96%	27.08%	58.07%	65.85%	20.50%
	3	6.25%	7.55%	10.42%			
	4	20.83%	24.53%	16.67%			
Best set of data		1	1	1			

Table 5.4 shows the average results by number of inputs and the average results of the models with a common input

PSO/FUZZY		Times better in choice			Average of indicators per		
		EP	EA	MSE	EP	EA	MSE
A	1	31.25%	25.00%	43.75%	64.19%	71.13%	17.36%
	2	31.25%	35.00%	31.25%	62.61%	71.75%	18.18%
	3	6.25%	0.00%	6.25%	59.80%	63.63%	19.72%
	4	31.25%	40.00%	18.75%	62.96%	70.50%	18.69%
B	1	37.50%	44.44%	56.25%	63.19%	71.75%	17.50%
	2	37.50%	27.78%	31.25%	62.54%	71.63%	18.17%
	3	6.25%	5.56%	12.50%	58.83%	64.63%	20.13%
	4	18.75%	22.22%	0.00%	61.92%	69.75%	19.52%
C	1	43.75%	38.89%	50.00%	64.83%	72.13%	17.00%
	2	25.00%	33.33%	31.25%	63.25%	72.75%	17.87%
	3	0.00%	0.00%	0.00%	58.58%	66.88%	20.24%
	4	31.25%	27.78%	18.75%	63.48%	73.75%	17.92%
Total times best	1	37.50%	35.71%	50.00%	**Final Average**		
	2	31.25%	32.14%	31.25%	62.18%	70.02%	18.53%
	3	4.17%	1.79%	6.25%			
	4	27.08%	30.36%	12.50%			
Best set of data		1	1	1			

Form the simulations above It can be concluded that The data sets from highest to lowest performance are 1, 2, 4 and 3. allows us to infer that pure signal processing is undoubtedly the best option, and that a signal averaged over 10 periods above do not have the information that is required in the classification process. The highest performing proposed model is number 1, followed by models 2, 7 and 3. All of these use three inputs and the sets of data 1 and 2. The proposed model with the lowest performance is number 14, followed by models 4, 10 and 12. These models had better results with data sets 3 and 4. Models with more than three inputs fail to overcome to the top four best models. This can be attributed to the size of the search space, since it is proportional to the number of entries in a fuzzy model, despite having a larger amount of data for classification. The model obtained with the best performance was obtained with genetic algorithms in model number 1, it had an MSE of 10% and it achieved a pure accuracy close to 88%. The most influential entries for classification success are EE, STE, and ZCR. Hit accuracy is a performance indicator for which a precisely optimized fuzzy model cannot be assured; pure accuracy is used for this purpose. The entry that best characterizes the gender of a voice signal is STE. The entry that least manages to characterize the genre of a voice signal is PSC. The GA, DE and PSO algorithms had close average performances. DE obtained the best average of performance indicators. The HS algorithm obtained the lowest performance. This means that the number of iterations that this requires to achieve convergence to global minimums must be greater than that of the other algorithms. The best options for the GA, HS, DE, and PSO algorithms are options C, C, B, and A, respectively. According to the interpretation of the best optimized model, the ZCR value tends to be lower in male voices. According to the interpretation of the best optimized model, a high value of ZCR together with a low value of STE better describes female voices. This scheme can be used to classify outputs other than gender and its description based on the inputs of the models as an advantage of its interpretability.

6. CONCLUSION

Speaker recognition is the process of automatically determining who is speaking by using the information contained in the sound signal. Speaker recognition is divided into speaker verification and speaker identification. Speaker verification is the process of determining whether a given voice sample belongs to the alleged person, and speaker determination is the process of determining which of the persons previously registered in the system. In addition, speaker recognition operations can be divided into two groups, independent of the text or dependent on the text. While different sentences or words are used in the training and testing stages of the system in text-independent speaker recognition systems, the same sentence or words are used in text-dependent systems both during training and testing. The usage areas of speaker recognition systems are quite common. For example, in recent years, it has been used in many fields such as telephone banking, voice calling, telephone shopping, database access services, remote voice control of computers and, one of the most important, forensic applications. As a pattern recognition problem, speaker recognition is a more difficult application compared to other problems, as the sound signal changes very frequently (not being stable) over time and is easily affected by factors such as ambient noise and weather conditions. The fact that the sound signal is not stable forces system designers to process the signal in short time intervals (10-20 ms) and this necessity causes the data sizes to be quite high. Large data sizes increase the working time of classifiers. Therefore, the classification method used in a speaker recognition system is expected to yield both high performance and fast results.

7. REFERENCES

[1] Alsulaiman, M., Alotaibi, Y., Ghulam, M., Bencherif, M. A., & Mahmood, A. (2010). Arabic speaker recognition: Babylon Levantine subset case study. Journal of Computer Science, 6, 381– 385. doi:10.3844/jcssp.2010.381.385

[2] Alsulaiman, M., Ghulam, M., Bencherif, M. A., Mahmood, A., & Ali, Z. (2013). KSU rich Arabic speech database. Information Journal, 16, 4231– 4254. http://catalog.ldc.upenn.edu/LDC2014S02

[3] Altıncay, H., & Demirekler, M. (2002). Why does output normalization create problems in multiple classifier systems? In Proceedings of CPR2002, 16th International Conference on Pattern Recognition, Quebec, Canada.

[4] Anusuya, M. A., & Katti, S. K. (2011). Front end analysis of speech recognition: A review. International Journal of Speech and Technology, 14, 99 – 145. doi:10.1007/s10772-010-9088-7

[5] Fukuda, T., & Nitta, T. (2003). A study on Japanese distinctive phonetic feature set for robust speech recognition. Autumn Meeting of the Acoustical Society of Japan, I, 9 – 10 (in Japanese).

[6] Fukuda, T., & Nitta, T. (2004). Orthogonalized distinctive phonetic feature extraction for noise-robust automatic speech recognition. IEICE Transactions on Information and Systems, E87-D, 1110– 1118.

[7] Gonzalez-Rodriguez, J. (2014). Evaluating automatic speaker recognition systems: An overview of the NIST speaker recognition evaluations (1996 – 2014). Loquens, 1, e007. http://dx.doi.org/10.3989/ loquens.2014.007

[8] Hassan, F., Kotwal, M. R. A., Rahman, M. M., Nasiruddin, M., Latif, M. A., & Huda, M. N. (2011). Local feature or mel frequency cepstral coefficients – Which one is better for MLN-based Bangla speech recognition? Springer-Verlag B, 2011, 154– 161, Berlin.

[9] Jayanna, H. S., & Mahadeva Prasanna, S. R. (2009). Analysis, feature extraction, modelling and testing techniques for speaker recognition. IETE Technical Review, 26, 181– 190.

[10] Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Communication, 60, 56 – 77. doi:10.1016/j.specom.2014.03.001

[11] Lawson, A., et al. (2011), Prague, Czech Republic Survey and evaluation of acoustic features for speaker recognition. ICASSP.

[12] Li, K. P., et al. (1966). Experimental study in SV using adaptive system. JASA, 40, 1441– 1449.

[13] Lopez-Moreno, I., Gonzalez-Dominguez, J., Plchot, O., Martı´nez-Gonza´lez, D., Gonzalez-Rodriguez, J., & Moreno, P. J. (2014). Automatic language identification using deep neural networks. IEEE International Conference on acoustics, speech and signal processing (ICASSP ’14), Florence, Italy.

[14] Mahmood, A., Alsulaiman, M., & Muhammad, G. (2012). Multidirectional local features for speaker recognition. ISMS 2012.

[15] Kota Kinabalu, Malaysia. Nitta, T. (1998). A novel feature-extraction for speech recognition based on multiple acoustic-feature planes. Proceedings of the IEEE ICASSP’98, I, 29 – 32.

[16] Nitta, T. (1999). Feature extraction for speech recognition based on orthogonal acoustic-feature planes and LDA. Proceedings of the IEEE ICASSP’99, I, 421–424.

[17] Qiu, A., Schreiner, C., & Escabi, M (2003). Gabor analysis of auditory midbrain receptive fields: Spectrotemporal and binaural composition. Journal of the Neurophysiology, 90, 456– 476. doi:10.1152/jn. 00851.2002

[18] Reynolds, D. (1995). Large population speaker identification using clean and telephone speech. IEEE Signal Processing Letters, 2, 46 – 48. doi:10.1109/97.372913

[19] Reynolds, D. A., Quatieri, T. F., & Dunn, R (2000). Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, 10, 19 – 41.

[20] Rosenberg, A. E., & Parthasarathy, S. (1996). Proceedings of the international conference on acoustics, speech, and signal processing, pp. 81 – 84 Speaker background models for connected digit password speaker verification.

[21] Senoussaoui, M., Kenny, P., Dehak, N., & Dumouchel, P. (2010). An i-vector extractor suitable for speaker recognition with both microphone and telephone speech. Proceedings of the Odyssey Speaker and Language Recognition Workshop, Brno, Czech Republic. Journal of Experimental & Theoretical Artificial Intelligence 11 Downloaded by [University of Nebraska, Lincoln] at 03:43 16 October 2015

[22] Sumithra, M. G., Thanuskodi, K., & Archana, A. H. J. (2011). A new speaker recognition system with combined feature extraction techniques. Journal of Computer Science, 7, 459– 465. doi:10.3844/ jcssp.2011.459.465

[23] Wildermoth, B., & Paliwal, K. K. (2000). Proceedings of the Australian International Conference on Speech Science and Technology (SST-2000), Canberra, Australia, pp. 324– 328 Use of voicing and pitch information for speaker recognition.

[24] Lucian Sulica. 2011. Hoarseness. Arch. Otolaryngol. Head Neck Surg. 137, 6 (June 2011), 616–619. DOI:http://dx.doi.org/10. 1001/archoto.2011.80

[25] Kirk P. H. Sullivan and Jason Pelecanos. 2001. Revisiting carl bildts impostor: Would a speaker verification system foil him? In Proceedings of the International Conference on Audio- and Video-Based Biometric Person Authentication (Lecture Notes in Computer Science), Vol. 2091. Springer. 144–149. DOI:http://dx.doi.org/10.1007/3-540-45344-X_21

[26] Bradford L. Swartz. 1992. Resistance of voice onset time variability to intoxication. Percept. Motor Skills 75, 2 (Oct. 1992), 415–424.

[27] Tiejun Tan. 2010. The effect of voice disguise on automatic speaker recognition. In Proceedings of the 3rd International Congress on Image and Signal Processing (CISP'10). IEEE. 3538–3541. DOI:http://dx.doi.org/10.1109/CISP.2010.5647131

[28] Shahrukh K. Taseer. 2005. Speaker identification for speakers with deliberately disguised voices using glottal pulse information. In Proceedings of the Pakistan Section Multitopic Conference. IEEE. 1–5. DOI:http://dx.doi.org/10.1109/INMIC. 2005.334384

[29] Oscar Tosi, Herbert Oyer, William Lashbrook, Charles Pedrey, Julie Nicol, and Ernest Nash. 1972. Experiment on voice identification. J. Acoustic. Soc. Amer. 51, 6B (June 1972), 2030–2043. DOI:http://dx.doi.org/10.1121/1.1913064

[30] Renetta Garrison Tull and Janet C. Rutledge. 1996. Automatic speaker recognition based on pitch contours. Proceedings of the Acoustical Society of America 131st Meeting—Lay Language Papers.

[31] Lior Uzan and Lior Wolf. 2015. I know that voice: Identifying the voice actor behind the voice. In Proceedings of the International Conference on Biometrics (ICB'15). IEEE, 46–51.

[32] Ratree Wayland, Scott Gargash, and Allard Longman. 1995. Acoustic and perceptual investigation of breathy voice. J. Acoustic. Soc. Amer. 97, 5 (May 1995), 3364. DOI:http://dx.doi.org/10.1121/1.413011

[33] Frederik Weber, Linda Manganaro, Barbara Peskin, and Elizabeth Shriberg. 2002. Using prosodic and lexical information for speaker identification. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP'02), Vol. 1. IEEE. 141–144. DOI:http://dx.doi.org/10.1109/ICASSP.2002.1005696

[34] Carl E. Williams and Kenneth N. Stevens. 1972. Emotions and speech: Some acoustical correlates. J. Acoustic. Soc. Amer. 52, 4B (March 1972), 1238–1250. http://www.ohio.edu/people/leec1/documents/sociophobia/williams_stevens_1972.pdf.

[35] Frank Wittig and Christian Mueller. 2003. Implicit feedback for user-adaptive systems by analyzing the user's speech. In Proceedings of the Workshop on Adaptivität und Benutzermodellierung in interaktiven Softwaresystemen (ABIS'03).

[36] Tian Wu, Yingchun Yang, Zhaohui Wu, and Dongdong Li. 2006. MASC: A speech corpus in mandarin for emotion analysis and affective speaker recognition. In Proceedings of the Speaker and Language Recognition Workshop (ODYSSEY'06). IEEE. 1–5. DOI:http://dx.doi.org/10.1109/ODYSSEY.2006.248084

[37] Zhizheng Wu and Haizhou Li. 2013. Voice conversion and spoofing attack on speaker verification systems. In Proceedings of the Signal and Information Processing Association Annual Summit and Conference (APSIPA'13). IEEE. 1–9. DOI:http:// dx.doi.org/10.1109/APSIPA.2013.6694344

[38] Naoaki Yanagihara. 1967. Significance of harmonic changes and noise components in hoarseness. J. Amer. Speech-Lang.- Hear. Assoc. 10 (Sept. 1967), 531–541.

[39] Eiji Yumoto. 1988. Quantitative assessment of the degree of hoarseness. J. Voice 1, 4 (Jan. 1988), 310–313. DOI:http://dx.doi. org/10.1016/S0892-1997(88)80003-1

[40] Eiji Yumoto, Wilbur J. Gould, and Thomas Baer. 1982. Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoustic. Soc. Amer. 71, 6 (June 1982), 1544–1549. http://www.ncbi.nlm.nih.gov/pubmed/7108029

[41] Elisabeth Zetterholm. 2003. Voice Imitation: A Phonetic Study of Perceptual Illusions and Acoustic Success. PhD Dissertation. Lund University, Lund, Sweden.

[42] Elisabeth Zetterholm. 2006. Same speaker—Different voices. A study of one impersonator and some of his different imitations. In Proceedings of the 11th Australian International Conference on Speech Science and Technology. 70–75.

[43] Elisabeth Zetterholm, Daniel Elenius, and Mats Blomberg. 2004. A comparison between human perception and a speaker verification system score of a voice imitation. In Proceedings of the 10th Australian International Conference on Speech Science and

Technology. Australian Speech Science and Technology Association, Sydney, Australia. 393–397. Retrieved from https://lup.lub.lu.se/search/publication/52907e52-0553-4228-a120-addc5e1f9d24.

[44] Cuiling Zhang and Bin Lin. 2017. Acoustic analysis of whispery voice disguise in Chinese. J. Acoustic. Soc. Amer. 141, 5 (2017), 3982–3982.

[45] Cuiling Zhang and Tiejun Tan. 2008. Voice disguise and automatic speaker recognition. Forensic Sci. Int. 175, 2–3 (April 2008), 118–122. DOI:http://dx.doi.org/10.1016/j.forsciint.2007.05.019

[46] Sue Anne Zollinger and Henrik Brumm. 2011. The Lombard effect. Curr. Biol. 21, 16 (Aug. 2011), R614–R615. DOI:http:// dx.doi.org/10.1016/j.cub.2011.06.003 Received September 2016; revised March 2018; accepted March 2018

[47] S. Kang, X. Qian, and H. Meng, "Multi-distribution deep belief network for speech synthesis," in Proc. ICASSP, 2013, pp. 8012–8016

[48] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," in Parallel distributed processing: explorations in the microstructure of cognition, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 194–281.

[49] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," Science, vol. 313, no. 5786, pp. 504–507, 2006.

[50] R. Salakhutdinov, "Learning deep generative models," Ph.D. dissertation, Univ. of Toronto, Toronto, ON, Canada, 2009.

[51] B. Kosko, "Bidirectional associative memories," IEEE Trans. Systems, Man, Cybern., vol. 18, no. 1, pp. 49–60, Jan. 1988.

[52] H. Oh and S. C. Kothari, "A pseudo-relaxation learning algorithm for bidirectional associative memory," in Proc. Int. Joint Conf. Neural Networks (IJCNN'92), 1992, vol. 2, pp. 208–213.

[53] M. Hattori, M. Hagiwara, and M. Nakagawa, “Quick learning for bidirectional associative memory,” IEICE Trans. Inf. Syst., vol. 77, no. 4, pp. 385–392, 1994.

[54] H. Zen, K. Tokuda, and T. Kitamura, “Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences,” Comput. Speech Lang., vol. 21, no. 1, pp. 153–173, 2007.

[55] T. Stafylakis, P. Kenny, M. Senoussaoui, and P. Dumouchel, “PLDA using Gaussian restricted Boltzmann machines with application to speaker verification,” in Proc. Interspeech, 2012.

[56] Y. Bengio, “Learning deep architectures for AI,” Foundat. Trends Mach. Learn., vol. 2, no. 1, pp. 1–127, Jan. 2009.

[57] G. E. Hinton, “A practical guide to training restricted Boltzmann machines,” in Neural Networks: Tricks of the Trade. New York, NY, USA: Springer, 2012, vol. 7700, pp. 599–619.

[58] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, “Distributed representations,” in Parallel distributed processing: explorations in the microstructure of cognition, D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA, USA: MIT Press, 1986, vol. 1, ch. 6, pp. 77–109.

[59] R. Salakhutdinov and G. E. Hinton, “Deep Boltzmann machines,” in Proc. Int. Conf. Artif. Intell. Statist., 2009, pp. 448–455.

[60] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” Speech Commun., vol. 27, no. 3, pp. 187–208, 1999.

[61] G. E. Hinton, “Products of experts,” in Proc. 9th Int. Conf. Artif. Neural Netw. (ICANN ’99) (Conf. Publ. No. 470), 1999, vol. 1, pp. 1–6, IET.

[62] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” arXiv preprint arXiv:1207.0580 2012.