UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING AND THE ENVIRONMENT

Institute of Sound and Vibration Research

Bio-inspired voice recognition for speaker identification

by

Konstantina Iliadi

Thesis for the degree of Doctor of Philosophy

October 2016

Abstract

Speaker identification (SID) aims to identify the underlying speaker(s) given a speech utterance. In a speaker identification system, the first component is the front-end or feature extractor. Feature extraction transforms the raw speech signal into a compact but effective representation that is more stable and discriminative than the original signal. Since the front-end is the first component in the chain, the quality of the later components is strongly determined by its quality. Existing approaches have used several feature extraction methods that have been adopted directly from the speech recognition task. However, the nature of these two tasks is contradictory given that speaker variability is one of the major error sources in speech recognition whereas in speaker recognition, it is the information that we wish to extract.

In this thesis, the possible benefits of adapting a biologically-inspired model of human auditory processing as part of the front-end of a SID system are examined. This auditory model named Auditory Image Model (AIM) generates the stabilized auditory image (SAI). Features are extracted by the SAI through breaking it into boxes of different scales. Vector quantization (VQ) is used to create the speaker database with the speakers' reference templates that will be used for pattern matching with the features of the target speakers that need to be identified. Also, these features are compared to the Mel-frequency cepstral coefficients (MFCCs), which is the most evident example of a feature set that is extensively used in speaker recognition but originally developed for speech recognition purposes.

Additionally, another important parameter in SID systems is the dimensionality of the features. This study addresses this issue by specifying the most speaker-specific features and trying to further improve the system configuration for obtaining a representation of the auditory features with lower dimensionality.

Furthermore, after evaluating the system performance in quiet conditions, another primary topic of speaker recognition is investigated. SID systems can perform well under matched training and test conditions but their performance degrades significantly because of the mismatch caused by background noise in real-world environments. Achieving robustness to SID systems becomes an important research problem. In the second experimental part of this thesis, the developed version of the system is assessed for speaker data sets of different size. Clean speech is used for the training phase while speech in the presence of babble noise is used for speaker testing. The results suggest that the extracted auditory feature vectors lead to much better performance, i.e. higher SID accuracy, compared to the MFCC-based recognition system especially for low SNRs. Lastly, the system performance is inspected with regard to parameters related to the training and test speech data such as the duration of the spoken material. From these experiments, the system is found to produce satisfying identification scores for relatively short training and test speech segments.

Table of Contents

List of Tables

List of Figures

Declaration of Authorship

I, Konstantina Iliadi,

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

"Bio-inspired voice recognition for speaker identification"

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:

   Conference paper

   Iliadi K. and Bleeck S., (2015). Speaker identification using auditory modeling and vector quantization, J. Acoust. Soc. Am., **138**, 1811.

Signed: Konstantina Iliadi

Date: 27 October 2016

Acknowledgements

First of all, I wish to express my gratitude and many thanks to my supervisor Dr. Stefan Bleeck for his supervision and help during various stages of the project. This research work would not have been possible without his guidance. I would also like to extend my thanks to my co-supervisor Dr. Ben Lineton for his help.

Special thanks to my lab colleagues and especially to Jessica Monaghan, Seon Man Kim and Xin Yang who, at various stages, have helped me in various interactions.

Outside the University of Southampton, my thanks go to Tom Walters for answering my questions on the box-cutting process and helping me find new directions for my project.

Finally, I wish to thank my parents and sister for their help and support during my studies. This work would have never been complete without them.

Definitions and Abbreviations

AIM Auditory image model

ASR Automatic speaker recognition

BMM Basilar membrane motion

DCT Discrete cosine transform

GPR Glottal pulse rate

MFCC Mel-frequency cepstral coefficient

NAP Neural activity pattern

SAI Stabilised auditory image

SNR Signal-to-noise ratio

SID Speaker identification

STI Strobed temporal integration

VQ Vector quantization

VTL Vocal tract length

Chapter 1

Introduction

1.1 Research objective

It is not uncommon that you receive a phone call where the caller starts talking and you realize you have the ability of immediately telling which acquaintance is speaking on the other end of the telephone. In this case, you have recognized the caller based only on one's voice rather than the content of one's speech. This is an example of speaker recognition that we perform in everyday life without realizing it. Another example is the impersonators that mimic voices of famous people for the entertainment industry. In both cases, people's voices are associated with their identities. The process of recognizing persons from their voices is called speaker recognition.

Speaker recognition is a natural procedure of human listeners. The most important parameter is the uniqueness of the human voice while the content of the speech usually does not play an essential role. The distinctiveness is because of various reasons such as the physiological differences of speech production organs like the larynx and vocal tract. Another additional factor that contributes to speaker recognition is the person's speaking style, i.e. the intonation or the use of specific words.

Many real-world applications benefit from speaker recognition such as access to buildings or telephone-based customer service systems. In general, there is an increasing need for person authentication in the world of information. Examples are applications ranging from credit card payments to border control and forensic investigations.

For decades, a person could be authenticated through two basic methods: something a person possesses, e.g. keys or credit cards and/or something a person knows, e.g. passwords or security questions. However, these ways have the drawback that items can be lost or stolen and passwords can be forgotten. Instead of them, there is another category of authentication methods that uses something that a person is such as fingerprint, voice, facial features or handwriting (www.biometrics.org). For example, one can use his/her voice and a built-in speaker recognition system can identify the speaker or verify if the speaker is the one being claimed. Another example is applying speaker recognition to an audio signal for forensic investigations in order to identify the persons of interest in the recording. This is called biometric authentication and it involves less problems given the fact that each person has a unique anatomy, physiology and habits. For that reason and given the increased computing power, the interest in the implementation of realistic biometric authentication has been increasing rapidly during the last years.

The design of automatic speaker recognition systems has been widely studied over the past few decades. A speaker recognition system comprises three processes which are feature extraction, speaker modeling and decision making. The feature extractor (or front-end) is the first component in an automatic speaker recognition system. Feature extraction transforms the raw speech signal into a compact representation that is more stable and discriminative than the original signal. Since the front-end is the first component in the chain, it determines the quality of the other components, i.e. speaker modeling and pattern matching parts.

Several feature extraction methods have been proposed for the speaker recognition task and most of them have been directly adopted from the speech recognition task. However, those tasks are contradictory by nature. Speaker variability is an error source for speech recognition whereas in speaker recognition, it is the type of information that needs to

be extracted.

Although several methods adopted from speech recognition work well, it is not always clear what kind of information the features capture from the speech signal. For that reason, it is necessary to use methods that provide some understanding of the extracted features. Another issue of automatic speaker recognition systems is the ability to achieve high performance in a well-matched condition. However, the performance drops significantly as speech is distorted by interference. In everyday environments, additive noise poses considerable challenges to such systems.

This thesis has two main purposes. Firstly, a new method for the feature extraction task is proposed and tested in conditions that simulate a quiet environment. After seeing the performance of the initial system design, a new technique is incorporated in order to go into the details of the feature extraction module. This procedure will help in further improving the existing system. Then, the developed speaker recognition system is evaluated in challenging conditions that simulate a real-world environment. Additionally, the second attempt of the thesis is to obtain insight, through this strategy, about which features are the important ones for successful speaker identification.

In general, the integration of this new method in the system design is the novel part of this work that also relates to acquiring knowledge of what is individual in a person's speech. Furthermore, the use of this knowledge helped in consolidating a system that combines good identification accuracy, robustness in the presence of noise as well as computational efficiency.

1.2 Outline of the thesis

The rest of the thesis is organized as follows. Chapter 2 presents the fundamental concepts of automatic speaker recognition and its applications. Firstly, an overview of the basic principles of ASR systems is given. Then, the methods that have been used for feature extraction, speaker modeling and decision making are reviewed. Additionally, the last section is a detailed overview of the existing literature about speaker identification systems in various environments.

Chapter 3 introduces the model that will be used as the front-end in the proposed speaker recognition system named Auditory Image Model (AIM). The model is a visual representation of all the stages that a sound goes through when it enters the human auditory system and it is explained analytically. In this research study, the MATLAB version of the AIM (aim-mat) that was written by Stefan Bleeck (and released in 2004) is used. Moreover, the techniques used for extracting the features from the auditory model are described in detail. The procedures involved in these techniques are part of research work by Lyon et al. (2010) and I worked on the MATLAB implementation of them.

Chapter 4 includes the first set of experimental work and it is about speaker identification in quiet conditions. The first part of this work consists of using a specific procedure in order to identify how the auditory model can be used, in the best possible way, to achieve satisfying levels of speaker identification accuracy. This knowledge is applied to the feature extraction step of the proposed system so that it can be further improved and compared to a baseline system that uses a state-of-the-art method as a front-end. This new process that has been incorporated as well as the acquired knowledge are original research that has been conducted during this study and I have worked on the MATLAB implementation of it.

Chapter 5 includes the second set of experimental work and it is about speaker identification in noisy conditions. After using the results of chapter 4 to improve the feature extraction stage, the developed design of the system is used for identification in real-world environments. The improvement of the feature extractor is based on the novel procedure used in the initial system design.

Finally, chapter 6 summarizes my conclusions, discusses the insights gained from this dissertation and makes suggestions on future research directions.

1.3 Contributions of the thesis

The auditory image is a model for the early stage representation of a signal entering the brain. This dissertation addresses the issue of robust SID from the perspective of using the image to extract features. In chapter 4, the design of the system consists of a feature extraction module, where the auditory model is combined with two methods that will be described analytically in chapter 3.

The first contribution is the incorporation of a new strategy for how the features of the speech material are used to create the speakers' reference templates. This method is used for extracting the features at specific positions in the auditory image. Except for its use for speaker matching, this approach to the problem of feature extraction allowed to analyze different parts of the SAI independently and resulted in specifying its most informative regions that indicate features that are more speaker-specific. This procedure is of importance since there were substantial redundancies in the SAI and it was essential to try and find a more dense representation of the signal with reduced data dimensionality.

Subsequently, the second contribution lies in the development of the box-cutting process and the improvement of the feature extraction stage. This has been done on the basis of identifying the patterns in the auditory image that are discriminative among speakers and selecting the area that includes them. The outcome has been a lower–dimensional SAI representation, which retains as much of the interesting information as possible, whilst creating an adequately compact feature vector that is also useful. Additionally, it makes a good comparison with the state-of-the-art feature extraction method, i.e. MFCC features, since their dimensionality is similar.

Chapter 2

Background

2.1 Automatic speaker recognition

Automatic speaker recognition (ASR) is a research area with significant development in terms of signal processing hardware and software. The purpose of automatic speaker recognition (or commonly termed as speaker recognition) is the extraction, characterization and recognition of information about the identity of a speaker.

2.1.1 Applications of speaker recognition

The main applications of speaker recognition technology can be summarized in the following list:

- Forensics
- Speaker authentication (or referred to as speaker verification)
- Environments with multiple speakers
- Personalized user interfaces

Forensic speaker identification is another important application of speaker recognition. If a voice sample has been recorded during the commitment of a crime, it can be compared with large datasets of speech samples of suspects in order to have an indication of similarity of the voices. This procedure can be very effective and contribute to both conviction and exoneration of suspects (Rose, 2002).

Person authentication (or verification) is one of the most obvious applications of any biometric system. Speaker recognition is already used by banks as an authentication method of the identity of customers (Myers, 2004). Additionally, it can be used in logging in electronic devices, entrances of physical facilities or border control. Also, if used as an authentication method, it can be combined with others like face recognition or fingerprints for more accurate results.

Multi-speaker environments can be found in many occasions and speaker recognition technology may be very useful when several speakers are included in the audio sample. Some examples are conference rooms, TV and radio broadcasting, court rooms and teleconferencing. These tasks that can be performed in such environments are speaker detection and speaker tracking. Speaker detection consists of deciding about the presence of a known speaker in a recording with multiple speakers while tracking is about locating, in the recording, a specific speaker's speaking activity.

Finally, personalized user interfaces are becoming more popular because of the developments in speech technology. An example is voicemail where the system could recognize the speaker and then adapt to one's needs or preferences.

2.1.2 Advantages and disadvantages of speaker recognition

One of the main benefits of speaker recognition is the fact that it is a natural process for people since speech is our main form of communication. As a result, it does not need the cooperation of the person that needs to be identified and it is non-intrusive from the user's standpoint.

Another advantage is the low cost of speaker recognition technology, given that the necessary equipment is usually a microphone compared to other systems such as fingerprint or retinal scanners. Additionally, the performance of ASR systems can be considerably high under specific conditions and speech seems to be the most popular tool in biometrics after fingerprint and face (Myers, 2004).

On the other hand, speaker recognition is not commonly accepted as a reliable authentication method. This is true to some extent given that a person's voice is not as unique as one's face or fingerprint. The fundamental difference is that the voice is more a behavioral biometric tool compared to the face or fingerprint, which are more physical features and can be measured directly from a person's body. A common example is impersonators. Impersonation or imitation is a type of voice disguise where the speaker tends to map one's voice in order to sound like another speaker. Disguise and imitation have been proven to degrade the performance of speaker recognition systems (Rose, 2002). Moreover, a speech wave is the outcome of movements of voice production body parts and can reflect their physical properties. For that reason, a speech wave cannot be produced exactly the same more than once.

As a conclusion, speaker individuality consists of several different parameters that supplement each other. However, it seems that humans use only a small subset of the available cues to identify a speaker. In general, speaker recognition technology may not be extremely reliable but it can be an asset if it is combined with other recognition methods for more robust results.

2.1.3 Elementary concepts and terminology

The most common characterization of automatic speaker recognition is the division into two different tasks: speaker identification and speaker verification tasks.

Speaker identification and verification

Speaker identification involves no identity claim and the system identifies the best speaker matched to the test speech signal. Speaker verification (also called authentication or validation) determines if the voice matches a particular registered speaker and it involves accepting or rejecting the identity claim of a speaker. The result is the probability of a match or a similarity measure. The similarity degree should exceed a certain (and predefined) threshold, which may be chosen to be the same for all speakers or speaker-specific. Ideally, the choice of the threshold(s) should be based on achieving a balance between the false acceptances (FA) (or referred as false positives (FP)) and the false rejections (FR) (or false negatives (FN)). False acceptance means that an impostor speaker is accepted while false rejection means that the correct speaker is rejected.

In general, speaker identification is a 1:N match where the voice is compared against N templates whereas speaker verification is a 1:1 match where a speaker's voice is matched to one template (also called "voice print" or "voice model"). Overall, the identification task is generally considered to be more difficult than the verification. This is instinctively reasonable: as the number of enrolled speakers increases, the probability of an incorrect decision increases as well (Doddington, 1985). Moreover, the performance of the verification task is, theoretically, not supposed to be influenced by the size of the population sample given that only two speakers are compared.

Open-set and closed-set identification

The speaker identification task is further classified into *closed-set* and *open-set identification*.

In the closed-set task, the speaker is determined from a set of registered (or enrolled) speakers and because of that limitation of the system, there is a risk of false identification. On the other hand, in the open-set identification, the speaker may not be in the database, which means that the target speaker is none of the registered speakers.

As a result, a closed-set system is used, firstly, to identify the speaker closest to the test speech data. Then, a verification system is used to compare the distance of this speaker with a chosen threshold and make a decision. The result can be either the identified speaker being accepted or an error message generated meaning that there is a “no-match” result.

Generally, the open-set identification is more demanding. In the closed-set task, the decision is made by the system simply on the basis of choosing the best matching speaker from the speaker database, despite the level of accuracy of the result. In the case of the open-set identification, there should be a predetermined threshold so that the similarity degree between the unknown speaker and the best matching speaker is within the threshold level. Also, speaker verification can be considered as a special case of the open-set identification with only one speaker in the database (N = 1).

Text-independent and text-dependent tasks

Another classification of speaker recognition is text-dependent and text-independent depending on the mode of operation. In the former case, the speech utterance that is used is known in advance whereas in the latter case, there are no assumptions about the speech sample.

In text-dependent recognition, the user is required to speak text that is spoken at enrolment (training stage) and can usually be a name, password or phrase. Text prompting is used in order to avoid fraud and the system requires the user to repeat again a random phrase or a list of numbers. Prompts can either be common across all speakers or unique.

In the case of text-dependent speaker verification, the utterance presented to the system can be either the same or different for every verification session. In the latter case, the system has a database of words or phrases and the user is prompted to utter the one that the system randomly selects. This procedure is called text-prompted speaker verification. The benefit of text prompting is that it makes it difficult for an impostor to know the utterance beforehand or use pre-recorded speech to play it back. Another option is that there may be a time lag between the utterance selection and the user uttering the phrase which makes it harder for an intruder to use a device that synthesizes the speaker's voice.

The text-independent system is non-invasive and does not require the user to actively answer prompts. It requires more training data and a longer enrolment phase. These systems are more frequently used because they require very little, if any, cooperation by the speaker. In this case, the text used during enrolment and testing is different. As a result, the enrolment can happen without the user's permission, which is the case for many forensic applications. In general, text-independent recognition is a more challenging task given that the text-dependent systems can be more accurate, since both the speech content and the voice can be compared.

2.1.4 Description of an ASR system

A speaker recognition system is based on the main principles of feature extraction and feature matching. Feature extraction relates to the extraction of important characteristics (or features) from a speech signal. The amount of data that will be extracted from the voice signal will be used to represent each speaker. An example of those characteristics is pitch, which is unique to different people. Other types of features will be described in the following section.

In text-independent systems, the features that appear in a person's speech are captured irrespectively of what one is saying. On the other hand, in text-dependent systems, the recognition of the speaker is based on the specific phrases that one is saying. Feature matching involves the process of identifying an unknown speaker through extracting features from one's voice and comparing them with those from a database of known speakers.

Generally, all speaker recognition systems have to complete two phases. The first stage is referred as the enrolment or training phase and the second one is the operation or testing phase. During the training phase, each registered speaker has to provide samples of their speech. The input speech is processed into two steps: firstly, the voice sample is condensed into numerical values that represent the characteristics of the vocal tract of each speaker and secondly, the data for every person are gathered into a single matrix that is called "template" or "codebook" and it is the reference model for that speaker. In the testing phase, the part of the architecture of the training system related to feature extraction is repeated.

In the case of speaker identification systems, the input speech signal is processed and then it is compared to the data stored in the codebook. The difference is used to make the recognition decision. For speaker verification systems, there is a threshold (speaker-specific) that also needs to be computed from the training samples and it is specified according to the system. During the testing phase, the input speech is matched with stored reference model(s) and the recognition decision is made. The basic structures of speaker identification and verification systems are shown in figures 2.1 and 2.2 respectively.

Speaker recognition systems can consist of different types of processing steps and those are chosen based on the application. Each system, text-independent or text-dependent, that may be used for identification or verification, has its own advantages and disadvantages. For that reason, different techniques may be required to create a system.

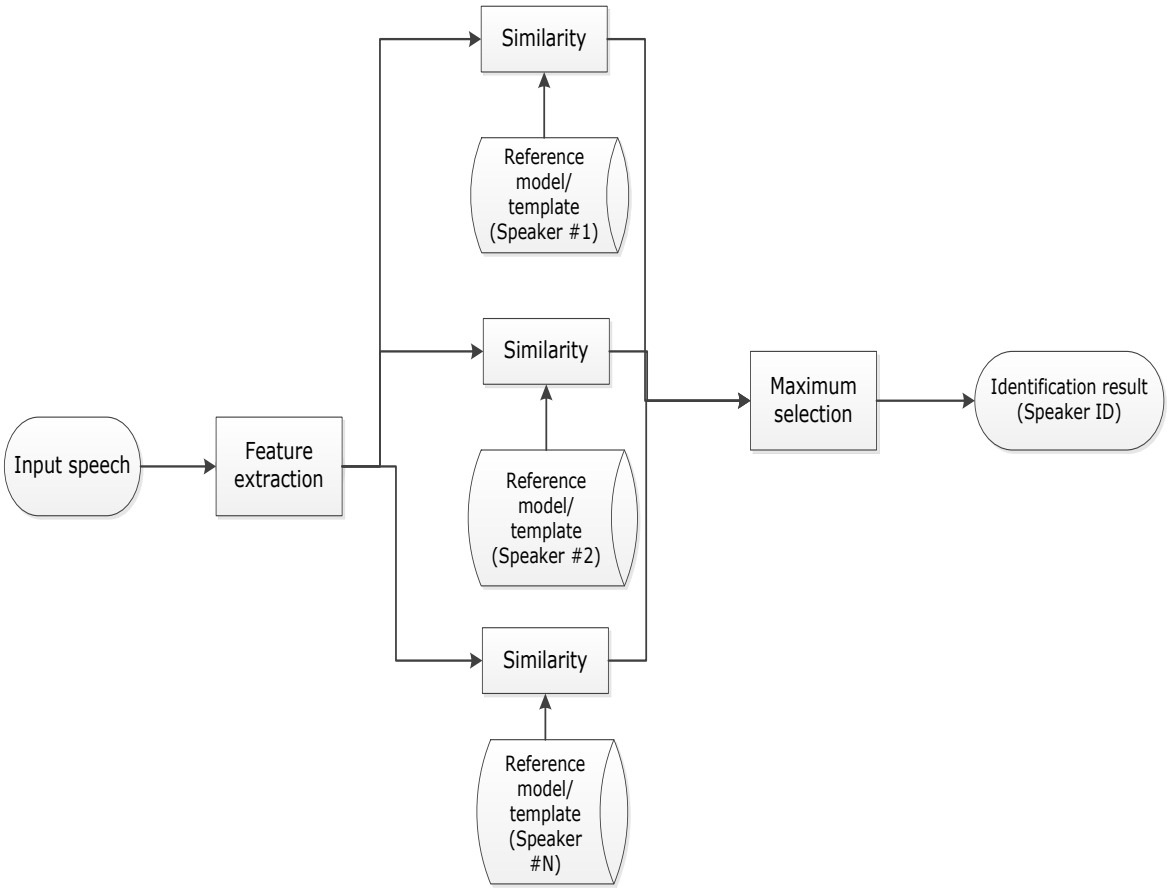


Figure 2.1: Block diagram of a speaker identification system

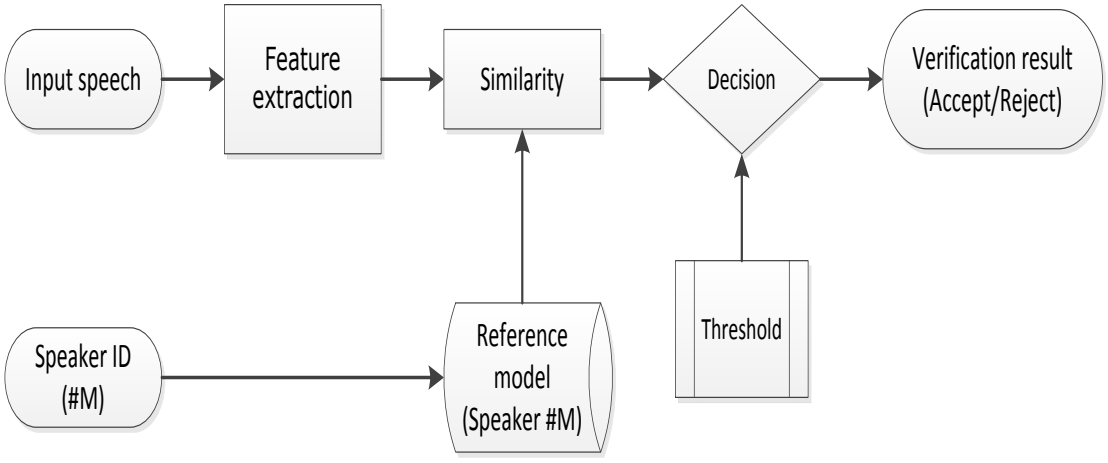


Figure 2.2: Block diagram of a speaker verification system

In the following chapters, different methods will be discussed for the development of a system that is classified as text-independent speaker identification system since its task is to identify the person speaking regardless of what is being said.

2.2 Feature extraction

Communication sounds among humans are produced by the resonance of a modified stream of air. The vocal apparatus is a term used in phonetics to designate all parts of human anatomy that can be used to produce speech. This includes the lips, tongue, teeth, hard and soft palates, uvula, larynx, lungs and others. Figure 2.3 shows a description of part of the voice organ that consists of the larynx and the vocal tract.

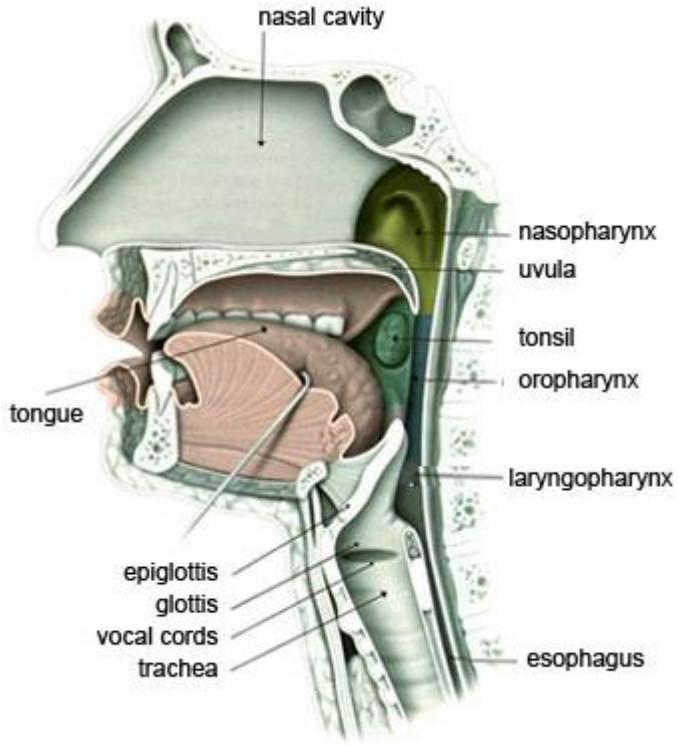


Figure 2.3: The human vocal apparatus (thebrain.mcgill.ca -)

Firstly, the air pressure from the lungs creates a steady flow of air through the trachea, larynx and pharynx. The vocal folds in the larynx open and close rapidly and that results in fluctuations in the air pressure known as sound waves.

These waves are modified by the resonances in the vocal tract according to the position and shape of the lips, jaw, tongue, soft palate and other speech organs. The resulting sounds have different features because of the variation in the resonances. Finally, the sound waves are scattered to the environment through the mouth and nasal cavities as speech.

From a speech production viewpoint, the vocal folds and glottis are the parts that are of interest. More specifically, the glottis is a small space between the vocal folds with triangular-shape. The airstream passes through the glottis to the vocal tract. The activity of the vocal folds forms the phonation type, whose major types are voiced speech (voicing), unvoiced speech (voicelessness) and whisper.

In the case of whisper and unvoiced speech, the vocal folds are apart from each other and the air flows from the lungs through the open glottis. The difference between them is defined by the extent of the glottal opening. When the glottal area is smaller, the result is air turbulence which generates the whispering sound. Inversely, when the area of the glottis is larger, there is much lesser turbulence of the airstream which results in the voiceless phonation.

In the case of voicing, the vocal folds open and close periodically. The coupling of the opening and closing phases interrupts the air flow from the lungs in a periodic manner and results in a series of glottal pulses that excite resonances of the vocal tract. This mechanism, which creates pulse-resonance signals, produces the voiced segments of speech. The resonances (or referred as formants) contain information about the vowel that is spoken. Also, these resonance patterns carry distinguishing information about the size of the body that produces them.

Vowels will almost always have four or more distinguishable formants but, the first two formants are most important in determining them. The first formant (F1) corresponds to vowel openness (vowel height). Vowel height is called the vertical position of the tongue relative towards the roof of the mouth or the aperture of the jaw. The second formant (F2) corresponds to vowel frontness (or backness). Vowel frontness (or backness) is called the position of the tongue during the articulation of a vowel relative to the front (or the back) of the mouth (International Phonetic Association (IPA)). On the other hand, the third (F3) and higher formants are assumed to be related more with speaker specificity. Additionally, Mokhtari (1998) suggested that the lower formants of vowels may carry speaker information as well despite the fact that their inter-speaker variation is smaller compared to that of the higher formants. As a result, phonetic and speaker information are mixed in the spectrum.

Furthermore, the perception of speaker size is related to the glottal pulse rate of the speaker (GPR), which is perceived as one's pitch (or fundamental frequency). Additionally, this type of information is conveyed to the listener by the frequencies of the resonances and their decay rates, which is directly related to the speakers' vocal tract length (VTL). Ives et al. (2005) conducted a series of experiments that showed that the size information in speech is available to the listener and when the VTL changes, it is possible to distinguish the difference in a person's size. For instance, a person with a longer VTL and a low GPR (which can be a male) is perceived as larger compared to someone with a shorter VTL and a high GPR (which can be a child). Additionally, Smith et al. (2005) demonstrated that VTL has a strong effect on the perception of speaker size. In their experiments, the participants commented on the speakers' height (on a scale from "very tall" to "very short") and on their gender and age (choosing from four choices: 'man', 'woman', 'boy','girl') given different combinations of GPR and VTL. The results demonstrated that the judgments about the size were strongly affected by the VTL and slightly affected by the GPR.

Nonetheless, these two properties may vary independently. For instance, speakers can change the pitch of their voice so people with different vocal tract lengths may speak the same phrase with the same pitch. Furthermore, speakers can also make small modifications of their VTL by lip rounding and by raising or lowering the larynx. These conditions alter the positions of the formant frequencies, which indicate if the vocal tract is shorter or longer, and this change in the vocal tract size results in a different impression about the speaker size. These factors can have an effect on the relation between formants and speaker height. As it has been shown before by Rendall *et al.* (2005), this relationship can be stronger for men but weaker for women whereas Gonzalez (2004) supported the inverse case. For the purpose of investigating the properties that make a speaker correctly identifiable, it is important to extract features that can show us which characteristics are necessary to process this task.

Feature extraction is the process of obtaining a set of parameters from the speech signal. Picone (1993) describes it as one of the basic operations of the signal modeling process that is also used in speech recognition systems. Signal modeling is the conversion process of speech signal into a set of parameters. Speech signals include different kinds of features but not all of them are useful for speaker recognition. Ideally, a feature should have the following characteristics in order to be important for discriminating a speaker (Kinnunen et al., 2009):

- Frequent occurrence in speech
- Large variability between speakers
- Small variability within a speaker
- Robustness against noise
- Difficulty in being imitated
- Ease in being measured from a speech signal
- Lack of influence from factors related to the speaker’s health or age

Another issue is the number of features that should be relatively low. Generally, the models that are used for the speaker templates in recognition systems cannot handle high-dimensional data. According to Jain et al. (2000), this happens because the number of required training samples for reliable representation grows exponentially with the number of features. This problem is called the curse of dimensionality. Also, low-dimensional features result in computational savings.

Features can be categorized in different ways. Kinnunen et al. (2009) has suggested a division of the features into the following categories that is based on their physical interpretation:

- spectral features,
- voice source features,
- prosodic and high-level features.

Generally, the choice of the features that one should use is a decision that depends on several factors. The aim of the application, the computing resources, the available speech databases and the cooperation of the speakers are some of them.

2.2.1 Spectral features

2.2.1.1 Short-term spectral features

Short-term spectral features are named in this way because of the fact that they are computed from short frames that last about 10-30 milliseconds. Speech signals change continuously because of articulatory movements, and therefore, they are broken down into frames of that duration.

Short-term analysis has been effective because of the quasi-stationary property of speech within this interval so that a feature vector can be extracted from each frame. These features usually describe the short-term spectral envelope, which is an acoustic correlate of timbre, i.e. the "color" of sound, as well as the resonance properties of the vocal tract. Reynolds et al. (2003) recommend the use of the short-term spectral features for speaker recognition research as they are easily computed and yield good performance.

Linear Prediction (LP) analysis

The logic in linear prediction (LP) analysis is based on the idea that adjacent samples of a speech waveform are highly correlated and thus, the signal can be predicted, up to a certain point, based on the past samples. The LP model assumes that each speech sample can be approximated by a linear combination of previous samples as shown in the following equation (Rabiner et al.,1993).

$$s[n] \approx \sum_{k=1}^{p} a_k s[n-k] \quad (2.1)$$

where *p* is the order of the predictor. The main target is to determine a set of predictor coefficients a_k so that the average prediction error or residual is as small as possible. The prediction error for the n^{th} sample is given by the difference between the actual sample and its predicted value:

$$e[n] = s[n] - \sum_{k=1}^{p} a_k s[n-k] \quad (2.2)$$

or alternatively,

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n] \quad (2.3)$$

When the residual $e[n]$ is small, equation 2.1 approximates $s[n]$ well. The total squared prediction error is defined as

$$E_n = \sum_n e[n]^2 = \sum_n (s[n] - \sum_{k=1}^{p} a_k s[n-k])^2 \quad (2.4)$$

The values of a_k that minimize the error E_n are estimated through the following equation

$$\frac{\partial E_n}{\partial a_k} = 0, \quad k = 1,2 \ldots p. \quad (2.5)$$

By expanding the equation 2.5 for different values of *k*, the optimal predictor coefficients are the solutions of the so called Yule – Walker or auto-regressive (AR) equations. Rabiner et al. (1993) described possible ways to solve the AR equations. Theoretically, any signal can be approximated with the LP model when the prediction error is small. The optimal model order depends on the type of information one wants to extract from the spectrum.

Rabiner et al. (1993) suggested the use of the frequency-domain interpretation of the LP in order to gain more insight. In that case, equation 2.1 can be turned into equality as follows

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + Gu[n] \quad (2.6)$$

where $u[n]$is the the excitation sequence and G is the gain with which the excitation signal is being scaled. In equation 2.6, the first term could be the feedback part of an IIR filter while the second term represents the input signal. Equation 2.6 is converted into the frequency domain through the Z-transform which results in the following equation:

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + GU(z) \qquad (2.7)$$

where a_k are the LP coefficients. The transfer function of the filter is obtained from equation 2.8 and is given by:

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1-\sum_{k=1}^{p} a_k z^{-k}} \qquad (2.8)$$

From the equation above, the poles and zeros of the filter will result in the parameters of the model for speech production. However, this filter has no zeros, and therefore, it is called an all-pole filter (Rabiner et al.,1993). Given that a pole corresponds to a local peak in the magnitude spectrum, the model has some restrictions as it appears that it models only the peaks of the spectrum (resonances of the vocal tract).

A generalized form of linear prediction is called Perceptual Linear Prediction (PLP). This type of analysis makes use of some of the psychoacoustics principles such as critical band analysis (in the Bark domain), equal loudness pre-emphasis and the intensity-loudness relationship *(*Hermansky, 1990). PLP (with 12 and 23 coefficients) has been successfully used in speaker identification but it seems to be

outperformed by the MFCCs (24 coefficients) (Reynolds, 1994) that will be described followingly. The reason for the lower performance of the PLP features is most likely to be the use of 17 bark-spaced filters for modeling the spectrum compared to the 24 mel-spaced filters used for the MFCCs.

Cepstral analysis

An alternative method to LP analysis is the cepstral analysis. In cepstral analysis, the magnitude spectrum is represented as a combination of cosine basis functions with varying frequencies. The cepstral coefficients are the magnitudes of the basis functions.

The cepstrum of a signal is defined as the inverse Fourier transform (IFT) of the logarithm of the spectral magnitudes and can be computed by the following equation:

$$c[n] = F^{-1}\{\log|F\ (frame)|\} \qquad (2.8)$$

where the coefficients *c[n]* are the Fourier series coefficients of the logarithm of the spectrum. In other words, the logarithm of the spectrum is expressed as an finite summation of cosines of different frequencies, and the cepstral coefficients are the magnitudes of the basis functions. The lower cepstral coefficients represent the slow changes of the spectrum while the higher coefficients the components of the spectrum that vary rapidly. In voiced speech, there is a periodic component in the magnitude spectrum which is the harmonic fine structure. The latter results from the vibration of the vocal folds. The slow variations are a result of the filtering effect of the vocal tract. The reason behind the choice of the logarithm of the spectrum is explained as follows.

As mentioned before, humans produce speech sounds through the vocal cords in the larynx that produce glottal pulses, which excite resonances in the vocal tract beyond the larynx. In modally voiced speech, it is assumed that the sound source is the airstream generated by the larynx while the vocal tract acts as a convolution filter. This is the source-filter model of speech production (Dudley, 1939).

Both of these components are inherently time-varying and assumed to be independent of each other. The cepstral analysis is useful for separating the excitation from the vocal tract shape. The concept is shown in figure 2.3.

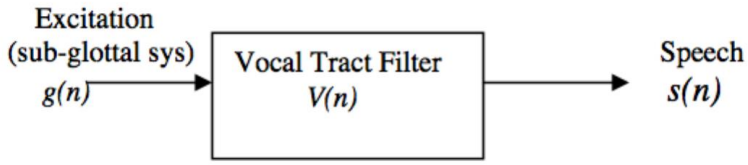


Figure 2.4: Source- filter model of speech production (Picone, 1993)

From the figure above, the speech signal is given by the following equation

$$s[n] = g[n] * V[n] \quad (2.9)$$

where $g[n]$ is the excitation signal and $V[n]$ is the vocal tract impulse response. In the frequency domain, the equation is transformed as follows

$$S(f) = G(f) \cdot V(f) \quad (2.10)$$

By taking the logarithms on both sides of the equation, it is obvious that the multiplicative components are converted into additive terms which is shown in equation 2.11.

$$\log|S(f)| = \log|G(f)| + \log|V(f)| \quad (2.11)$$

The equation above shows that the convolution of two signals can be expressed as the addition of their cepstra. The latter is a very important property because it can be applied to convert signals, combined by convolution (such as the excitation and the vocal tract filter), into sums of their cepstra, and consequently, they can be separated.

Atal (1974) introduced the concept of linear prediction cepstral coefficients (LPCCs) for speaker recognition. In this work, it was demonstrated that LPCCs were better than the linear prediction coefficients (LPCs) and other features such as pitch and intensity. The system performed text-dependent and text-independent speaker identification using 12 cepstral coefficients. Also, the system was tested for different speech utterances spoken by the same group of speakers for all of the 6 repetitions of the training and testing stages. Finally, it is worthy of note that the results were obtained for only 10 speakers, which is a small number of people.

Mel-Frequency Cepstral Coefficients (MFCCs)

The Mel-frequency Cepstral Coefficients (MFCCs) are one of the most popular features in speech and audio processing. Mermelstein (1976) is typically credited with their development. MFCCs are mostly used in speech recognition but they are also common in speaker recognition.

The Mel scale is defined as a perceptual scale of pitches that are equal in distance from one another (Stevens et al.,1937). A popular formula to convert hertz into mel is (O'Shaughnessy, 1987):

$$m = 2595 \log_{10}\left(1 + \frac{f}{700}\right) = 1127 \ln\left(1 + \frac{f}{700}\right) \quad (2.12)$$

MFCCs are coefficients that collectively make up a Mel-frequency cepstrum (MFC). The difference between the cepstrum and the Mel - frequency cepstrum is that in the MFC, the frequency bands are equally spaced on the Mel scale. This approximates the human auditory system's response more closely than the linearly spaced frequency bands used in the normal cepstrum. In summary, MFCCs are derived by the following steps :

- Estimate the power spectrum (through the Fourier transform) of short frames of the signal.
- Apply the Mel scale to the power spectra (using triangular overlapping windows) and sum the power in each filter.
- Estimate the logarithms of all of the Mel filterbank energies.
- Estimate the Discrete Cosine Transform (DCT) of the log Mel filterbank energies.
- The amplitudes of the resulting spectrum, i.e. the DCT coefficients, are the MFCCs.

Figure 2.5 shows a block diagram of all the steps followed for the derivation of the MFCCs.

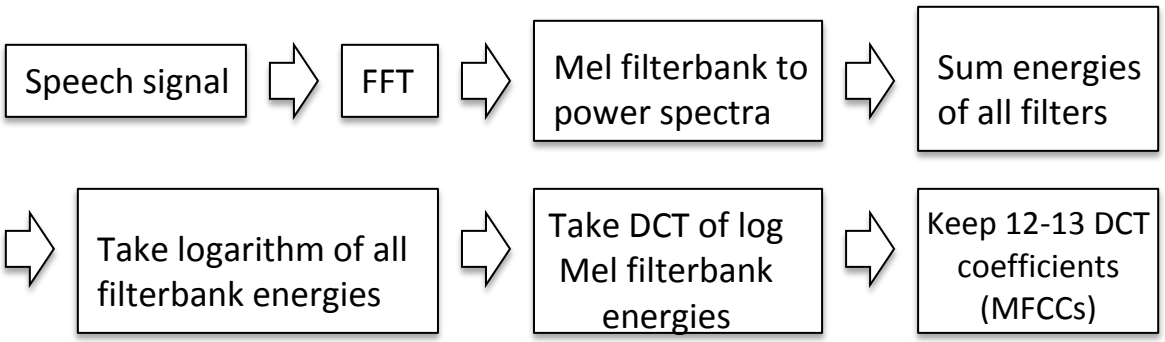


Figure 2.5: Block diagram that illustrates the derivation of the MFCCs

In most implementations, the standard number of filters that are used ranges from 26 to 40. Typically, the MFC is low-pass filtered and the first 13 coefficients are retained while the rest of them are discarded. Generally, the DCT is very effective in lowering the dimensionality of the spectrum in the MFCCs (Walters, 2011). Also, the cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal and contains information about the vocal tract of a speaker. This is displayed in the MFCC representation, when a change occurs in the size of the source, through the shift of the spectrum along the frequency channel dimension.

On the other hand, the MFCCs have a disadvantage related to the DCT. When the DCT is performed, the outcome is a number of cosine functions that oscillate at different frequencies and cannot change phase. Consequently, this restricts them to obtain their maxima at zero point. Ideally, when there is a change in the speaker size, the shift of the spectrum should be reflected on the cosine functions through a change in their phases while their amplitudes remain constant. Since this is not feasible, only the amplitudes of the cosines may vary and the DCT operation does not permit the cosine functions to shift with acoustic scale (Patterson, 2010). Considering that the cepstral coefficients carry a combination of information related to both the VTL and the type of vowel, this means that, for a certain VTL, the maxima of a specific cosine correspond to the formant peaks of a vowel but, when the VTL changes, they cannot shift to the new formant peaks. Furthermore, their performance for the speaker identification task drops significantly in the presence of noise, where there is a series of peaks at various frequencies.

Lastly, the MFCCs are widely used in speaker recognition systems because of the advantages that were described above. Since they are the state-of-the-art method for the speaker identification task, they will be employed, in this research study, for evaluating the proposed front-end in both quiet and noisy conditions.

2.2.1.2 Delta and delta-delta features

In the previous sections, each feature vector consisted of spectral parameters that are assumed to be a representation of a short-term quasi-stationary signal. However, there is no time information encoded in these features. As a person speaks, the articulators change position with a certain rate that depends on the speaking style, speaking rate and speech context (Kinnunen, 2004). These dynamic changes are shown on the spectrum as changes in the formant frequencies and can be indicators of the speaker.

A widely used method to encode some of the dynamic information of spectral features is through the estimation of the 1st and 2nd order time derivatives, which are called *delta* (Δ) and *delta-delta* (Δ^2) coefficients respectively (Rabiner, 1993). They are computed as the time differences between the adjacent vectors that contain feature coefficients usually added up with the initial coefficients on the frame level, yielding a higher-dimensional feature vector.

For instance, if 13 Mel-frequency cepstral coefficients are appended with their time derivative estimates, the dimensionality of the new feature vectors is 13 + 13 = 26 coefficients. Delta and delta-delta features are also known as differential and acceleration coefficients. The MFCC feature vector describes only the spectral envelope of a frame, but it is possible that speech could also have information in the dynamics of the MFCCs over time. To calculate the delta coefficients, the following formula is used:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2 \sum_{n=1}^{N} n^2} \qquad (2.13)$$

In the equation above, d_t is a delta coefficient, from frame t computed in terms of the static coefficients c_{t+n} to c_{t-n}. A typical value for *N* is 2. Delta-delta coefficients are calculated in the same way, but they are calculated from the deltas, not the static coefficients.

Overall, the short-term spectral features (LPC, LPCC, PLP, MFCC) and the delta features, which describe the spectral dynamics, are based on modeling the power spectrum. For the speaker identification task, the main issue in these features is to what extent they are affected by noisy effects and how easy it is to dissociate the speech characteristics from the noise. Specifically, in the case of delta (and delta-delta) features, it is possible that they are quite sensitive to noise. This happens because they are derivatives of static features and poor performance of static features may imply that its dynamics is not discriminative enough to correctly identify a speaker.

Finally, all of these features represent mainly the speaker-specific information due to the vocal tract. In the next sections, features that are related to speakers' glottal source and behavioural traits will be described.

2.2.2 Voice source features

The voice source (or glottal flow) features are the ones that characterize the glottal excitation signal of voiced sounds such as the fundamental frequency (or pitch) and the glottal pulse shape. The spectrum of the source generates the harmonic fine structure and spectrum roll-off whereas the vocal tract operates as a transfer function that modifies the overall spectral envelope (Kinnunen, 2004).

The fundamental frequency varies among individuals and can be considered as an element of individuality in the voicing mechanism. In general, it is widely known that adult males have lower pitches than young children and adult women (Smith et al., 2005). Nevertheless, the pitch can be controlled and altered by speakers. For that reason, it is not

always very useful for the process of speaker identification.

Furthermore, the shape of the glottal pulse depends on the how much the vocal folds open and the duration of the closing phase. These parameters are directly affected by the tension of the vocal folds and influence the quality of the voice that can be characterized as modal, breathy, creaky or pressed (Espy-Wilson et al., (2006), Laver (1994)). This type of information relates to voice quality and it can be useful in speaker recognition systems. However, it is a more instinctive idea of the differences among speakers and Eskelinen-Roenkae et al. (1999) showed that it is difficult to obtain a reliable measure of voice quality.

Generally, it is not possible to measure these features directly due to the vocal tract filtering effect. Plumpe et al. (1999) tried to estimate the glottal flow through using inverse filtering based on linear prediction. Even though the glottal features seemed to contain speaker-specific information, it appeared that their measurement is not easy, especially in the case of noisy speech.

As a conclusion, voice source features are not as discriminative as vocal tract features for speaker recognition but better accuracy could be obtained through their combination.

2.2.3 Prosodic and high-level features

Prosody refers to the suprasegmental aspects of speech. The term suprasegmental refers to the time span of the acoustic analysis. Suprasegmental parameters cover several segments of speech, like syllables, words and phrases, and reflect differences in speaking style, language background, sentence type and emotional conditions. Examples of prosodic features are intonation, stress, speaking rate and rhythm (Laver, 1994). The prosodic parameters that are typically measured are the intensity and the fundamental frequency.

In the case of speaker recognition, f_0 carries both physiological and behavioural traits. However, a drawback of the fundamental is the fact that it is a one-dimensional feature so it is not expected to be very discriminative. For that reason, variations of the f_0 have been used and they were assumed to carry important speaker information. Rose (2002) suggested that the mean value of f_0 can be considered as an acoustic correlate of the larynx size whilst the temporal variations of pitch are related to the manner of speaking. Shriberg et al. (2005) suggested the long-term temporal variations of f_0 while Carrey et al. (1996) showed the importance of long-term pitch and energy information for speaker recognition.

Additionally, the combination of features related to f_0 with spectral features may prove to be effective especially in noisy conditions. Peskin et al. (2003) demonstrated that the combination of prosodic features, like long-term pitch, with spectral features provided notable improvement as compared to only the pitch features. Except for the features that have been described so far, there are also other characteristics that relate to behavioural traits and were used for speaker recognition such as word duration, intonation, speaking rate and speaking style (Mary et al., 2004; Shriberg et al., 2005).

Furthermore, high-level features attempt to capture characteristics of speakers during a conversation. Doddington et al. (2001) initiated the use of a speaker's characteristic vocabulary, which is called *idiolect,* to characterize speakers. The reason is that speakers do not differ only in their timbre and pronunciation, but also in their lexicon, i.e. the type of words they tend to use in their conversations. The concept of the modeling process of these features was the conversion of each utterance into a sequence of tokens and the specification of patterns that occurred concurrently. The patterns of tokens characterized the differences among speakers. The tokens that were suggested were words (Doddington et al., 2001) or changes in pitch and energy (rising/falling) that had a more prosodic character (Shriberg et al., 2005). However, high-level features require more complex front-ends, such as automatic speech recognizer.

Overall, it is known by intuition that suprasegmental features carry speaker-related information. For instance, intonation, stress and rhythm vary among speakers. Also, these features can convey information to the listener about the speaker's emotional state, accent, dialect (or language) or social status. The main motivation for using them in speaker recognition is that they are not affected by noise and transmission lines as much as the spectral features (Reynolds, 2002). For that reason, they can be useful for telephone-based applications.

On the other hand, suprasegmental features are a challenge for speaker recognition systems given the fact that the speaker can control effects that influence them. One of their disadvantages is that they depend on the speaker's emotional state and attitude (Kinnunen, 2004). Furthermore, they are less discriminative and easier to impersonate. For example, the imitation of the pitch contour of the target speaker is a case that was demonstrated by Ashour et al. (1999). Hence, they are not considered to be as reliable as the spectral parameters. Lastly, another difficulty is their measurement. An example is the f_0 estimation using shorter segments of speech (i.e. in the order of milliseconds) that has to be preceded the estimation of the pitch contour (Kinnunen, 2004).

In conclusion, all of the different feature extraction techniques that have been described above can be summarized in three categories: spectral features that reflect speaker information due to the vocal tract, excitation source features and long-term features that represent speaker information because of behavioural traits. Among all of them, spectral features are the most widely used ones in speaker recognition. This possibly happens because there is less intra-speaker variability for spectral features and many tools for their analysis are available. Nevertheless, it could be beneficial for the feature extraction task to employ feature extraction methods for excitation source and behavioural traits but the main restriction is the lack of suitable tools.

In addition, spectral representations are severely affected by interfering noise in speaker identification systems, which results in degradation of their accuracy levels (Reynolds, 1994). As a result, it seems that it is important to investigate the use of new features that are more speaker – dependent and can be robust to distortions. The core part of this reseach study is to examine the use of features obtained by an auditory model, which will be described in the following chapter, that make the front-end of the proposed speaker identification system.

2.3 Speaker Modeling

The speaker modeling part performs a reduction of the feature data by modeling the distributions of the feature vectors. The objective of modelling techniques is to generate speaker models using speaker-specific feature vectors.

A speaker model is a reference template for every speaker that is trained. All the speaker models are stored in a speaker database. Such models will have enhanced speaker-specific information at reduced data rate. The purpose of this process is to use this database in order to make the final decision about the identity of the target speaker by comparing unknown feature vectors to all the models and selecting the best matching model.

The state-of-the-art speaker recognition systems employ different modeling thechniques. In this section, some popular speaker modelling techniques will be described. These methods have evolved concurrently throughout the years with the spectral features. In the first subsection, Vector quantization (VQ) will be explained analytically as it is the one that will be used for the design of the proposed speaker identification system. The second subsection consists of a summary of other methods that are widely used for this task.

There are two main approaches for estimating the feature distributions that are speaker-dependent: parametric and non-parametric approaches. In the parametric approach, a certain type of distribution is fitted to the training data by searching the parameters of the distribution that maximize some criterion. The non-parametric approach makes minimal assumptions about the distribution of the features.

Finally, the module of the decision logic is based on pattern matching. This step consists of computing a similarity score for the unknown speaker's feature vectors and all the speaker models in the database. The similarity (or dissimilarity) measure depends on the type of the speaker models.

2.3.1 Vector Quantization (VQ)

In this section, Vector Quantization (VQ) will be explained as it is the method that will be used for the purpose of this research study. VQ is a non-parametric method and it is one of the most popular approaches to text- independent speaker recognition. It is also known as *centroid model* and it was initially used for speaker recognition in the 1980s (Soong et al., 1987) but its main use was originally for data compression (Gersho et al., 1991).

In the case of speaker identification systems, VQ is used at the enrolment session for creating the speaker database. The speaker models are formed by clustering the extracted feature vectors of each speaker in K non-overlapping clusters. Theoretically, it could be possible to use all of the training feature vectors as the reference template for every speaker. However, for computational reasons, the number of vectors is usually reduced. Each cluster is represented by a code vector c_i, which is the *centroid* or *center of the cluster*. The resulting set of code vectors $\{c_1, c_2, \dots, c_k\}$is called a codebook, and it serves as the model of the speaker.

With regard to the codebook generation, the two important issues are the chosen method for generating the codebook and the size of the codebook (or equivalently the number of code vectors). The codebook size is significantly smaller than the training set. As a result, the amount of data is reduced but at the same time, the information of the original distribution is preserved (Gersho et al., 1991).

The function for speaker matching in a speaker identification system using VQ is estimated by the *quantization distortion.* Given a codebook $\mathrm{C} = \{\mathrm{c}_1, \mathrm{c}_2, \ldots \mathrm{c}_\mathrm{K}\}$ obtained from the training session and another set of feature vectors from the testing process $X = \{\mathrm{x}_1, \mathrm{x}_2, \ldots \mathrm{x}_\mathrm{T}\}$, the *quantization distortion* between a feature vector x_i generated from the target speaker and the codebook C is defined as:

$$d_q(x_i, C) = \min_{1 \le k \le K} d(x_i, c_k) \qquad (2.14)$$

where $d(\cdot,\cdot)$ is a distance measure that is mostly chosen to be Euclidean or Euclidean squared distance (Gersho et al., 1991). The code vector c_k for which the distance is minimized is the nearest neighbour of the feature vector x_i in the codebook C. Furthermore, the *average quantization distortion* is defined as the average of the individual distortions:

$$D_Q(X, C) = \frac{1}{T} \sum_{t=1}^{T} \min_{1 \le k \le K} d(x_i, c_k) \qquad (2.15)$$

The smaller the value of the distortion is, the higher the likelihood for X and C to originate from the same speaker. Figure 2.5 shows the concept of the VQ process for matching the feature vectors from the target (unknown) speaker with the codebook of a speaker from the database on the basis of achieving the minimization of the distance values between them.

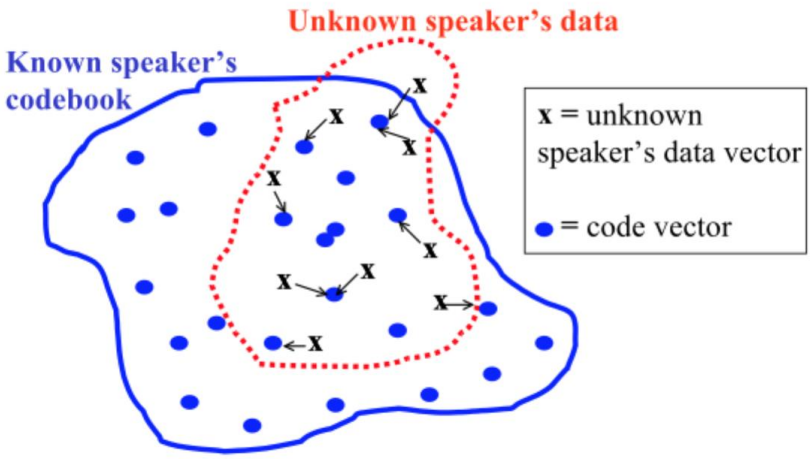


Figure 2.6: Illustration of the VQ-based speaker matching. The concept is to match the unknown speaker's feature vectors (indicated by x) with the neighbouring centroids of a known speaker's codebook (blue dots) that achieve as minimum distance as possible (Kinnunen et al., 2004)

For the generation of codebooks, there are two types of methods: unsupervised and supervised learning algorithms. In unsupervised learning, the speakers' codebooks are trained independently of each other and there is no human expert to assign the features to clusters. In this case, the feature distribution is the one that will determine the clusters given the fact that there is no supervisor to provide guidance. On the other hand, in supervised learning, there are overlaps between the codebooks and there is a supervisor that imposes on the data. Usually, the unsupervised methods are used since they include less parameters that need to be controlled by the user. The most popular unsupervised codebook generation algorithm, which is also one of the simplest ones conceptually, is the K-means algorithm. Figure 2.7 shows the concept of a codebook construction for VQ using K-means clustering and figure 2.8 is a diagram showing the steps of the algorithm for obtaining the reduction of dimensionality from N feature vectors to K centroids.

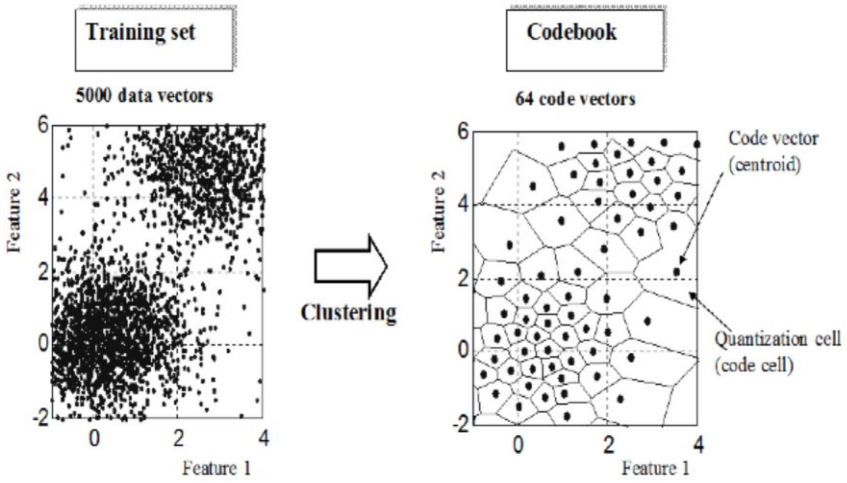


Figure 2.7: Codebook construction for VQ using the K-means algorithm. The original training set (left) consists of 5000 vectors and the features are reduced to a set of 64 clusters represented by their code vectors (centroids) (right) (Kinnunen et al., 2009)

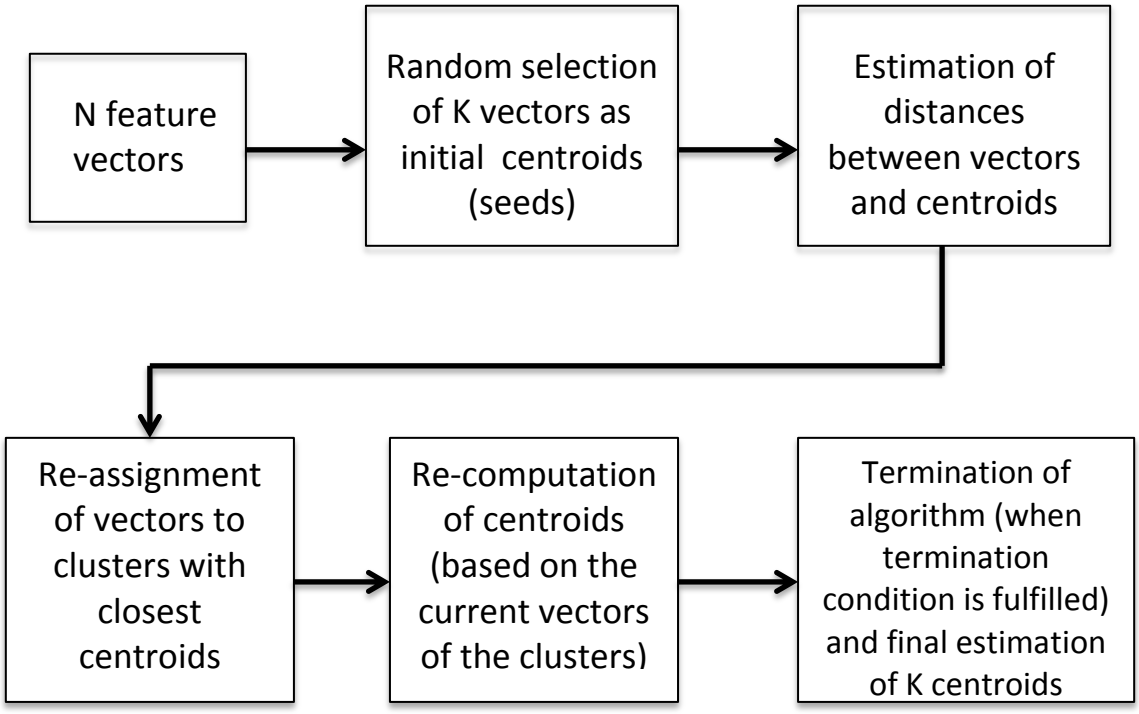


Figure 2.8: Block diagram of the steps of the VQ process (using K-means algorithm) for obtaining dimensionality reduction from N feature vectors to K centroids (N > K)

K-means algorithm

Clustering algorithms are used for grouping sets of features into clusters. The purpose of the algorithm is the creation of clusters that have internal coherence while they are different from each other. As a result, the data inside a cluster should be as similar as possible but, at the same time, they should be as dissimilar as possible from those in other clusters.

One type of distinction is between hard and soft clustering algorithms. Hard clustering means that each vector belongs only to one cluster. The opposite happens in soft clustering where each vector belongs to several clusters. This is called fractional membership. K-means is the most widely used hard clustering algorithm, which is also efficient (Manning et al., 2009).

The clustering process involves two important issues. The first is the choice of the number of clusters or *cardinality* of a clustering, which is denoted by K. In this case, the user of the algorithm needs to specify K, which is also the codebook size. Most of the times, K is usually guessed based on experience but there are also heuristic methods for choosing it (which will not described here as it is outside the scope of this thesis).

Then, the second issue is finding a good starting point for the clustering. In clustering, there are many possible partitions but it is hard to enumerate all of them until the best is found. Usually, the algorithm starts from an initial codebook of size K, which consists of randomly selected vectors from the training feature set. This is the first step of K-means and these initial cluster centers are called the *seeds*.

Afterwards, the algorithm iterates by repeating two steps until the codebook does not change. The first step is the re-assignment of the feature vectors to the cluster with the closest centroid and the second one is the re-computation of each centroid based on the current members of its cluster. The objective is to minimize the average squared Euclidean distance between the vectors and their centroids or, equivalently, to maximize the similarity between them.

In general, similarity and distance are both used to describe relatedness between features (Manning et al., 2009). The best option of a cluster, in K-means, is a sphere with the centroid being the center of gravity. Figure 2.9 by Manning et al. (2009) shows an example of the *K*-means algorithm. The codebook size is equal to 2. Then, the algorithm moves the centroids around in space through the two successive steps that have been described before so that average squared Euclidean distances are minimized. The algorithm converges after 9 iterations.

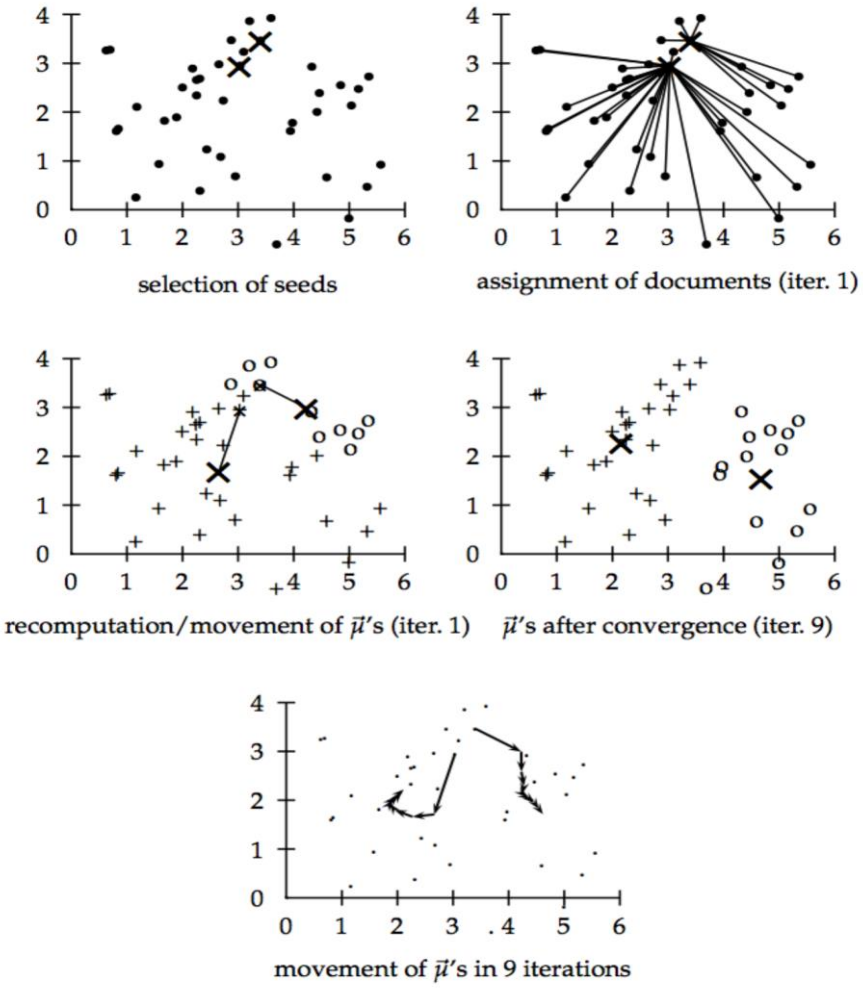


Figure 2.9: An example of the *K*-means algorithm for *K*=2 (Manning et al., 2009). The positions of the two centroids (indicated by X) move around the feature space until their distances with the feature vectors are minimized after 9 iterations.

2.3.2 Other speaker modeling techniques

Gaussian Mixture Model (GMM)

A Gaussian Mixture Model (GMM) is a parametric model of the distribution of features which has become a popular method in speaker recognition. It can be considered as an extension of the VQ model, but the main difference is that the clusters are overlapping. As a result, a feature vector is not assigned to the nearest cluster but it has a non-zero probability of originating from each cluster.

In the case of GMM-based speaker recognition systems, a speaker model consists of K Gaussian distributions parametrized by their prior probabilities or mixture weights w_i, mean vectors μ_i and covariance matrices Σ_i. The model is denoted as $\lambda = \{\lambda_1, \lambda_2, \dots, \lambda_K\}$, where $\lambda_i = \{w_i, \mu_i, \Sigma_i\}$ are the parameters of the i^{th} component (Reynolds et al., 1995). The parameters of the model are typically estimated using the Expectation-Maximization (or EM) algorithm, which is a generalization of the K-means algorithm. For the speaker modeling part, the target is the estimation of the parameters of the GMM λ that match the distribution of the training feature vectors. The matching function in GMM is defined in terms of likelihood and the aim is to find the model parameters that maximize it given the test data.

For a set of T training vectors $X = \{x_1, \dots x_T\}$ and the assumption of statistical independence between the vectors, the GMM likelihood is given by the following expression (Reynolds et al., 1995):

$$p(X|\lambda) = \prod_{t=1}^{T} p(\vec{x}_t|\lambda) \qquad (2.16)$$

where $p(\vec{x}_t|\lambda)$ is the Gaussian mixture density defined as:

$$p(\vec{x}_t|\lambda) = \sum_{i=1}^{K} w_i \, g(x|\mu_i, \Sigma_i) \quad (2.17)$$

where $w_i, i = 1, \ldots, K$ are the mixture weights or prior probabilities of each Gaussian component, and $g(x|\mu_i, \Sigma_i), i = 1, \ldots, K$ are the Gaussian component densities. The mixture weights (or prior probabilities) satisfy the constraint $\sum_{i=1}^{K} w_i = 1$. Each component density is a multivariate Gaussian function of the following form (Duda et al., 2000):

$$g(x|\mu_i, \Sigma_i) = (2\pi)^{-D/2} |\Sigma_i|^{-1/2} \exp\{-\frac{1}{2}(x - \mu_i)^{\mathrm{T}} \Sigma_i^{-1} (x - \mu_i)\} \quad (2.18)$$

where *D* is the number of variables, μ_i is the mean vector and Σ_i is the covariance matrix of each Gaussian mixture density.

There can be several variations on the GMM described in the above equation. Firstly, the covariance matrices can be full rank or constrained to be diagonal. Additionally, parameters can be shared among the Gaussian components, such as having a common covariance matrix for all components (Reynolds et al., 1995). The model configuration, i.e. the number of Gaussian components, the type of covariance matrices, and parameter tying is often determined by the amount of data available for estimating the GMM parameters and the application.

GMMs are often used in speaker recognition systems, due to their ability to represent large sample distributions and approximate densities that are randomly shaped. However, a disadvantage of GMM is that it needs sufficient data to model the speaker well (Reynolds et al., 1995). Additionally, there are many parameters that need to be adjusted by the user which makes an identification system more computationally complex.

Variations of Vector Quantization (VQ)

Another method to generate VQ codebooks for speaker recognition is the LBG algorithm (Soong et al., 1987) which gives good recognition performance for large test data and large codebooks. The algorithm is also known as Linde-Buzo-Gray algorithm according to its inventors (or generalized Lloyd's algorithm) (Linde et al., 1980). The concept of it is closely related to the k-means algorithm, i.e. it finds the centroid of each set in the partition, and then re-partitions the input based on which one of the centroids is closest. However, the difference in the LBG algorithm is that its input is a continuous geometric region rather than a discrete set of points. Therefore, in the re-partitioning step, the LBG algorithm uses Voronoi diagrams (a partitioning of a plane into areas based on distance to points in a specific subset of the plane) rather than only determining the nearest centroid to each one of a finite set of points (as the *K*-means algorithm does).

An alternative to VQ is Fuzzy Vector Quantization (FVQ), which has been used by Lin et al. (2006) as a classifier for speaker recognition. Its performance was very good because of its working principle, i.e. in FVQ, each feature vector is associated with all the clusters on a varying degree described by the membership function whereas in VQ, each feature vector has an association with only one of the clusters. However, the FVQ seems to be effective for little training data (Lin et al., 2006).

Support Vector Machine (SVM)

Another special theory for classification, is the Support Vector Machine (SVM). SVM is basically used for a two-class problem but it could be extended to solve a multi-class problem. The working principle of it is the mapping of the input data into a high-dimensional space until a separating hyper-plane, which maximizes the margin of separation between these two classes, is found. Also, it has been combined with

GMM for better results (Campbell et al., 2006). However, it is more popular for the speaker verification task since it is a binary classifier.

2.3.3 Discussion

Overall, there are a variety of techniques for the speaker modeling module. Given certain conditions, some of them are preferred compared to others. According to literature, in the case of text-independent speaker recognition, GMM is the mostly used one. Nevertheless, the EM-algorithm is quite complex from the implementation point of view. For instance, it is necessary that the minimum values of the variances of the Gaussian components are set in advance so that numerical problems are avoided. This means that very large values of the multivariate Gaussian density are computed (which is called overflow) when there is not enough training data or when using noise-corrupted data (Reynolds et al., 1995).

On the other hand, the VQ approach does not have the problem of controlling parameters. As mentioned before, its deficiency is that there is no cluster overlapping (and thus, the density function that is represented by the code vectors is not continuous.) However, it has been shown by Singh et al. (2000) that this approach gives comparable results with the GMM-based speaker recognition with much simpler implementation and computational efficiency.

Furthermore, several other approaches to speaker modeling in text-independent speaker recognition have been proposed, including, for example, support vector machines or variations of VQ that have been described above. Overall, the choice of the modeling technique depends on various parameters such as the nature of the task (speaker identification or verification) or the size of the training data. In section 2.5, an overview of existing speaker recognition systems will give more insight on how some of these methods perform when they are combined with specific feature extraction methods that have been explained in section 2.2.

After having reviewed some of the most commonly used feature extraction and speaker modeling methods for speaker recognition systems, the important question is how a new set of features can perform in the context of the speaker identification task. Thus, the focus of this research study is a new front-end, which will be described analytically in the next chapter. For that reason, it is salient to use a speaker modeling technique that is efficient and does not involve many adjustable parameters since the main attempt of this work is the evaluation of the novel feature extractor with regard to identification accuracy. As a result, the VQ-based approach seems to be more suitable for our system design because it combines competitive accuracy, simple implementation and computational speed. Also, it will not make the system design more complicated since it has less parameters that are controlled by the user.

2.4 Speaker matching and decision logic

In speaker recognition systems, the final step is the testing stage that involves matching and decision logic. At this stage, the feature vectors of the test speaker are compared with the speaker models in the database. The matching gives a score that indicates how close the test feature vectors are to the reference templates. The decision is taken on the basis of the matching score and it depends on the task (speaker identification or verification) in contrast to the feature extraction and speaker modeling procedures that are the same for different recognition tasks.

Given a database of N speakers, a set of speaker models $S = \{S_1, S_2, \dots, S_N\}$ and a set of the target speaker's feature vectors $X = \{\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{T}\}$, the matching score can be notated as $score(X, S_i)$ for the model of the i^{th} speaker.

For both identification and verification tasks, distance measurement techniques or probabilistic scoring are used in order to match the test feature vectors and the reference models. These include the estimation of Euclidean distance (Rosenberg et al., 1975) or variations of it such as the logarithm of distances (Atal, 1974). Another method was the angle measurement between vectors in order to evaluate their separation (Glenn et al., 1967). Also, Reynolds et al. (1995) used the concept of likelihood ratio for speaker identification which was mentioned in section 2.3.2.

In the case of closed-set speaker identification, the decision is directly the index j of the speaker that obtains the maximum score over the entire speaker database:

$$j = \arg(\max_i(score(X, S_i))) \qquad (2.19)$$

In the open-set identification task, the decision is given as follows:

$$decision = \begin{cases} j = \arg(\max_i(score(X, S_i))), if\ score(X, S_j) \geq T_j \\ no\ match, otherwise \end{cases} \qquad (2.20)$$

where T_j is the decision threshold that is chosen according to the application. Thus, the best matching speaker is found and if the score of this speaker is above the decision threshold, then this result is accepted. If not, the outcome of the recognition is that the speaker is no one. Furthermore, the performance of a speaker identification task can be computed directly, by the following equation, as the ratio of the number of correctly identified speakers to the total number of speakers that have been considered for the testing phase.

$$SID\ Accuracy\ (\%) = \frac{Number\ of\ correctly\ identified\ speakers}{Number\ of\ testing\ speakers} \times\ 100\%$$

(2.21)

On the other hand, in the case of verification, the decision is given as follows:

$$decision = \begin{cases} accept\ speaker\ j, if\ score(X, S_j)\ \geq\ T_j \\ reject\ speaker\ j, if\ score(X, S_j)\ < T_j \end{cases} \quad (2.22)$$

where T_j is the verification threshold that can be same or different across the entire number of speakers. The choice of threshold is made on the basis of trade-off between the false positives (FP) and the false negatives (FN). FP means accepting an impostor while FN refers to the rejection of the correct speaker. The performance of a speaker verification task is measured in terms of equal error rate (EER), which is defined as the error rate where the false positive rate (FPR) is equal to the false negative rate (FNR).

In the case of our research study, the operation of the proposed system falls into the category of closed-set identification. Consequently, the decision logic is based on the indexes of the speakers. Furthermore, the feature matching is achieved through the estimation of Euclidean distances. Lastly, the identification performance will be assessed through the SID accuracy decribed in equation 2.21.

2.5 Overview of speaker identification systems

In this section, existing approaches concerned with robustness of speaker identification (SID) systems are reviewed. As mentioned previously, speaker recognition systems comprise of two stages: the training (or enrolment) and the testing (or operation) phases. When speakers are trained, clean speech is usually used so that the codebooks are created. However, for the testing stage, the speech signals may be either in quiet environments or under noisy conditions.

Furthermore, the architecture of speaker recognition systems may have many variations. The speaker features that encode speaker specific characteristics can be extracted with different methods whilst the speakers can be modeled using various classifiers. Some of the combinations of the methods that have been implemented so far will be described in this section in order to understand the evolution of identification systems over the years. Additionally, there are other parameters that may affect the result of identification accuracy, such as the number of speakers being identified, the duration of the used speech material as well as the similarity or difference between the texts of the test and the reference samples (text-dependent ot text-independent identification).

2.5.1 SID systems in quiet conditions

First of all, some of the most important results that have been presented in literature for SID systems in quiet environments are reviewed in this section. Also, table 2.1 summarizes all of them in a chronological order (as they are described in the following paragraphs).

Pruzansky (1963) conducted the first identification study in 1963 using spectral energy patterns as features. They yielded good identification performance (89% as shown in table 2.1) and their usefulness for speaker recognition was confirmed. A study by Glenn et al. (1967) suggested that acoustic parameters that are produced because of nasal phonation can be highly efficient for speaker identification (97% and 93% correct identification for 10 and 30 speakers respectively as presented in table 2.1).

Furthermore, the concept of linear prediction was introduced by Atal in 1974. In his work, it was shown that LPCCs were better than LPCs for a group of 10 speakers. The accuracy of the system was also tested for different durations of test speech (0.5 and 2 sec as indicated in table 2.1). For the duration of 0.5 sec, the system performance was estimated in both text-dependent and text-independent conditions where it is obvious that this factor has an impact on the outcome (98% vs. 72% accuracy).

Reynolds (1994) studied and compared different features, such as MFCCs, linear frequency cepstral coefficients (LFCCs), LPCCs and perceptual linear prediction cepstral coefficients (PLPCCs) for robust speaker recognition. Among these features, MFCCs and LPCCs gave better performance than the others. Despite the fact that the MFCCs and LPCCs are used to extract the same vocal tract information, their performance differs because of the different principles involved in extracting them. In these experiments, the database consisted of 10 conversations from 51 adult male subjects that were divided into two different groups of 26 and 25 speakers (because the recordings took place in different locations, i.e. San Diego and New Jersey). Also, in this study, there was a special condition about the used database. For the group of 26 speakers, there was some unknown change in the recording equipment or telephone lines that resulted in an acoustic mismatch between the first five and second five recordings (for that reason, the speech material was divided into these 2 sets of five recordings) and that had an effect on identification performance.

In the case of the group of 26 talkers, it is obvious from table 2.1 that the accuracy was better when the training and test speech came from only one side of the divide whereas the performance dropped if the training and testing data came from different sides of the divide. The decision for the experimental process was to examine both the within the divide (WD) and across the divide (AD) performance. For consistency reasons, the conversations were split the same way for the group of 25 talkers even though there was no mismatch among the sessions.

Band energy values (i.e. energy values from certain frequency regions) have also been used as features by Besacier et al. (2000). The experiments were conducted for 630 speakers on the TIMIT and NTIMIT databases, which is the largest set of speakers that has been used in literature to our knowledge. The TIMIT corpus of read speech contains recordings of eight major dialects of American English, in which every speaker reads ten phonetically rich sentences. The corpus was designed by the Massachusetts Institute of Technology (MIT), SRI International (SRI) and Texas Instruments, Inc. (TI). NTIMIT (Network TIMIT) is a telephone bandwidth version of TIMIT (Linguistic Data Consortium). The accuracy percentage in table 2.1, i.e. 93.7%, is the best achieved result with relation to the number of speakers. However, it is worth mentioning that the system consisted of the two most popular methods for feature extraction and speaker modeling, i.e. MFCCs and GMM that have been proven to be very efficient for the SID task.

Speech parameters based on amplitude modulation (AM) and frequency modulation (FM) are another method which has been proposed by Grimaldi et al. (2008). In this study, it was demonstrated that speakers could be discriminated from different instantaneous frequencies (IF) on the basis of formants and harmonics in the speech signal. The importance of IF descents from the fact that speech is a nonstationary signal so the spectral characteristics vary with time.

Another case of a large set of 356 speakers (NIST 2003 speaker database) was used by Madikeri *et al.* (2011). The Mel Filterbank Slope (MFS) feature was investigated and compared to conventional MFCCs. This feature emphasizes formants as well as glottal shape and, therefore, indirectly uses the glottal pulse shape for speaker identification. Given the large number of speakers, the accuracy of this system (60.4%) may be smaller but it is noteworthy since it proposes an alternative feature extraction method.

Finally, Shirali-Shahreza et al. (2011) and Kumar et al. (2011) preferred the LBG algorithm (due to Linde, Buzo and Gray) to implement VQ for speaker modeling while using variations of MFCCs as features. The achieved identification accuracy was very good for both cases (100% and 80% correct recognition as shown in table 2.1). However, the systems have been tested on small datasets (32 and 10 speakers) whilst using the state-of-the-art method for extracting features.

In general, all of these systems have used spectral features like band energies, spectrum, cepstral coefficients etc that mainly represent the speaker-specific information due to the vocal tract. The majority of these features are based on some type of modeling of the power spectrum. If the details of the spectrum are not smoothed excessively or too much detail is not given to spurious spectral information, then, the performances of the features should not differ a lot. However, this is not the case as indicated by some of the accuracy levels in table 2.1. For instance, in the studies by Atal (1974) and Reynolds (1994), LPCC features (with 12 coefficients in all cases) perform differently for the different groups of 10, 26 and 25 speakers. The lack of similarity is more obvious for the groups of 26 talkers (99%(WD) and 80%(AD)), where the division of the speech data because of the channel and noise effects seems to be a restriction for these features since the performance drops significantly. In addition, given that these studies used the same classifier, i.e. GMM, it is reasonable to assume that the difference in performance may be a result of the choice of parameters involved in the GMM.

Consequently, it appears that the features can be importantly tied with the classification procedure and that can be a limiting factor for the achieved recognition levels. This is also shown by the results of the studies in table 2.1 that used MFCCs as the feature extractor (Reynolds (1994); Kumar et al. (2011); Shirali-Shahreza et al. (2011)) and different speaker modeling techniques (VQ and GMM). As a consequence, spectral features can be dependent on the employed classifier and that can lessen their effectiveness for speaker identification. Also, Reynolds (1994) showed that they are not immune to noisy effects and that will be further examined in the next section.

Affiliation	Front-end	Speaker Modelling	Number of speakers	Accuracy
Pruzansky (1963)	Spectral energy patterns	GMM	10	89%
Glenn et al. (1967)	Spectrogram patterns of [n](nasal phonation)	Direct template matching	10 30	97% 93%
Atal (1974)	LPCC	GMM	10	93% (text-indep/ 2 sec test speech) 72% (text-indep/ 0.5 sec test speech) 98% (text-dep/ 0.5 sec test speech)
Reynolds (1994)	MFCC LPCC LFCC PLPC MFCC LPCC LFCC PLPC	GMM GMM	26 25	**WD/AD** 100%/86% 99%/80% 100%/84% 95%/74% **WD/AD** 63%/66% 62%/59% 63%/64% 65%/55%
Besacier et al. (2000)	MFCC	GMM	630	93.7%
Grimaldi et al. (2008)	AM-FM (IF)	GMM	16	92% (3% error)
Madikeri et al. (2011)	MFS	GMM	356	60.4%
Shirali-Shahreza et al. (2011)	MFCC	VQ (LBG)	32	100%
Kumar et al. (2011)	MFCC	VQ (LBG)	10	80%

Table 2.1 : Summary of SID systems in quiet conditions

2.5.2 SID systems in noisy conditions

Noise robustness of speaker recognition systems has been widely studied over the past decade. Noise creates a mismatch between clean training data and noisy test data and makes the identification task more difficult.

Shao *et al.* (2007) proposed the Gammatone Filter Cepstral Coefficients (GFCC), from the combination of a gammatone filterbank and a binary T-F mask obtained through computational auditory scene analysis (CASA). The advantage of it is that the frequency spacing of the filterbank is closer to the cochlear filtering. Another case is the system of Zhao et al. (2012) where it was shown that GFCC outperformed the MFCC features and provided noise-robust SID. The system was tested on the 2002 NIST Speaker Recognition Evaluation (SRE) corpus that contains 330 speakers and it is one of the largest number of speakers that has been used for SID experiments. The performance of the system has been tested under different types of noisy conditions (stationary and non stationary) at various SNR levels from -6 to 18 dB (at 6 dB intervals). In order to assess if the noise robustness of the GFCC features was specific to the NIST SRE dataset, the experiment was repeated for another 330 speakers randomly chosen from the TIMIT corpus (Zhao et al.,2013) for only one SNR (0 dB) and one type of noise (factory noise). The benchmark performance for the baseline MFCCs was 3.94%, which was very low compared to GFCCs. As a result, it is obvious that GFCCs are much more noise-robust than MFCCs and that is consistent with the initial findings of Zhao et al.(2012).

Another concept that has been pursued is the addition of noise to clean training data. Ming et al. (2007) suggested the training of speakers in simulated noisy conditions. Models from different training conditions were combined during testing. However, in this study, there was no specific information about the noise that was used for the training data.

Another case where clean and noisy speech of 20dB was used for training was by Wang et al. (2010). In addition, they applied spectral subtraction before extracting MFCC features and combined them with phase information and deletion of frames with low energy/SN. The achieved levels of SID accuracy are satisfactory but the experiments have been implemented using high SNRs (i.e 10 and 20dB), which does not guarantee the robustness of the system for lower SNRs. Furthermore, given that for both sizes of speaker sets, the highest performances are attained for the same type (stationary/ nonstationary) and level (20dB) of noise (i.e. 98.7%(stat) and 96.9%(nonstat) for 35 talkers as well as 97.9%(stat) and 98.3%(nonstat) for 270 talkers as presented in table 2.2), it appears that matched training and test conditions upgrade the performance of SID systems despite the presence of noise.

The approaches discussed so far address noise in feature level through suggesting noise-robust features. Nevertheless, there are alternative approaches with new recognition methods in order to improve noise robustness such as missing feature methods (namely marginalization and reconstruction). The approach of missing data techniques focuses on the construction of a a time-frequency (T-F) mask to categorize each T-F point on whether is it dominated by speech or noise. However, they will not be explained further here as it is outside the scope of this thesis.

Pullella *et al.* (2008) exploited the combination of missing data with feature selection for robust SID. The training session involved clean speech mixed with stationary white noise of SNRs ranging from -5 to 20 dB while stationary white noise and nonstationary factory noise of the same SNRs were added to the testing speech. Furthermore, to identify which T-F units are reliable, Shao et al. (2008) employed the coupling of the GFCC features with the missing feature reconstruction method. Finally, Ming *et al.* (2007) adopted missing feature techniques to improve robustness by ignoring the heavily corrupted subbands and focus on the rest of them.

Nevertheless, an important limitation of the missing feature techniques is that they are efficient for partial noise corruption (Ming et al., 2007). As a consequence, the SID systems that use them may not be reliable enough for various real-world applications since this condition does not relate to realistic situations where noise affects most of the frequency content of speech signals.

All of the above methods that have been used in SID experiements are summarized in table 2.2, in chronological order, with details about the techniques used for feature extraction and speaker modeling as well as specific results about SID accuracy for specific noisy conditions.

From the table below, it is obvious that there is a variation of feature extraction methods that have been used in literature while the GMM seems to be the most popular choice for the speaker modeling module. However, most of these methods seem to be performing satisfactorily for SNRs above 0 dB while their performance degrades a lot for SNRs below 0 dB. Another parameter, as previously mentioned, is the size of the speaker database. When the number of speaker becomes very large, the achieved accuracy is high only for higher SNRs such as 18 or 20 dB. The only exception is the result obtained by the experiments of Zhao et al. (2012), for 330 speakers and -6 dB SNR, where the recognition score reached 29.85%.

Affiliation	Front-end	Speaker Modeling	Number of speakers	Training session	Testing session	Accuracy
Ming et al.(2007)	Multicondition training, Missing feature method	GMM	630	Unknown noise (simulated)	10dB (engine) 20 dB (restaurant)	26.43% 93.89%
Shao et al.(2007)	GFCC	GMM	34	Clean	-12 dB(ssn) 6 dB (ssn)	3% 97%
Shao et al. (2008)	GFCC, Missing feature method	GMM	34	Clean	-12dB(ssn) 6 dB (ssn)	9.83% 98.67%
Pullela et al.(2008)	Log-spectral features, Missing feature method	GMM	31	White noise (various SNRs)	-5 dB (factory) 20 dB (factory) -5 dB (white noise) 20 dB (white noise)	5% 80% 35% 80%
Wang et al.(2010)	MFCC, Phase information	GMM	35	Clean 20 dB (stat) 20 dB (nonstat)	20 (stat) 20 dB (stat) 20 dB (nonstat)	94.9% 98.7% 96.9%
Wang et al.(2010)	MFCC, Phase information	GMM	270	Clean 20 dB (nonstat) 20 dB (stat)	20 dB (nonstat) 20 dB (stat)	94% 98.3% 97.9%
Zhao et al.(2012)	GFCC	GMM	330	-6 (ssn) 18 (babble)		29.85% 90.61%
Zhao et al.(2013)	GFCC	GMM	330	0 (factory)		16.36%

Table 2.2 : Summary of SID systems in noisy conditions

2.6 Discussion

In this chapter, the main aspects involved in the design of a speaker recognition system have been examined methodically. At first, the fundamental concepts and terminology of the field of speaker recognition as well as its applications have been explained. Also, the general structure of recognition systems for both speaker identification and verification have been described. Then, the two basic modules of feature extraction and speaker modeling have been investigated as well as the concepts for measuring the performance of speaker identification and verification systems. Lastly, existing speaker identification systems that have been used in both quiet and noisy environments have been reviewed.

Firstly, the feature extraction methods that have been commonly used so far for the recognition task have been explained. These methods capture spectral characteristics of the speech signals representing information due to the vocal tract and they are based on the theory of the source-filter model. Based on this model, the spectrum of the source generates the harmonic fine structure and the spectral tilt whereas the vocal tract filtering effect modifies the spectral envelope. Overall, most of the proposed features for speaker recognition tend to focus on modeling the spectral envelope. However, apart from the smooth spectral shape, an important amount of speaker information is included in the spectral details (Kinnunen, 2004). Therefore, in addition to the spectral envelope, it seems that the key point is to consider the fast variations of the signal that are reflected in its fine structure. In this manner, it is possible to benefit from the information created by the voice source as well as from other parts of the speech production organs. In consequence, it is important to utilize a feature set that can involve such information.

In this study, the proposed feature extractor consists of a model that takes into consideration the temporal information of speech signals. The main attempt is to investigate if this kind of features may be more speaker-specific so that the influence of other parameters such as noise effects can be eliminated as much as possible. For that reason, the proposed speaker identification system will be evaluated in both quiet and noisy conditions.

In the case of speaker modeling, the proposed feature extractor will be combined with the VQ-based approach through using the K-means algorithm. As mentioned previously in section 2.3, the choice is based on its competitive accuracy, simple implementation and efficiency. Additionally, VQ is a non-parametric modeling approach so minimal assumptions are made for the feature distribution. This is important since the goal is not to optimize the classifier, but to evaluate the features.

Furthermore, after reviewing existing speaker identification systems, it seems that there is not yet best combination of a specific feature extractor and a speaker modeling method but the end result is a synthesis of various parameters, such as size of speaker population, choice of features, type of classifier as well as duration of used speech material. Consequently, it is another interesting question to investigate how the outcome of the identification system, which uses this novel front-end, is affected by the variation of different parameters.

In addition, another notable observation for all of the identification systems that have been described in section 2.5, is that they were not randomized in terms of the speaker populations. For all of the experiments, the enrolment and the testing sessions were repeated using the same groups of speakers. In each repetition, the speech material changed and the same speakers read different utterances. As a result, it seems that there is not enough knowledge on how a system may perform if the groups of speakers vary. This is another aspect of the research work in this thesis and it has been taken into consideration in the experiments that have been conducted.

In conclusion, speaker recognition is a challenging task depending on various parameters. This has become obvious from the existing approaches that have been reviewed in this chapter. The nature of speech creates difficulties, given that it is a non-stationary signal with spectral characteristics that vary with time. Depending on the message embodied in the speech itself and on the anatomy of the speaker's vocal cavities, the final result of the act of speaking is a very complex pressure signal containing many different forms of information.

Finally, in the following chapters, we will present our proposed method for speaker identification in both quiet and noisy conditions. The aim is to address speaker identification in realistic conditions that can be as challenging as possible according to the choice of parameters.

Chapter 3

The Auditory Image Model (AIM)

In chapter 2, an overview of the fundamental aspects of speaker recognition has been given. After reviewing some of the existing speaker identifications systems, it has become obvious that most of the popular feature extraction methods do not take into consideration all the speaker information included in speech signals. This happens because they focus on modeling the spectral envelope, which represents the smooth spectral changes, and they do not take into account the fast varying details represented by the fine structure of the signal (except from the delta features that may capture such information but are dependent on the static coefficients). For that reason, it is interesting to investigate how a new feature set, which considers the temporal aspect of speech, may perform as part of a voice recognition system that is used for speaker identification.

In this chapter, the methodology that has been followed in order to implement the feature extraction module of the proposed system is explained. Each one of the sections describe one of the processes that are used in a sequence in order to execute the feature extraction principle.

The first section decribes the auditory model named Auditory Image Model (AIM) that will be part of the novel front-end. The second section consists of the explanation of the two methods that complete the new feature extractor. These methods are called box-cutting and downsampling and they are the tools for extracting information from the AIM.

For both the training and the testing stages, the architecture of the proposed front-end is the same. All the modules that make up the auditory model will be described analytically in section 3.1. Section 3.2 consists of the details for retrieving the features from the auditory image.

3.1 The Auditory Image Model

The purpose of the feature extraction stage is to extract the speaker-specific information in the form of feature vectors at reduced data rate. The feature vectors represent the speaker-specific information due to one or more of the following: vocal tract, excitation source and behavioral traits. A good feature set should have a representation due to all the components of speaker information.

The Auditory Image Model (AIM) is considered to be a computational model in the time domain that describes some of the perceptions, which are produced in the hearing system, and can be associated with the ascending auditory pathway. These auditory perceptions in the brain result from the combination of sounds entering the ear canal with information and context from memory. The model allows us to illustrate the auditory image concept that is constructed by the brain. The acronym is used for both the conceptual model in the brain and the computational version of this model in order to create the auditory images of the incoming sounds.

Generally, an auditory image is produced when a sound is represented in terms of tonotopy and time-interval information and that representation can be considered to be an auditory image model. So far, there have been several models that have described the auditory image. This includes the autocorrelation models of Meddis *et al.* (1991) or Meddis *et al*. (1997) as well as models for the implementation of auditory images specifically (Patterson *et al*. (1992); Patterson *et al*. (1995); Patterson *et al*., 1996).

The important stages of the AIM are the spectral and temporal analysis performed by the auditory system. They consist of the frequency analysis of incoming sound performed by the cochlea, which sets up the tonotopic dimension of auditory perception, and the mid-brain analysis regarding the timing information in the individual frequency channels that create the time-interval dimension of auditory perception. A computational version of AIM, called aim-mat, consists of all the processing modules implemented in MATLAB (Bleeck et al., 2004). The functions of the AIM, which will be used in this study, are the following steps and they are described in the next sections:

1. the basilar-membrane motion (BMM) in the cochlea,
2. the neural activity pattern (NAP) in the auditory nerve and cochlea nucleus,
3. the identification of the "important" peaks (named "strobe points") of the signal and the preservation of its fine structure for every identified peak in order to construct the auditory image (Strobe Temporal Integration),
4. the stabilized auditory image (SAI), which is the fundamental description of the auditory perception.

Table 3.1 describes the architecture of the AIM and how it relates to the physiology and signal processing of the human auditory system.

Process	Frequency analysis	Sharpening	Feature detection	Temporal integration
Physiological structure	Cochlea	Brainstem/Thalamus/Cortex		
aim-mat module	BMM	NAP	STI	SAI

Table 3.1: Architecture of the AIM related to the physiology and signal processing analysis of the auditory system (Bleeck et al., 2004).

3.1.1 Basilar membrane motion (BMM)

For any sinusoidal stimulation, the response of the basilar membrane is a traveling wave which moves along the BM from the base towards the apex. The amplitude of the wave increases at first and then decreases abruptly. For different frequency sounds, the response is strongly affected by its mechanical properties that vary from base (narrow and stiff) to apex (wider and much less stiff) (Moore, 2003). The frequency of stimulation is the one that defines the position of the peak in the vibration pattern. Maximum displacement of the BM near the base (with little movement on the rest of the membrane) is produced by high frequency sounds whilst for low frequency sounds, the maximum is reached before the apex (with a vibration that extends along the BM).

Generally, sounds of different frequencies produce maximum displacement at different places along the BM. The frequency that corresponds to maximum response at a specific point on the BM is known as the characteristic frequency (CF) for that place (Moore, 2003). Each point on the BM can be considered as a bandpass filter with a specific centre frequency (corresponding to the CF) and bandwidth (Moore, 2003).

According to the previous description about the shape and resonance of the BM, it appears that there is a tonotopic organization of the sensitivity to different frequency ranges. The latter can be modelled with a bank of band-pass filters that overlap. These filters are known as auditory filters. Each different point along the BM corresponds to a filter with a different center frequency and they determine the frequency selectivity of the cochlea, which relates to the ability of a listener to discriminate between different sounds. For instance, in the case of a noisy background, the listener is trying to detect a signal by using a filter with a center frequency close to that of the signal.

In our case, the BMM module simulates the spectral analysis of the auditory system using a linear, gammatone auditory filterbank. This type of auditory filter will be explained in the next section. Lastly, the auditory system performs the compression that takes place on the BM within the auditory filters, i.e. the outer hair cells (OHCs) are used as mechanical feedback to the signal conveyed by the inner hair cells (IHCs).

3.1.2 Gammatone filter

In the time-domain auditory models, a bank of gammatone auditory filter simulates the frequency analysis of the basilar membrane (Patterson et al., 1992; Patterson et al., 1995; Patterson et al., 1996). The impulse response of the gammatone is

$$g_t\ (t) = at^{n-1} \exp(-2\pi ERB(f_c)t) \cos(2\pi f_c t + \varphi)\ (t > 0) \quad (3.1)$$

where $\alpha, n, f_{c,}\ \phi$ □ are parameters for the amplitude, the filter's order, the center frequency and the phase of the carrier correspondingly. ERB is the equivalent rectangular bandwidth of the filter. According to the analysis by Glasberg et al. (1990), the bandwidth of the filter corresponds to a fixed distance on the BM. The following equation describes the ERB of the human auditory system (Glasberg et al., 1990).

$$ERB\ (f_c) = 24.7 + 0.108 f_c \quad (3.2)$$

The filter is named "gammatone" because of the fact that the envelope formed by the power function and the exponential is a gamma distribution function. Also, the cosine carrier is a tone in the auditory range. The gammatone filter is a linear filter and the amplitude characteristic is

approximately symmetric on a linear frequency scale. Figure 3.1 shows the impulse response of the gammatone filter.

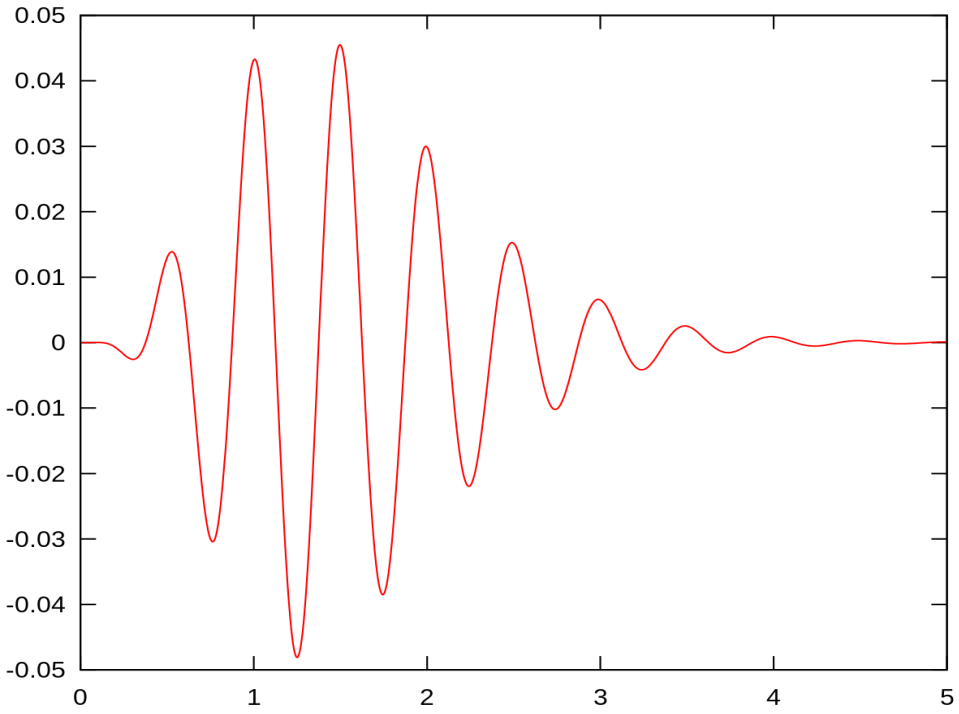


Figure 3.1: The impulse reponse of a gammatone filter

The channels of the auditory filterbank correspond to the number of auditory filters that are supposed to be a continuous series of overlapping filters. So far, there has been a suggestion for 24 successive critical bands that correspond to the 24 1/3 octave bands that the human ear can perceive. With regard to AIM, the gammatone fitlerbank will be used for the purpose of our experiments. The number of filters will be determined in chapter 4.

3.1.3 Neural activity pattern (NAP)

Firstly, the neural activity is simulated using the default module ‘hcl’ that consists of three operations: the half-wave rectification, the compression and the low-pass filtering.

In the case of half-wave rectification, either the positive or the negative half of the sound wave is passed while the other half is blocked. In the case of the AIM, it is used to represent the unipolar response of the hair cells. The next operation is the compression that simulates the cochlear compression. This means that a large range of input sound levels is compressed into a smaller range of responses on the basilar membrane. The compression is a process that helps the auditory system cope with the large dynamic range of sounds. The default process is the square root compression since that is closer to what the auditory system applies. However, it is also possible to use logarithmic compression since the humans do not perceive loudness on a linear scale and this is the one that will be used for this research study.

The third operation involved in the simulation of NAP is the low-pass filtering, which corresponds to the progressive loss of phase locking as the frequency increases above a certain value. In AIM, the standard low-pass filter has a cutoff frequency at 1200Hz.

3.1.4 Strobed temporal integration (STI)

The next stage of the AIM is the specification of important peaks in the NAP that are called strobe points. The process has been created based on knowledge from previous research that showed that the fine timing information in the NAP is carried on to the later stages of the auditory pathway (Patterson, 1994b). For that reason, Patterson et al. (1995) suggested that the auditory temporal integration cannot, in general, be simulated by a temporal average process, since averaging over time destroys the temporal fine structure within the averaging window. Also, Patterson et al. (1992) supported that it is the fine structure of periodic sounds that is preserved rather than the fine structure of noises.

As a result, they showed that this information could be preserved by, firstly, finding peaks in the neural activity as it flows from the cochlea and then, measuring time intervals from these strobe points to smaller peaks. The final step is to form a histogram of the time-intervals, one for each channel of the filterbank. These steps form the Strobed Temporal Integration (STI) process (Patterson et al, 1992).

In STI, every channel has a strobe unit that monitors the level of activity instantly. When a large peak is identified, the entire record in that channel of the buffer is transferred to the corresponding channel of a static image buffer, where the record is added, point for point, with whatever is already in that channel of the image buffer (Bleeck et al., 2004). The SAI module uses the strobe points to convert the NAP into an auditory image.

The model includes the strobe finding algorithm named ‘sf2003’. The algorithm uses an adaptive threshold in order to isolate strobe points. A strobe is defined when the neural activity rises above that threshold. The time of a strobe corresponds to a peak of the NAP pulse. After issuing a strobe, the threshold initially starts increasing on a parabolic path and then, follows a linear decrease to avoid spurious strobes. The duration of the parabola depends on the center frequency of the channel and its height is proportional to the height of the strobe point. After the parabolic section of the adaptive threshold, its level decreases linearly to zero in 30 ms. Figure 3.2 shows the mechanism of STI in order to identify the strobes of a signal and figure 3.3 presents the steps of the process that takes place for every frequency channel in order to construct the auditory image.

In AIM, the STI creates the time-interval dimension of the auditory image, which is different from the time dimension. For periodic or quasi-periodic sounds, the STI matches the integration period to the period of the sound and it produces a stable image of the repeating temporal pattern. These patterns are vertical ridges that relate to the repetition rate of the source and indicate where the resonances start in each frequency channel. In

this manner, it is possible to separate, in the auditory image, the glottal pulse rate from the resonance pattern because of the vocal tract.

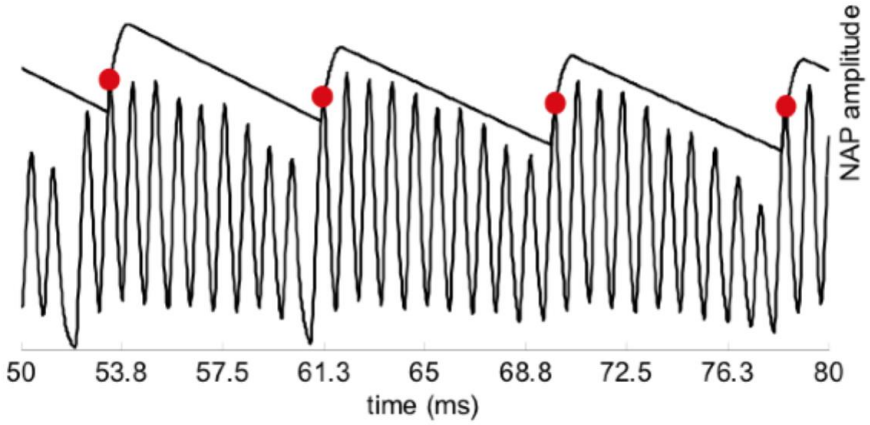


Figure 3.2: The mechanism of Strobed Temporal Integration (STI) (Walters, 2011)

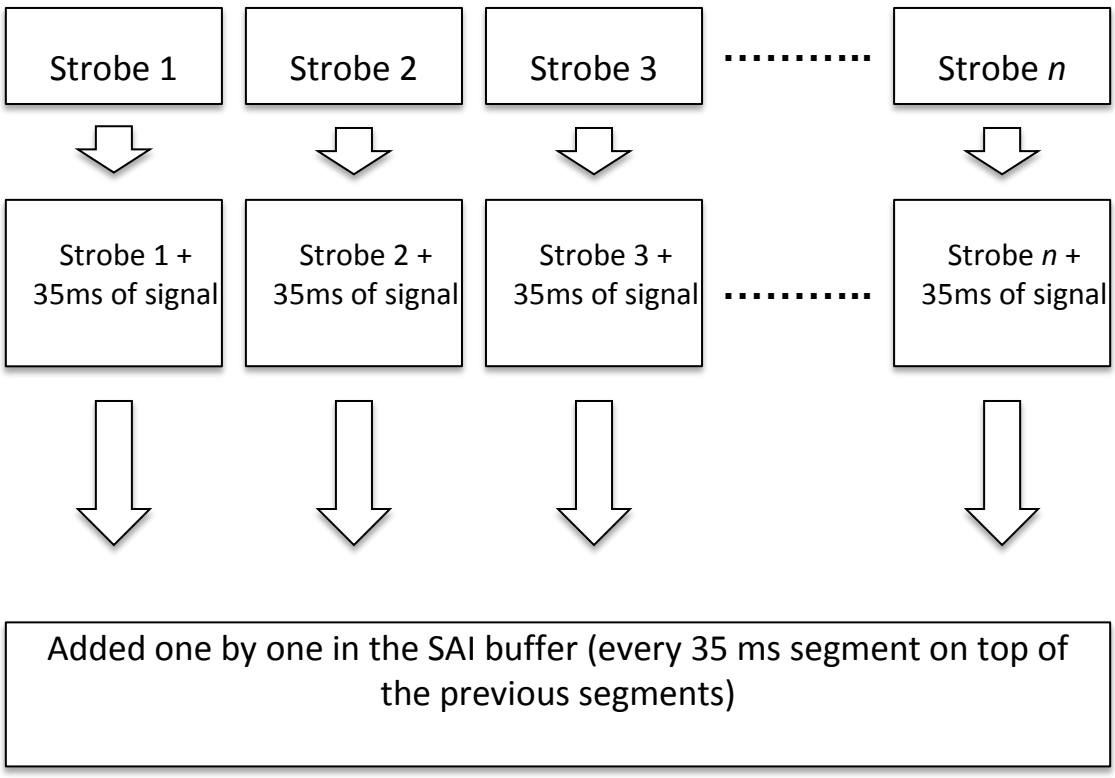


Figure 3.3: The steps of the process for specifying strobes and using them for the construction of the auditory image. This process is repeated several times for all frequency channels of the filterbank in the SAI.

The STI can be considered as an altered type of autocorrelation. In the case of autocorrelation, a signal is cross-correlated with itself at different points in time. The outcome is symmetrical to the zero point. On the other hand, in the case of STI, the signal is cross-correlated with a function that is zero everywhere except at the strobe points. The advantage of it is that it is less computationally complex than autocorrelation since one of the signals contains mostly zeros (Walters, 2011). Additionally, the outcome is not symmetrical to the zero point and any pattern of temporal asymmetry of the incoming signal is preserved. The benefit of these asymmetrical structures that are created in the image is that they are indicative of characteristics of pulse-resonance sounds (Patterson et al., 1998). Lastly, the identification of strobe pulses and the formation of the histogram of time-intervals (one for each channel of the filterbank) enable the segregation of the pulse and resonance information which helps to retrieve information in both temporal and spectral context.

3.1.5 Stabilized auditory image (SAI)

After the completion of the STI, the SAI module uses the strobes to convert the NAP into an auditory image. The SAIs are generated through the ‘ti2003’ algorithm which operates in the following steps.

Firstly, the temporal integration process starts when a strobe is issued and the NAP values, following the strobe, are scaled and added into the corresponding channel of the SAI. The time interval between the strobe and a given NAP value defines the position where the NAP value is entered in the SAI. If there are no succeeding strobes, the process continues for 35ms and then terminates. If more strobes appear in the 35 ms segment, which usually happens in music and speech, then each strobe initiates a temporal integration process, but the weights on the integrations are constantly adjusted so that the level of the SAI is normalized to the NAP level.

More specifically, if a new strobe appears, the weights of the older integration processes are reduced in order to contribute relatively to the SAI. The weight of the n^{th} process is 1/n, where n is the strobe number in the segment and 1 corresponds to the most recent strobe. The weights of the strobes are normalized so that the sum of the weights equals 1 constantly. The purpose of that is to keep the overall level and the spectral profile of the SAI more closely to that of the NAP. The final result of this process is the SAI, which is the decaying sum of all these events in 35 ms. Generally, when the sound enters the ear, the image is built up and and as the sound goes off, the image fades away as well.

Figures 3.4 - 3.7 show an example of the analysis performed by the AIM for the vowel /ae/ (Bleeck et al., 2004). The vertical ridge that corresponds to the pulse rate of the signal is visible and, as the pulse rate changes, it shifts on the time – interval axis. The formants appear as “waves” running horizontally from the vertical pitch ridge. As the VTL changes from long to short, the formants shift up in the frequency dimension and get narrower in the time-interval dimension. In the case of the SID system that has been designed in this study, the main interest is the SAIs (an example is shown in figure 3.7) that are generated from the AIM once a speech signal enters the system. Then, these auditory images are used for extracting the features that are necessary for identifying that specific speaker.

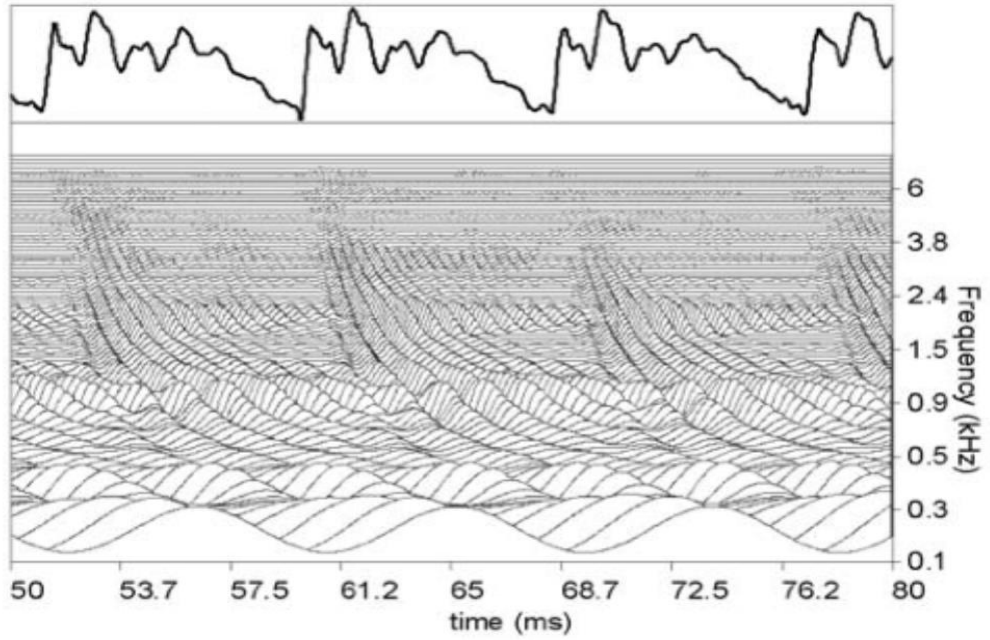


Figure 3.4: Basilar membrane response to the vowel /ae/. The upper panel is the original signal in the time domain. The horizontal axis in the panel is time. The vertical axis is cochlear channel, from low frequency at the bottom to high frequency at the top. Each line represents the traveling wave on the basilar membrane that corresponds to each channel of the filterbank (Bleeck et al., 2004)

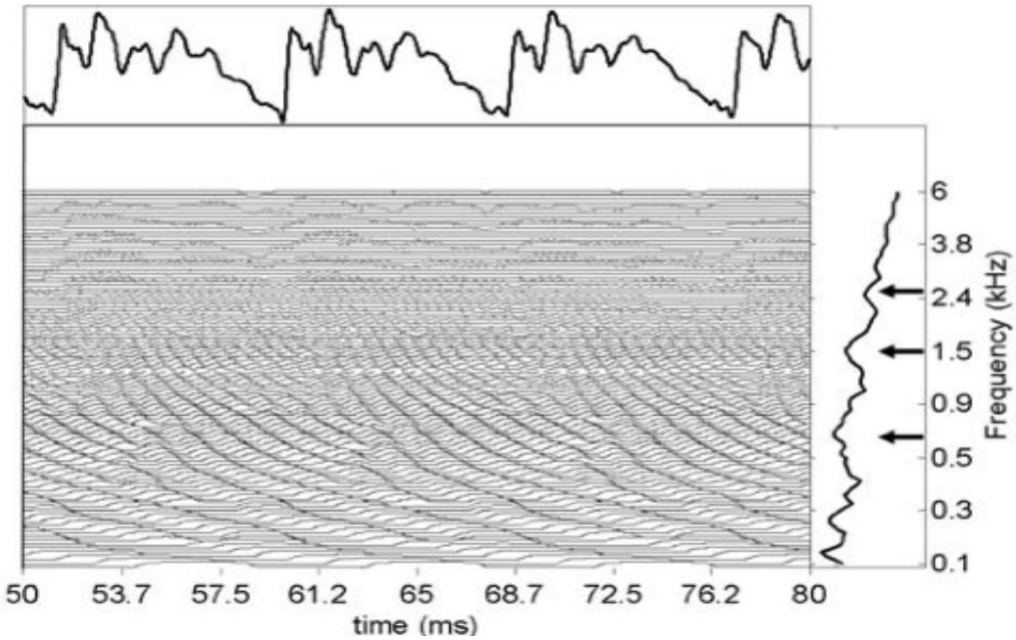


Figure 3.5: Neural activity pattern generated from the basilar membrane motion shown in figure 3.4 for the vowel /ae/. The output of the gammatone filterbank is half-wave rectified, compressed and low-pass filtered (Bleeck et al., 2004)

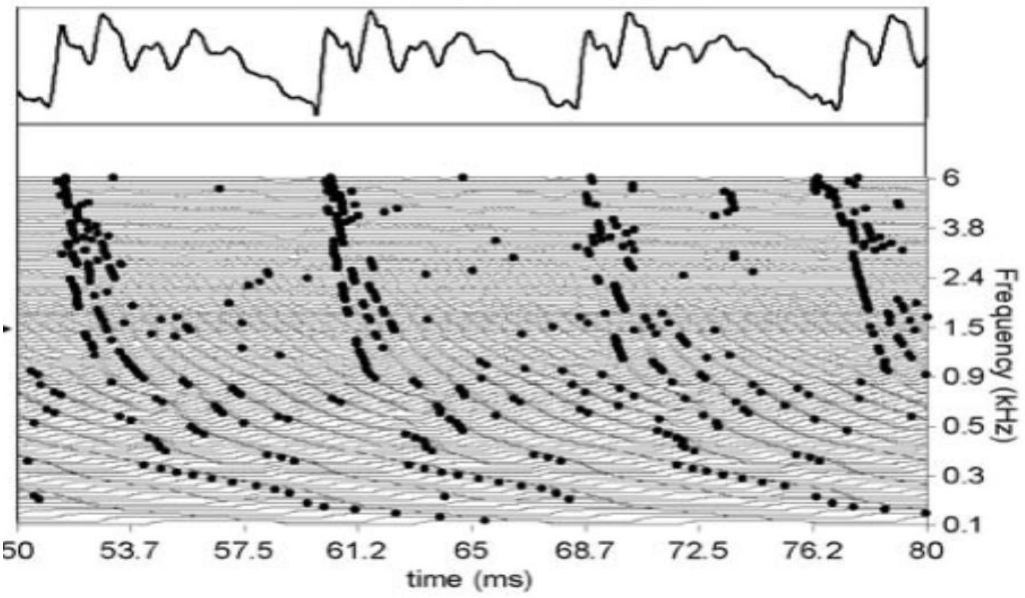


Figure 3.6: The strobe points as they are identified on the NAP of the vowel /ae/ in figure 3.5. Each black dot is a strobe that occurs when the NAP rises above the threshold (Bleeck et al., 2004).

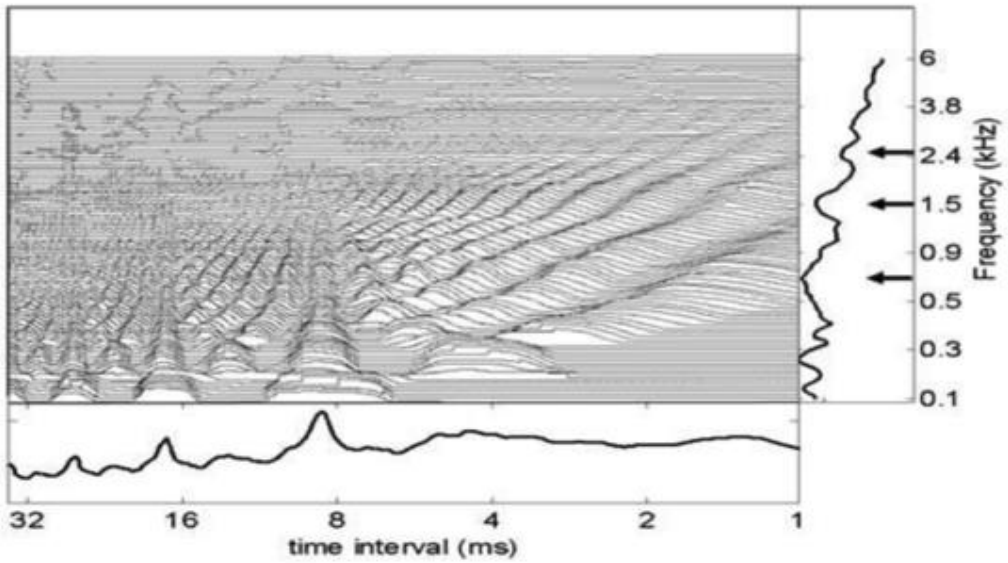


Figure 3.7: The SAI of the vowel /ae/ using the 'ti2003' module. The bottom panel is the temporal profile, which is estimated as the average over all channels for every point in time. The righ panel is the spectral profile, which is the average over time for every channel. The arrows show the locations of formants (Bleeck et al., 2004)

3.2 Feature extraction

In this section, the methods that have been used for the complete design of the new front-end, named box-cutting and downsampling, are described. In speaker recognition systems, it is important to use effective algorithms for representing sounds. These representations should capture the features that humans use to distinguish between different speakers. In order to achieve that, the box-cutting method is used to extract the features from the SAI through cutting it into a number of boxes. Additionally, an issue that affects the efficiency of such systems is the large dimensionality of the extracted features. For that reason, the downsampling process is applied on the features contained inside all of the boxes and it helps to achieve computational savings. Both procedures have been created by Lyon et al. (2010)) and I worked on the MATLAB implementation of them.

3.2.1 Box-cutting process

Firstly, the auditory image can be used in the feature extraction stage in order to obtain feature vectors that represent it. These vectors can be processed for identifying spectro-temporal patterns that appear in a SAI. The resulting representation of an image may appear as a histogram of those patterns. In general, the patterns can be identified at different positions in the auditory image. The specific location of a pattern depends on the characteristics of the sound source.

For speaker recognition, the information can be identified in smaller and larger scales in the SAI. At large scales, the temporal structure of the sound can provide information about the pitch whilst at smaller scales, there is information about the resonances following each pulse. The latter can be an indication for the vocal tract length (VTL).

Therefore, it is preferred to look for patterns in various locations and different scales of the SAI rather than looking only in the whole image. The process is based on defining a set of overlapping rectangles of different scales that cover the SAI frame.

Firstly, the initial rectangle size has been chosen to be 16 samples in the time - interval dimension by 32 filterbank channels. From this baseline pair of box sizes, both dimensions are multiplied in order to increase by powers of 2. The multiplication finishes at the point where the dimensions of the boxes do not exceed the dimensions of the SAI frame. For every pair of box dimensions, the SAI space is tiled with boxes, starting at the 0 ms point in the time-interval dimension. In the cochlear channel dimension, the box tiling occurs with a shift of half box height each time, i.e. 50% overlap. Figure 3.8 shows the process of tiling the SAI space with boxes and figure 3.9 shows an example of doubling both dimensions of the rectangles.

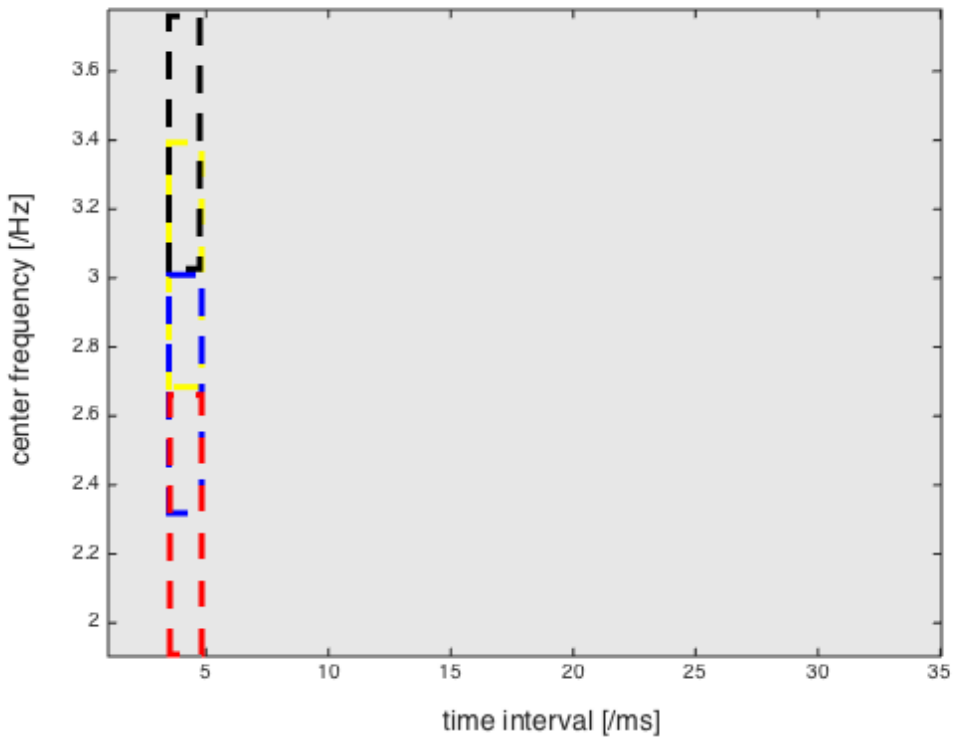


Figure 3.8: The concept of cutting the SAI frame in boxes with overlap equal to half box height

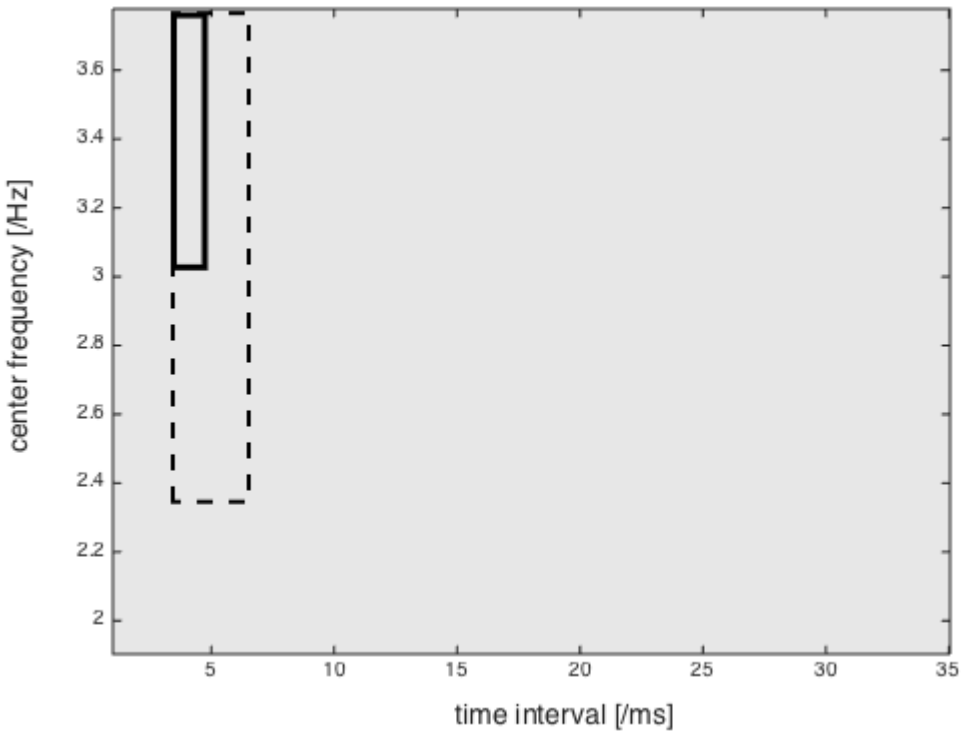


Figure 3.9: The doubling of the dimensions of the baseline box

The concept behind different box shapes and sizes is to capture different types of information. If a box is short and wide, it limits the temporal patterns into a small frequency region but it can capture the local spectral pattern. For tall and narrow boxes, the overall spectral pattern is captured at different temporal resolutions. Boxes with intermediate size may capture a combination of those features. This may be helpful in the case of presence of different sounds that correspond to different regions of the SAI. This procedure may also be used in a way that can provide some insight about regions of the image and it will be described in the next chapter.

3.2.2 Downsampling process

After completing the box-cutting process, the content of each rectangle is independently processed with a concept called downsampling. The aim of this procedure is to reduce the dimensionality and achieve computational efficiency of the system.

In order to achieve that, the margins of each box are computed by averaging the elements over each of the two dimensions. The result is two one – dimensional vectors, with dimensions equal to the dimensions of the smallest box, i.e. 32 and 16 elements. Then, the final feature vector is formed by concatenating them and consists of 48 elements. As a result, the image inside every box is downsampled to a one-dimensional feature vector. The outcome of this stage is the creation of a set of features that are in a more compact form because of the rescaling of the bigger boxes into the size of the baseline ones. Figure 3.10 shows the effect of downsampling on the auditory image and figure 3.11 shows an analytical diagram of the steps of the feature extraction using the combination of box-cutting and downsampling.

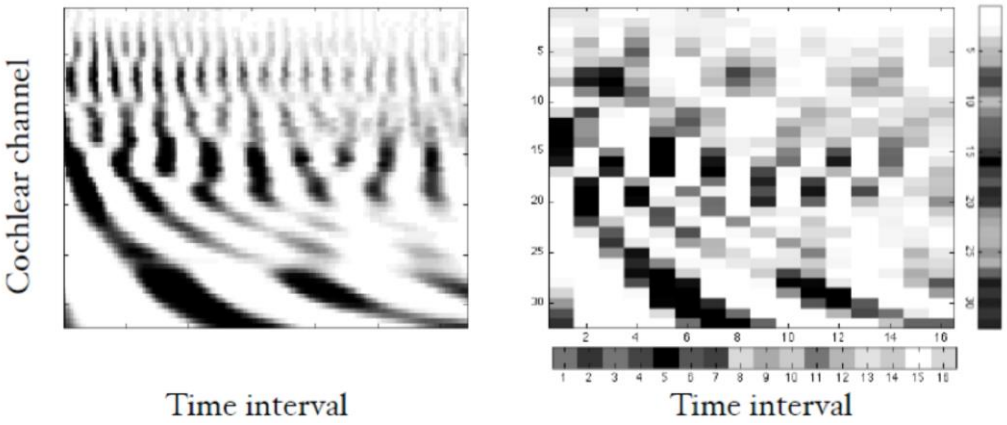


Figure 3.10: SAI frame with full resolution (left) and after downsampling (right) to 32 x 16 pixels (coarser resolution) (Walters, 2011)

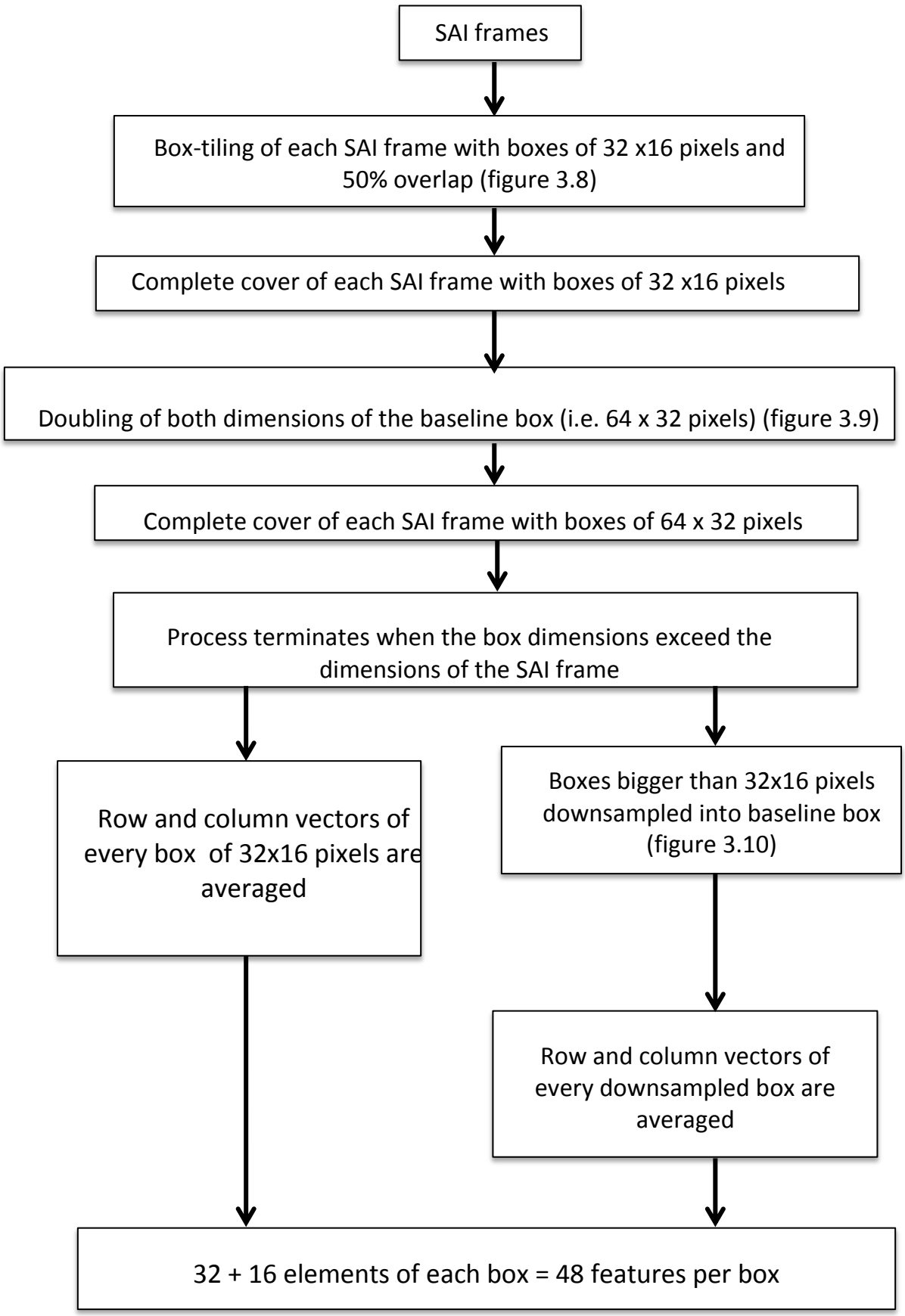


Figure 3.11: Diagram with the multiple steps of the process that consists of both box-cutting and downsampling

In conclusion, the combination of these two methods (box-cutting and downsampling) with the auditory image shape the new feature extractor that is proposed, in this research work, for the speaker identification task. The complete architecture of the system and the results from its application in speaker recognition for different conditions will be described in the two following chapters.

Chapter 4

Speaker Identification in Quiet Conditions

4.1 Introduction

In chapter 3 of this thesis, the parts that make up the feature extractor of the proposed speaker identification system have been analytically explained. The front-end is composed by the AIM, which generates the SAIs, and the box-cutting and downsampling methods that extract the auditory features. The output of the feature extraction stage is the speaker features, which are equal to the total number of the extracted boxes after being downsampled.

In this chapter, the feature representation from the SAI is hypothesized to produce high accuracy levels for speaker identification in quiet conditions. Also, the features will be compared to the MFCC features, which are the state-of-the art method for speaker recognition, and we hypothesize that they will produce similar or better results. In addition, we assume that there may be an optimal filterbank size that may produce better results for this specific task. The motivation behind it is that the filterbank size is a parameter of the model that affects the patterns created in the SAI as well as the outcome of the box-cutting process. The latter relates to the feature dimensionality since the feature vectors are equal to the number of extracted boxes from the image. As a result, it is expected that there is a particular number of filters that may produce the best possible SID accuracy.

In order to test these hypotheses, the proposed SID system that consists of the AIM with the box-cutting and downsampling processes, as described by Lyon et al. (2010), is used. The system performance is estimated for different filterbanks and speaker databases in order to find the optimal case for this task.

Furthermore, we hypothesize that there may be a subset of the extracted SAI features that contains the information that is more speaker-specific. These features are expected to have lower dimensionality, which results in computational efficiency of the system, and provide knowledge about what makes a speaker more discriminable. The reason for making this assumption is that there is data redundancy in the SAI since there is a very large increase in the data rate as the input signal is processed through the model. More specifically, in an AIM simulation, a single waveform in the time-domain is divided into a number of frequency channels. Each one of these channels has the same data rate as the original waveform and the SAIs that are generated have a much higher data rate. As a result, the challenge is to reduce the data rate while retaining as much of the important information as possible. In order to achieve that, the VQ process, combined with the box-cutting module, will be used in such a way that it is possible to extract information about which features are speaker-dependent. This knowledge is of importance for developing the box-cutting and improving the existing system design which will be explained in chapter 5.

The rest of this chapter is organized as follows. Section 4.2 describes the modules of the system as they have been implemented for both the AIM-based and the MFCC-based systems. Section 4.3 contains the results of the suggested methods for different speaker databases. Finally, the results of this experimental set are discussed in section 4.4.

4.2 System Overview

In this section, the modules of the proposed system are described analytically. Furthermore, the arrangement of the baseline system that uses MFCCs as a front-end is also explained. This understanding will help in the further development of the AIM-based system so that the improved system can be used in the next experimental set.

4.2.1 AIM-based system

The proposed system uses AIM, box-cutting and downsampling as the front end. The architecture of the system is presented in figure 4.1. The gammatone filterbank was used as the cochlea model for the auditory processing. In this experimental set, the filterbank consisted of 32, 64 and 96 bands, spanning a frequency range from 80 Hz to 6 KHz. The NAP was calculated using half-wave rectification of the filterbank output. The strobe detection is performed through the 'sf2003' algorithm as described in chapter 3.

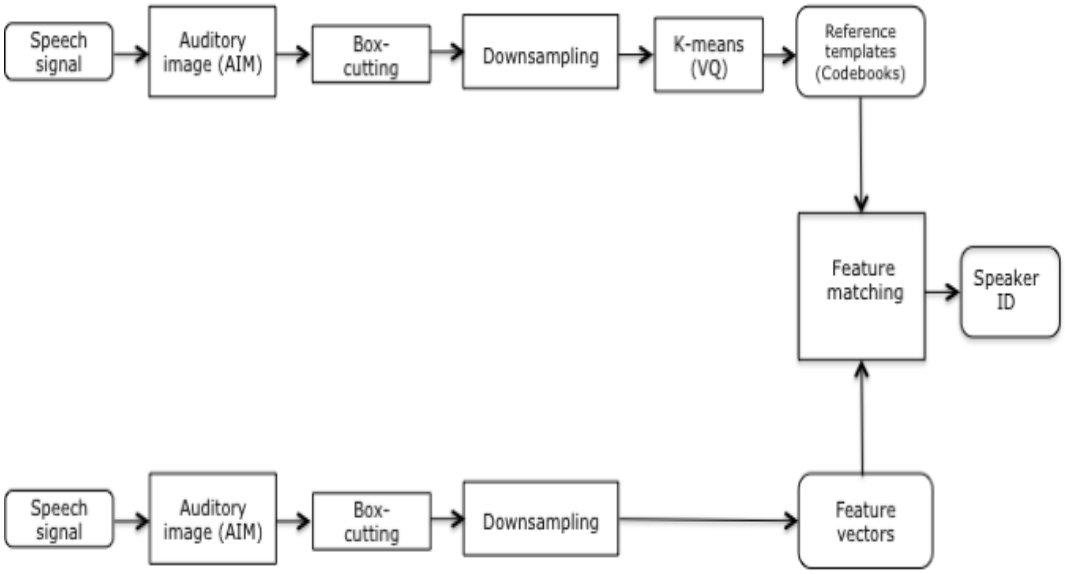


Figure 4.1: Schematic design of the SID system using AIM as a front end

The AIM uses the ‘ti2003’ for SAI generation: when a strobe occurs, the signal following the strobe is added into the buffer, starting from the zero time-interval point. This process continues for 35 ms after the strobe has occurred which leads to an AIM-SAI width of 35 ms. As time passes, more strobes occur which are also added in the buffer. If multiple strobes are active, i.e. more than one strobe is present in a 35ms window, the signal following each strobe is weighted by a quantity inversely proportional to the number of active strobes, before being added to the buffer. This happens so that the overall level of the SAI remains equal to the level of the cochlea model output. The output of AIM is the 35ms SAI and that is the input for the box-cutting step.

As described before, the purpose of box-cutting is to identify patterns that usually appear in a SAI. In this system, the SAI is tiled with boxes of dimensions that are powers of 2. The initial box dimensions are 32 filterbank channels by 16 samples in the time-interval dimension. For each pair of box dimensions, the SAI frame is covered with boxes that start at the zero point in the time-interval dimension and shift in the frequency channel dimension by half box height each time. From the baseline box size, both dimensions are doubled up to the largest possible box that does not exceed the dimensions of the SAI frame.

Given these parameters, the SAI that uses 32, 64 and 96 frequency bands is cut into a total of 44, 154 and 264 rectangles respectively. As described before, the image in each box is downsampled to the size of the smallest box. As a result, all of the rectangles are reduced to 48 values each with the downsampling process. The values of all the experimental parameters used are shown in Table 4.1. The size of the filterbank is the parameter that was varied in this group of experiments and affects the parameters invoved in the box-cutting module, i.e. number of extracted boxes.

Parameter Set of AIM-SAI			
Time-interval (STI)	35ms		
Filterbank size	32	64	96
Total number of boxes	44	154	264
Smallest box dimensions	32 x 16		
Means per box (K)	64		

Table 4.1: Parameters used for the experiments with the SAI

After the completion of the process of transforming the SAI frames into feature vectors, the speaker modeling step follows. Theoretically, it could be possible to use all of the feature vectors as the reference template for every speaker. However, given the dimensionality of the data, it is important to find a data representation with reduced dimensions. This is achieved through Vector Quantization, which is implemented using the K-means algorithm, as it has been described in chapter 2.

The choice of VQ has been made on the basis that it is a non-parametric modelling approach. Therefore, minimal assumptions are made about the underlying distribution of the features. In general, a specific modelling technique may be better for a specific set of features but not good enough for another set. In the case of this system, it is assumed that the results could be generalized for other modelling techniques, such as the GMM.

Given that the purpose of box-cutting and downsampling is based on identifying patterns in the auditory image and representing each SAI frame, the same concept is followed in the classification procedure. In practice, this means that the clustering algorithm has to take into account the content of each box. During the training session of a particular speaker model, a codebook is learnt for each one of the multiple boxes over the total number of the extracted SAI frames. As a result, every box

will have its own codebook created and each speaker template will consist of a number of codebooks equal to the number of boxes.

Our hypotheses were tested on two speech corpora, a small set of 30 speakers and a large of 180 talkers, which will be described analytically in the next section. The initial step of the process is the selection of all the 48 – element feature vectors that represent each box over all of the training frames. The average duration of the 30-speaker speech corpus used for training is 14.2 sec and the window length is 10 ms. Thus, the average number of frames is 1420. This results in 1420 x 44 boxes = 62480 feature vectors when the AIM uses a filterbank of 32 frequency channels, 1420 x 154 boxes = 218680 for a 64-channel filterbank and 1420 x 264 boxes = 374880 for a 96-channel filterbank. Considering that the average training speech duration for the corpus of 180 speakers is 20.7 sec, the average total number of frames is 2070. Before the vector quantization, the feature dimensions are 91080 (2070 x 44 boxes) for the 32-center frequency SAI, 318780 (2070 x 154 boxes) for the 64-center frequency SAI and 546480 (2070 x 264 boxes) for the 96-center frequency SAI.

After concatenating all of these feature vectors for the total number of frames, we have the representation of the entire speech signal. Then, the K-means clustering algorithm is used for vector quantization. The number of centroids is chosen to be 64 and that will result in a codebook of size 64 for each one of the rectangles. In general, we expect that using larger codebooks improves recognition performance. For the minimum codebook sizes, a number between 16 and 64 is acceptable, depending on the feature set, the dimensionality and the amount of training data (Kinnunen et al., 2004). In our experiments, the choice of 64 codewords seems to be reasonable as it is the maximum of the range of possible minimum values and reduces dimensionality to a significant degree. This parameter will be kept constant throughout all of the experiments so that we can focus on the performance of the extracted auditory features.

The outcome of the VQ process is a representation of the speakers with a significantly reduced dimensionality. For the 32-center frequency SAI, the total number of feature dimensions is equal to 64 x 44 boxes = 2816 while for the SAI with 64 and 96 frequency bands, the results are 9856 (64 x 154 boxes) and 16896 (64 x 264 boxes) feature vectors respectively.

In conclusion, the final result of the enrolment session is a number of speaker models equal to the number of trained speakers. Each one of these templates consists of a number of codebooks equal to the total number of boxes for every case that has been mentioned above. The size of each one of these codebooks is equal to 64 x 48 elements. Figure 4.2 shows the steps for creating the reference template of one speaker from the feature vectors that are extracted from one's speech. The process is repeated as many times as the number of trained speakers.

Subsequently, in the testing phase, the feature extraction is repeated in the same way as in the enrolment session. The features are computed for every box as 48 - element feature vectors. At this stage, the main concept is to see how well each codebook encodes the features for the test speaker. To achieve that, the values that can reconstruct each frame of the test speaker using each one of the trained speaker models have to be estimated.

Firstly, for every frame and every box, the Euclidean distance is computed between every centroid in the codebook for that specific box, and the current feature vector for that box. For each one of the frames, the minimum of these distances is the reconstruction value for that frame using the codebook for that specific box.

Afterwards, the process is repeated over the total number of frames and those values, for each box, can be averaged over all of the frames, i.e. for the whole speech utterance. This results in the mean reconstruction value for every box. The process can be repeated for each one of the trained speakers.

The speaker that is most likely to be the target speaker is the one who has the largest number of boxes corresponding to the smallest average reconstruction value. The procedure that has been described above leads to the speaker matching step of the recognition system.

Having completed that, it is possible to assess the system in terms of identification accuracy. The procedure for the stage of speaker matching is explained analytically in figure 4.3.

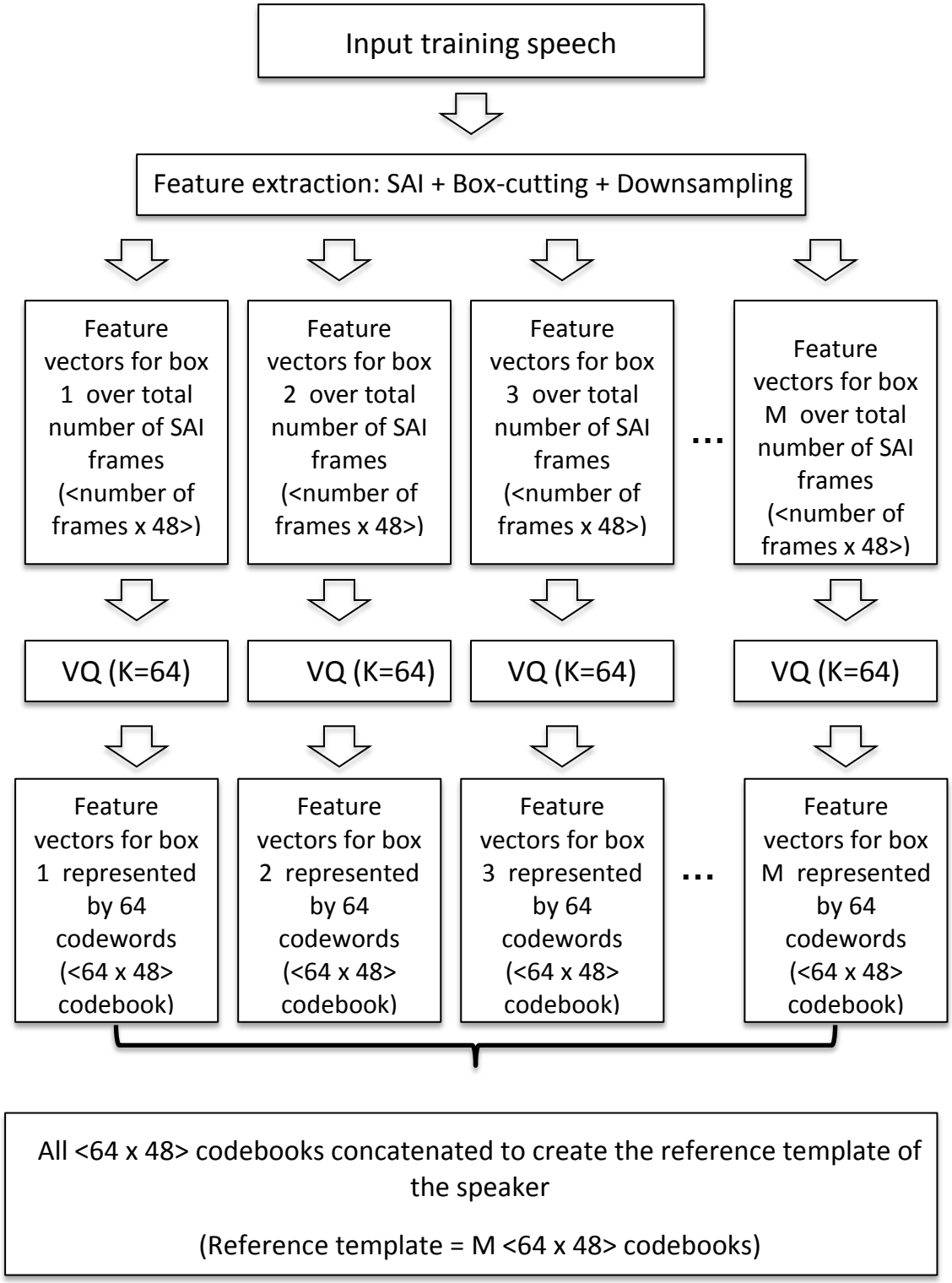
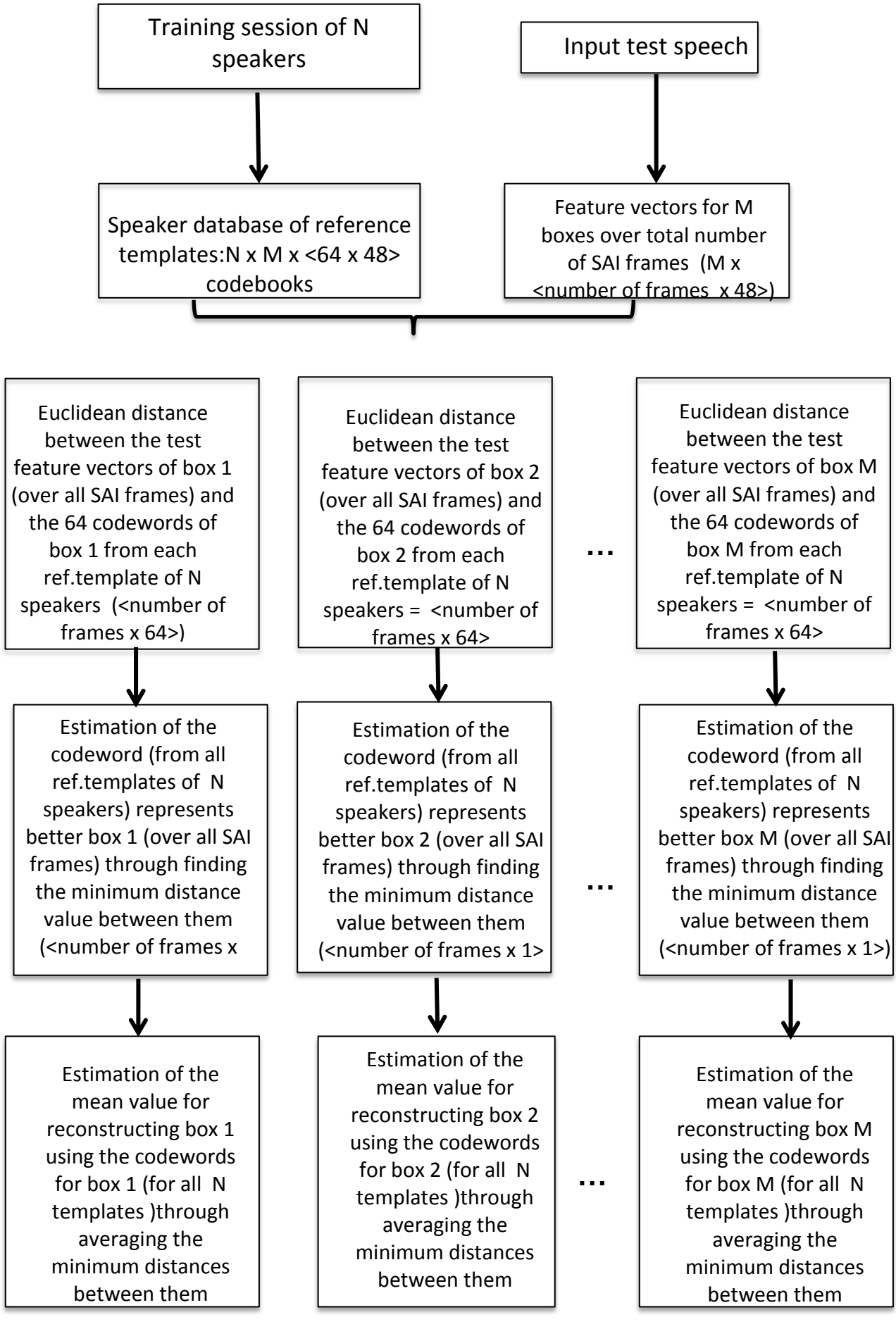


Figure 4.2: The steps of the enrolment (training) session for creating the template of 1 speaker that contains one codebook for every extracted box (total number of boxes = M) from the speaker's speech signal. The same process is repeated for the total number of speakers in the database.


Training session of N speakers
Input test speech
Speaker database of reference templates:N x M x <64 x 48> codebooks
Feature vectors for M boxes over total number of SAI frames (M x <number of frames x 48>)
Euclidean distance between the test feature vectors of box 1 (over all SAI frames) and the 64 codewords of box 1 from each ref.template of N speakers (<number of frames x 64>)
Euclidean distance between the test feature vectors of box 2 (over all SAI frames) and the 64 codewords of box 2 from each ref.template of N speakers = <number of frames x 64>
...
Euclidean distance between the test feature vectors of box M (over all SAI frames) and the 64 codewords of box M from each ref.template of N speakers = <number of frames x 64>
Estimation of the codeword (from all ref.templates of N speakers) represents better box 1 (over all SAI frames) through finding the minimum distance value between them (<number of frames x
Estimation of the codeword (from all ref.templates of N speakers) represents better box 2 (over all SAI frames) through finding the minimum distance value between them (<number of frames x 1>
...
Estimation of the codeword (from all ref.templates of N speakers) represents better box M (over all SAI frames) through finding the minimum distance value between them (<number of frames x 1>)
Estimation of the mean value for reconstructing box 1 using the codewords for box 1 (for all N templates)through averaging the minimum distances between them
Estimation of the mean value for reconstructing box 2 using the codewords for box 2 (for all N templates)through averaging the minimum distances between them
...
Estimation of the mean value for reconstructing box M using the codewords for box M (for all N templates)through averaging the minimum distances between them

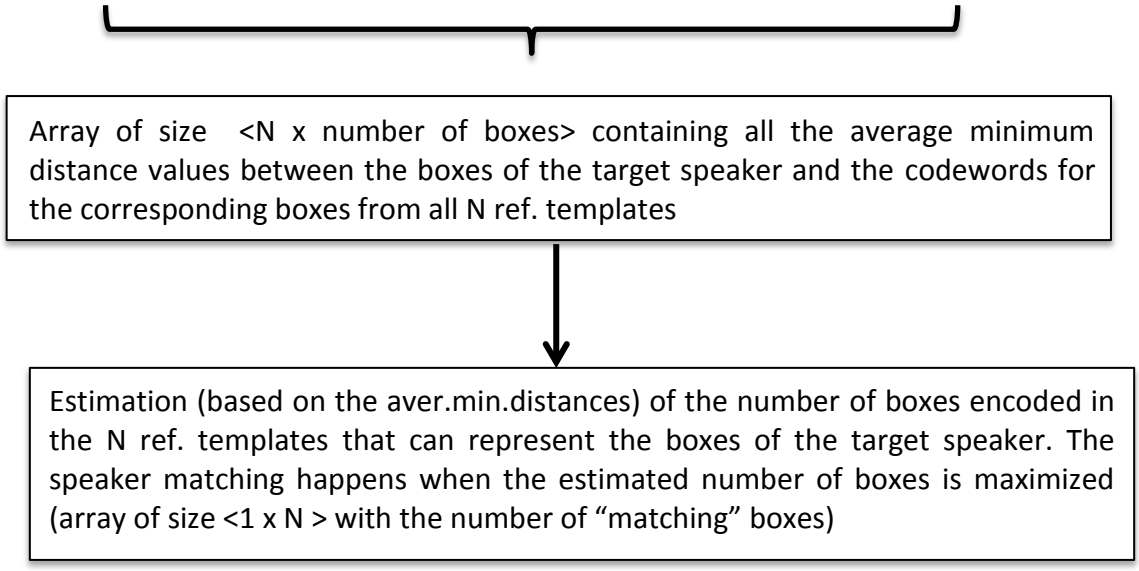


Figure 4.3: The steps of the speaker matching stage for 1 test (target) speaker and N trained (enrolled) speakers (i.e. N reference templates with M codebooks each). The same process is repeated for the total number of test speakers.

For the succeeding step of this study, the process for testing our hypothesis about finding a lower-dimensional feature representation from the SAI is explained. Based on the assumption that differences among speakers can rely on low-level features, which means that they focus on specific individual components of a system operation, we use the very high-dimensional SAI feature representation (acquired by the combination of the methods included in the feature extractor and the VQ) and we let the learning algorithm specify the most discriminative features among speakers. The advantage is that we gain insight on traits of the SAI that are more speaker-specific and it becomes feasible to improve the proposed feature extractor.

As mentioned previously, the matching between speakers happens when the majority of the boxes, which are extracted by the SAIs of the target speaker, fulfill the criterion of the minimization of the average reconstruction value from the codebooks for those boxes of a specific speaker in the speaker database.

Alternatively, between speakers that do not match with each other, those average reconstruction values should be maximized. Through estimating those maximum values, it is possible to specify the boxes that correspond to them. These boxes are the most discriminative among all of the speakers. Furthermore, the position of these boxes on the SAI can indicate the areas that are most informative for identifying a person. This knowledge is important for modifying the box-cutting process and selecting those boxes instead of all of them as it was initially done. The results of this process for various parameters of the SAI as well as the speech corpora that are used will be presented in section 4.3.

4.2.2 MFCC-based system

Mel-cepstrum is probably the most commonly used feature in speech recognition and has become the state-of-the-art method for parametrization in speaker recognition systems as well. However, this might not be the best representation of speech sounds especially under challenging conditions such as noisy environments (which will be investigated in the experiments presented in the next chapter).

In this research study, the system that uses MFCCs as a front end will be used as a comparison to the proposed system. For this system, the speaker modelling task is completed through the same type of classification that was used for the AIM-based system. The main difference is that there is not a box-cutting process involved so the VQ procedure is not repeated for multiple boxes as before. The architecture of the MFCC-based system is presented in figure 4.4.

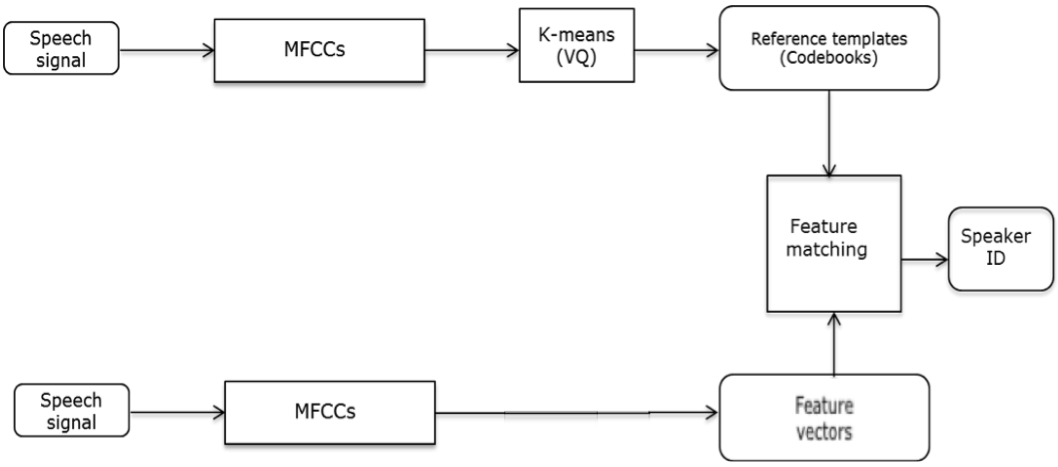


Figure 4.4: Schematic design of the SID system using MFCCs as a front end

Analytically, during the training stage, a feature vector consisting of 40 cepstral coefficients is extracted for each frame (which has a duration of 25ms). The complete feature representation for the whole speech utterance will be of size equal to (number of frames) x 40. Afterwards, during speaker modeling, a codebook is learnt for every speaker over the total number of frames. As before, the number of centroids is 64 and the size of the codebook for that speaker has 64 x 40 elements and it is learnt over the whole utterance. After completing the enrolment of all of the speakers, the final estimate is a number of speaker reference templates equal to the number of speakers. Each model consists of one codebook. Figure 4.5 shows the steps for creating the reference template of one speaker from the MFCC features that are extracted from one's speech. The process is repeated as many times as the number of trained speakers.

Then, in the testing phase, the concept is to see how well each codebook for every speaker encodes the features of the test speaker. So, for every frame, the 40 - dimensional feature vector is computed and compared to each one of the 64 codewords. The comparison is made by computing the Euclidean distance between the feature vector and each one of the centroids. This results in a matrix of distances to each centroid.

After that, the minimum distance between one of the centroids and the feature vector for each frame is computed. This specific centroid is considered to be the most representative centroid and the reconstruction value for that frame using that codebook is that distance value. Afterwards, the same procedure is repeated for all of the other frames and the outcome is a list of all of these distances. Then, the next step is to estimate the mean value of these distances, which is the average reconstruction value for the whole speech utterance.

Finally, the decision about the most likely target speaker is based on repeating this process for each one of the speaker models and estimate the mean reconstruction values for all of them. The speaker model with the smallest value is the most probable to match the unknown speaker. The steps of the procedure for speaker matching is explained analytically in figure 4.6. After using both of these systems for the SID task, the results are presented and explained in the next section.

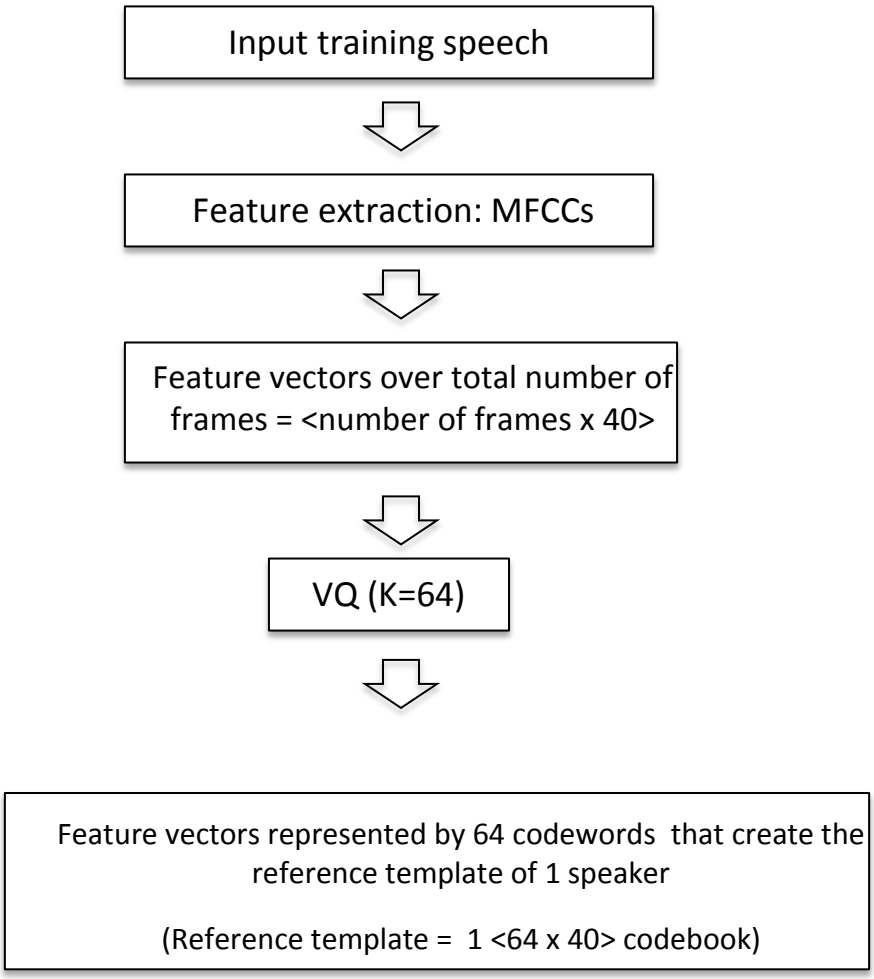


Figure 4.5: The steps of the enrolment (training) session for creating the template of 1 speaker that contains only 1 codebook. The same process is repeated for the total number of speakers in the database.

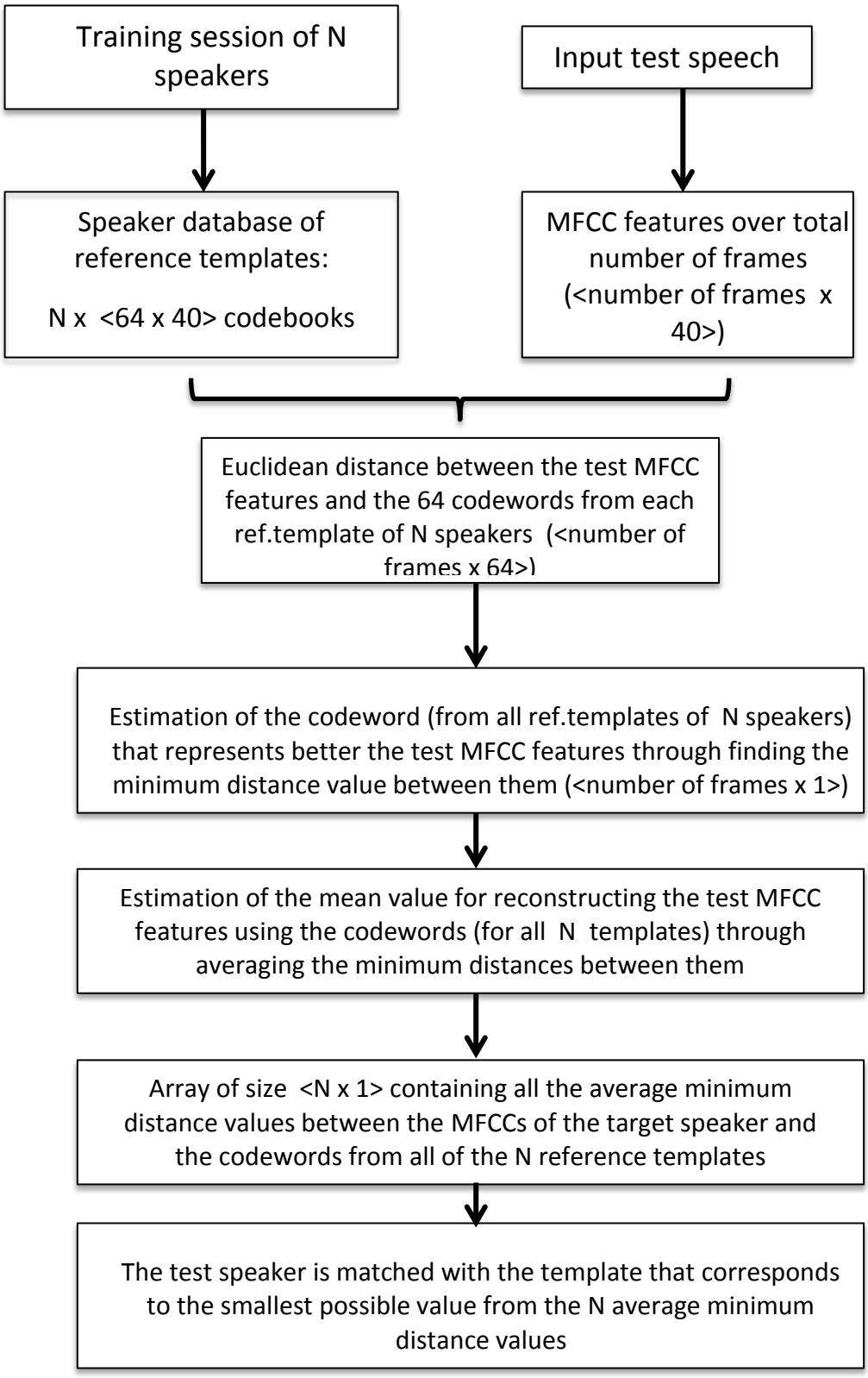


Figure 4.6: The steps of the speaker matching stage for 1 test (target) speaker and N trained (enrolled) speakers (i.e. N reference templates and N codebooks). The same process is repeated for the total number of test speakers.

4.3 Evaluation

4.3.1 The Data Sets

The speech corpus that was used is a multilingual corpus named EUROM1 (Chan et al., 1995). This database consists of recordings in 7 European languages (Danish, Dutch, British English, French, German, Norwegian, Swedish).

For the experiments of this thesis, specific parts of this database were used for creating the different speech corpora. The EUROM1 corpus was collected by the "Enabling Technology and Research" working group within ESPRIT-project 2589 "Speech Assessment Methodology (SAM)". One of the major tasks of this working group has been the production of standard European speech databases. The corpora for each one of the languages was validated by institutions at the associating European countries.

The EUROM1 corpus contains approximately 60 speakers from each European country. The number of males and females varies for each language. The total number of speakers was used to create three different sets of speakers consisting of many talkers, few talkers and very few talkers. All speakers were prompted to read the same material. The recordings were done in an anechoic room. The data was digitized using a sampling frequency of 20 KHz and a quantization resolution of 16 bits per sample.

For this experimental set, two speaker databases were created consisting of 30 and 180 talkers. Each one of them consists of 3 groups of 10 and 60 speakers respectively. For all of them, the speech material is in the form of small passages of 5 sentences. The speech signals have been pre-processed for pause removal. The characteristics of the audio files used for the training and testing stages are summarized in the following tables for both corpora.

Corpus of 30 speakers	
Language	English
Speakers	30 (12 F+ 18 M)
Speech type	Read speech
Sampling frequency	20 KHz
Training speech type	Clean
Training speech duration (avg)	14.2 sec
Test speech type	Clean
Test speech duration(avg)	15 sec

Table 4.2: Summary of the speech corpus consisting of 30 speakers

Corpus of 180 speakers	
Language	English, French, Swedish
Speakers	180 (90 F+ 90 M)
Speech type	Read speech
Sampling frequency	20 KHz
Training speech type	Clean
Training speech duration (avg)	20.7 sec
Test speech type	Clean
Test speech duration(avg)	20.6 sec

Table 4.3: Summary of the speech corpus consisting of 180 speakers

4.3.2 Performance on the EUROM1 corpora

In this section, we evaluate the SID system that consists of the AIM and the box-cutting method as it was described by Lyon et al. (2010). At first, our hypothesis was that the proposed system will give high accuracy levels. Secondly, we hypothesized that the results may be similar or better to the ones obtained by the MFCCs. Furthermore, we assumed that, based on the identification results, it may be possible to characterize the best possible filterbank size for the task at hand. In this experimental set, the system was tested for 32, 64 and 96 frequency channels. The system performance was assessed with the SID accuracy, which is equal to the ratio of the number of correctly identified speakers to the total number of speakers that have been considered for the testing phase.

As previously described, in the speaker matching stage, the boxes that correspond to the extracted feature vectors from the test speaker are compared to the codebooks of these boxes in the speaker database using the criterion of minimization of the average reconstruction value between them. When the majority of the boxes of the test speaker fulfil this criterion in relation to the codebooks of a specific speaker template, then these two speakers are matched. In practice, confusion matrices among the speakers were created for every group of speakers that was used for this experimental set. For each one of the groups of the 30-speaker corpus, square confusion matrices of order 10 were made. Similarly, for the corpus of 180 speakers, the resultant confusion matrices are 60-by-60.

Tables 4.4, 4.5 and 4.6 present the results for one of the groups of 10 speakers from the 30-speaker speech corpus for 32, 64 and 96 frequency channels. The same procedure has been followed for all of the groups of both speech corpora. In the confusion tables, the estimation of the number of boxes that are matched among speakers is presented. In the case of correct identification, the diagonal of every confusion matrix is expected to contain the largest possible number of boxes in comparison

with the 44, 154 and 264 boxes that are extracted from the SAI that consists of a bank of 32, 64 and 96 filters respectively.

Actual Speaker Index	Hypothesized Speaker Index									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	44	0	0	0	0	0	0	0	0	0
2	0	39	0	0	0	0	0	0	0	5
3	0	0	40	0	0	4	0	0	0	0
4	0	0	0	44	0	0	0	0	0	0
5	0	0	0	0	44	0	0	0	0	0
6	0	0	0	0	0	44	0	0	0	0
7	0	0	0	0	0	0	38	0	0	6
8	0	0	0	0	0	0	1	43	0	0
9	0	0	0	0	0	0	0	0	44	0
10	0	0	0	0	0	0	0	1	0	43

Table 4.4: Confusion matrix of the SID results for a group of 10 speakers (from the 30-speaker speech corpus) using the features from a 32- center frequency SAI. The total number of extracted boxes is 44. The pattern matching indicates that all of the 10 speakers are correctly identified since the diagonal contains the largest numbers compared to the 44 extracted boxes.

Actual Speaker Index	Hypothesized Speaker Index									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	138	1	0	1	7	1	0	0	1	5
2	0	105	1	0	24	7	0	2	0	15
3	8	0	98	4	0	26	0	7	1	10
4	23	2	1	114	11	3	0	0	0	0
5	4	11	0	0	134	0	0	0	0	5
6	0	3	5	0	0	142	1	0	0	3
7	0	11	6	0	31	6	75	4	0	21
8	20	0	0	1	0	0	0	133	0	0
9	6	5	4	0	17	4	0	10	106	2
10	2	1	18	0	1	7	5	0	0	120

Table 4.5: Confusion matrix of the SID results for a group of 10 speakers (from the 30-speaker speech corpus) using the features from a 64-center frequency SAI. The total number of extracted boxes is 154. The pattern matching indicates that all of the 10 speakers are correctly identified since the diagonal contains the largest numbers compared to the 154 extracted boxes. The presence of more off-diagonal terms is because of the K-means algorithm that creates the codebooks for every box. The variation in the way the feature vectors are grouped (clustered) each time results in differences between the matching of the codewords of the boxes in the reference templates and the test features for the corresponding boxes.

Actual Speaker Index	Hypothesized Speaker Index									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	259	0	0	0	0	0	0	0	5	0
2	0	264	0	0	0	0	0	0	0	0
3	0	6	227	0	0	1	8	20	0	2
4	0	0	0	264	0	0	0	0	0	0
5	0	2	0	0	261	0	0	0	0	1
6	0	0	0	0	0	263	0	1	0	0
7	0	0	0	0	0	0	255	5	0	4
8	0	0	0	0	0	0	0	264	0	0
9	0	0	0	0	2	0	0	0	261	1
10	0	0	0	0	0	0	0	0	0	264

Table 4.6: Confusion matrix of the SID results for a group of 10 speakers (from the 30-speaker speech corpus) using the features from a 96-center frequency SAI. The total number of extracted boxes is 264. The pattern matching indicates that all of the 10 speakers are correctly identified since the diagonal contains the largest numbers compared to the 264 extracted boxes.

After computing all the confusion matrices that contain the results for the matching patterns, the SID accuracy is estimated for every group. Then, the mean identification accuracy is estimated for the data sets of 30 and 180 speakers. The error of the identification score is calculated as the standard error of the mean.

In SID systems, there are different sources of error that are related to the speaker itself and/or to technical conditions. Firstly, the intra-speaker variation can affect a speaker's voice. This means that a person's health condition (e.g. flu) or emotional state (e.g. stress) may influence one's voice. Another case that has an impact on voice quality is smoking or the use of drugs (Kinnunen, 2004). Additionally, another example of intra-speaker variation is the fact that a speech wave cannot be produced exactly the same, when the same person speaks in different sessions, since it is the end result of movements of voice production body parts. Secondly, voice disguise is another type of error source where a person intentionally changes his/her voice so that it cannot be matched with a speech sample that has been previously produced by him/her. Impersonation or imitation is a specific type of voice disguise where the speaker tends to map one's voice in order to sound like another person. Voice disguise and imitation have been proven to degrade the performance of SID systems (Rose, 2002). Lastly, other error sources are associated with technical factors such as distortions of poor-quality microphones, reverberations or additive environmental noise. These error sources are the most usual ones and they cause the circumstances of the training and testing phases to be different. These mismatched conditions are regarded as the most challenging source of error (Reynolds, 2002) and it will be examined in chapter 5.

Tables 4.7 and 4.8 summarize the results for the SID accuracy and the identification error for both speech corpora and for different filterbank sizes. The test speech utterances were different from the training ones since the system performs text-independent recognition.

Corpus of 30 speakers				
System configuration	**SAI (32 cf)**	**SAI (64 cf)**	**SAI (96 cf)**	**MFCC**
SID Accuracy (%)	100	100	100	100

Table 4.7: SID accuracy (%) of the SAI-based system (for varied filterbank size) and MFCC-based system (corpus consisting of 3 groups of 10 speakers)

Corpus of 180 speakers				
System configuration	**SAI (32 cf)**	**SAI (64 cf)**	**SAI (96 cf)**	**MFCC**
SID Accuracy (%)	84.4	89.4	87.1	90.5
Error of SID (%)	2.4	2	2.94	3.1

Table 4.8: SID accuracy (%) of the SAI-based system (for varied filterbank size) and MFCC-based system (corpus consisting of 3 groups of 60 speakers). The error of SID is the standard error of the mean(estimated as the error among the levels of SID accuracy of the 3 subsets of 10 speakers)

From the results above, it is obvious that the system achieves high accuracy levels that are similar to the ones obtained from the MFCCs. Additionally, it is clear that the SID accuracy decreases when the size of the speaker population increases.

Furthermore, an interesting observation is that the number of cochlear channels can influence the system performance. This becomes more obvious for the larger set of 180 talkers, where the identification score obtained from the 64-center frequency SAI is better compared to the ones from the system configurations that use the SAI with 32 or 96 center frequencies.

Moreover, in table 4.8, the results for the SID accuracy between the 64 and 96 center frequencies and those between 32 and 96 center frequencies do not appear to be significantly different (overlapping standard errors). Nevertheless, the choice of 64 center frequencies is reasonable since it has produced the highest average SID accuracy. As a result, our assumption that there may be a specific number of filters that can produce better SID accuracy can be supported. This finding will be combined with the results about the informative SAI regions, which are presented in the following section, so that a decision is made for the further modification of the existing system.

4.3.3 Specification of informative SAI features

In this section, our hypothesis about finding a subset of the extracted SAI features that contains speaker-specific information and, at the same time, has lower dimensionality is investigated. As described in chapter 3, the SAI is like a ‘movie’ with multiple frames. Each frame is two dimensional, i.e. the cochlear channel dimension and the time interval. In the SAI, the changes of the glottal pulse rate (GPR) correspond to a change in the horizontal spacing of the vertical pitch ridges. At the same time, the change of the resonance scale (formants) associate to the changes in the vertical location of the resonance structure. Therefore, the SAI separates, to a certain extent, the two types of information into the two dimensions of the auditory image.

From the first part of this experimental set, it seems that the results for the high-dimensional SAI feature representation over multiple boxes are promising. In this part of the experiments, the VQ process that has been described in section 4.3 is used and the learning algorithm helps in identifying the features that are most distinctive. As explained before, this is achieved by specifying the boxes that fulfil the criterion of maximum average reconstruction value between the extracted boxes and the

codebooks for those boxes. These are the ones that are more speaker-specific and constitute particular traits about them.

The procedure was repeated 3 times for different groups of 10 and 60 speakers. In addition, the experiments have been conducted for 32, 64 and 96 center frequencies of the filterbank. The concept is to see the distribution of patterns, when there is variation in the parameters of the frequency dimension. Figures 4.7 – 4.12 show where these areas are located on the SAI for all of the speakers of the chosen corpora. In the following figures, the x-axis is the time-interval dimension and the y-axis is the frequency channel dimension of the SAI. The rectangles that are plotted are the ones that have been specified by the combination of the VQ and the box-cutting procedures and they fulfil the criterion that has been described before. Each rectangle has a size of 16 samples in the time - interval dimension by 32 filterbank channels and contains part of the patterns of the auditory image.

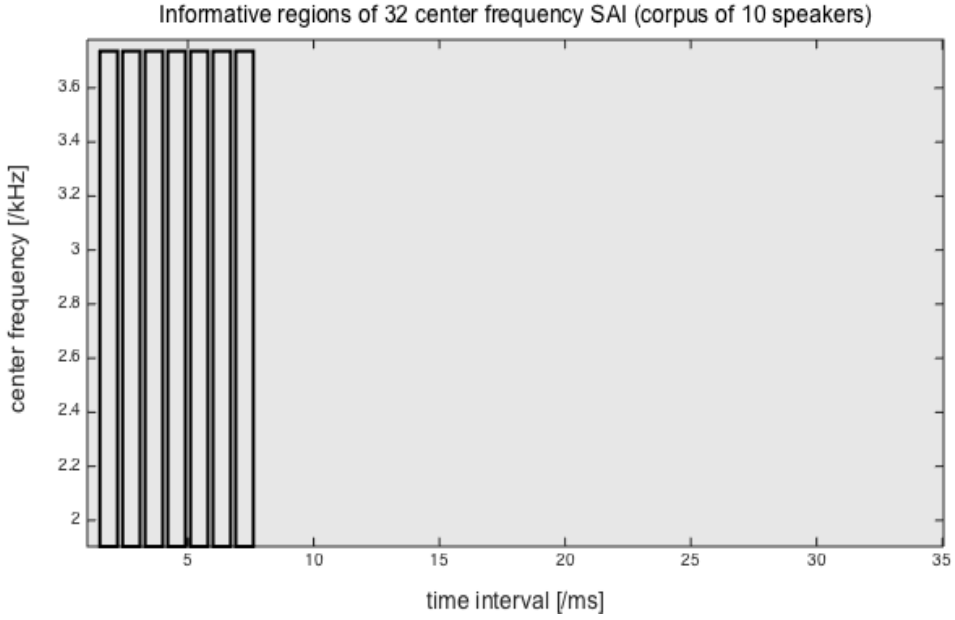


Figure 4.7: Specification of the informative regions of the 32-center frequency SAI (for all trials using 10 speakers). The 7 tall and narrow boxes cover the filterbank and the area between 1.6 and 7.2 ms.

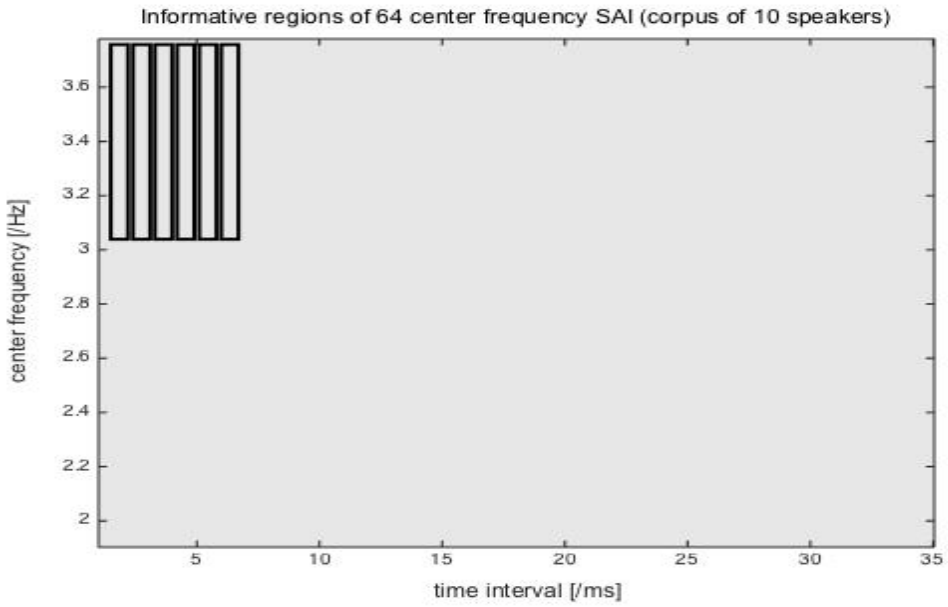


Figure 4.8: Specification of the informative regions of the 64-center frequency SAI (for all trials using 10 speakers). The 6 short and narrow boxes cover part of the filterbank (above 1KHz) and the area between 1.6 and 6.4 ms.

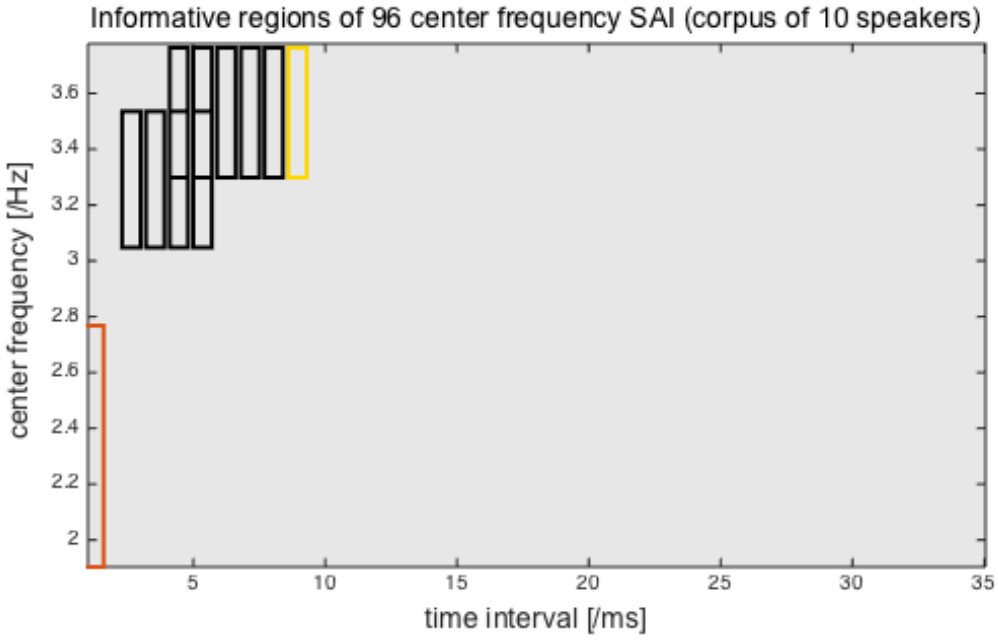


Figure 4.9: Specification of the informative regions of the 96-center frequency SAI (for all trials using 10 speakers). The 11 boxes cover various parts of the filterbank (low and high frequencies) and extend up to 8.8 ms. The black boxes appear to be informative in all trials while the colored boxes (yellow and orange) are discriminative as well for the first and second trial respectively.

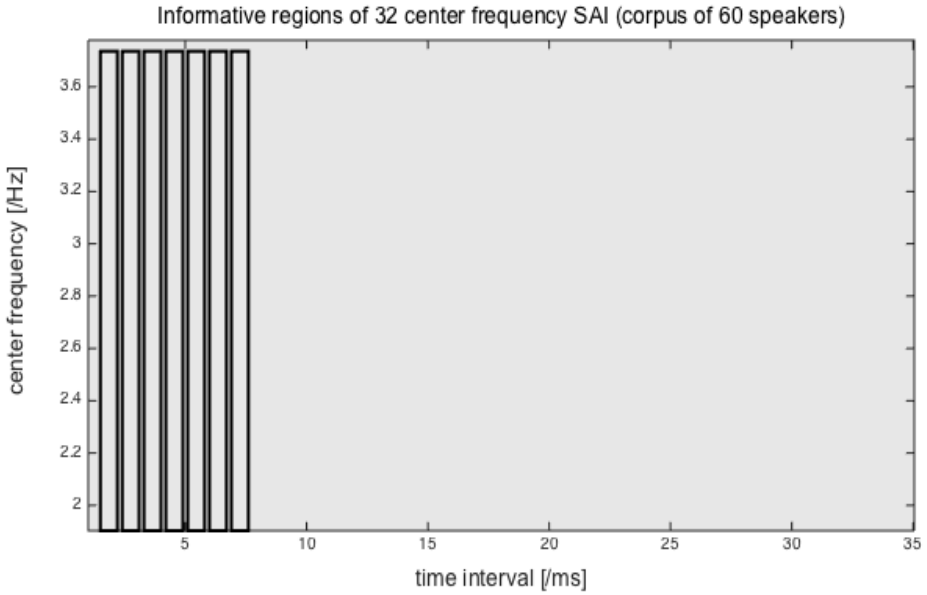


Figure 4.10: Specification of the informative regions of the 32-center frequency SAI (for all trials using 60 speakers). The 7 tall and narrow boxes cover the filterbank and the area between 1.6 and 7.2 ms.

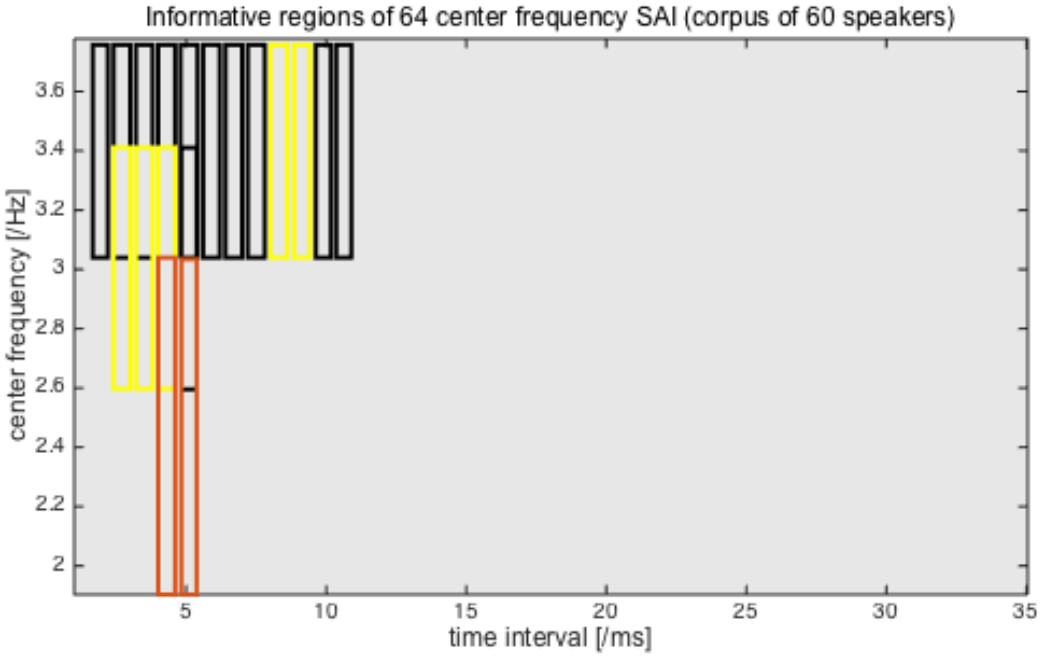


Figure 4.11: Specification of the informative regions of the 64-center frequency SAI (for all trials using 60 speakers). The 18 boxes cover various parts of the filterbank (low and high frequencies) and extend up to 11.2 ms. The black boxes appear to be informative in all trials while the colored boxes (yellow and orange) are discriminative as well for the first and second trial respectively.

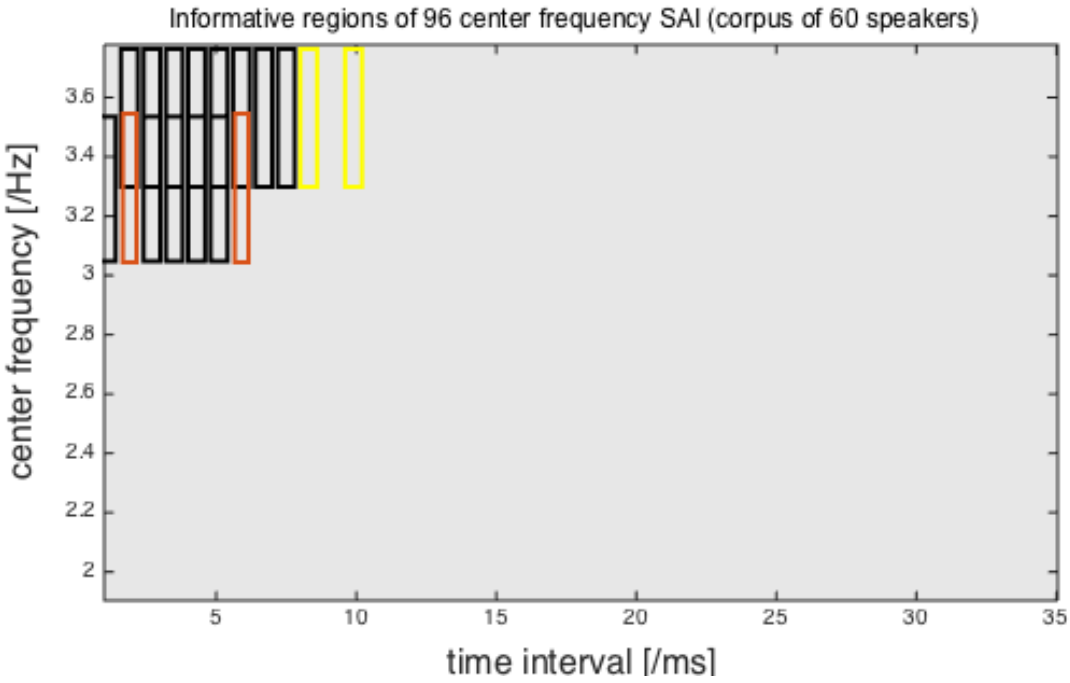


Figure 4.12: Specification of the informative regions of the 96-center frequency SAI (for all trials using 60 speakers). The 17 short and narrow boxes cover part of the filterbank (above 1KHz) and extend up to 10.4 ms. The black boxes appear to be informative in all trials while the colored boxes (yellow and orange) are discriminative as well for the first and third trial respectively.

From the figures above, it is obvious that the most informative areas are located at, approximately, the first 10 ms of the SAI in terms of the time interval dimension. Usually, the first glottal pulse, which forms the first pitch ridge, lies in that time span. As it is known, the glottal pulses are produced by the vocal cords in the larynx and they excite resonances in the vocal tract beyond the larynx. Given that the larynx is quite individual, there is variation among people of different gender and age in how their pitch is distributed. For example, females and children have smaller vocal folds and their overall pitch is higher than that of an adult male.

Pitch is considered to be a source of individuality in the voicing mechanism. Nevertheless, it is not always a valid estimate about a speaker since it can be altered by the person. An example is the case of speakers that modify, i.e. lower or raise, the average pitch of their voice, which results in giving the impression of a person with different size and/or age. Additionally, Kunzel (1989) supported that the GPR is not, importantly, correlated with speaker size in adults when other variables like age and sex are controlled.

Generally, the glottal pulse rate and the resonance scale can vary independently while the perceived message remains the same. This means that despite a person's pitch alteration, the resonances of one's vocal tract will not change because the anatomy remains the same. For this reason, it seems that the formants are more coordinated to a speaker's size and the patterns that associate with them are more critical for speaker identification. From the figures, it is apparent that the boxes, which are located beside each other on a horizontal level, include part of the structures that have been created as a result of resonances. Consequently, they contain information that is speaker – dependent.

As mentioned before, the first glottal pulse usually lies in the time interval of around 10 ms. In view of that, another interesting observation is that the shape of it can carry important speaker information. During voicing, there are differences in the degree of opening or closing of the vocal folds because of their tension. For some speakers, there is complete glottal closure whilst for others, the glottis closes partially. The latter results in a rapid roll-off of the relative magnitude after the glottal pulse (since the glottis stays almost open) and the impression of the produced speech could be characterized as "breathy" (Laver, 1994; Kinnunen, 2004). This influences the general voice quality of speakers, which is a more instinctive notion of the differences among them. Given that it depends on intuition, it is difficult to be measured in a reliable way

(Eskelinen-Roenkae et al.,1999). Still, it can be an indication that can be combined with other features for identifying a person.

Furthermore, a very appealing finding of the VQ process is that the high frequency range of speech may contain meaningful speaker information. Zhou et al. (2011) has worked on speaker recognition and compared the performance between cepstral coefficients that are derived using linear frequency and mel-frequency scales. One of his conclusions was that that higher frequencies should not be overlooked. In figures 4.7 - 4.12, the informative boxes that capture SAI features support the fact that apart from the lower formants, the higher formant frequencies can be distinctive as well.

In particular, when the SAI has a bank of 32 filters, the boxes cover the complete frequency range including lower (F1,F2) and higher (F3 and above) formants. The size of this filterbank seems to be quite small because the patterns are merged in single boxes shown in figures 4.7 and 4.10. Nonetheless, it seems that the use of 64 and 96 filters provides better frequency resolution since there are a number of smaller boxes on various positions. From the figures 4.8, 4.9, 4.11 and 4.12, it is obvious that the patterns contained in those boxes are more local and oriented in the frequency range above 1KHz. Additionally, the boxes in figure 4.11 cover a small low – frequency area as well. These findings support what has already been stated in literature that, except for the two lowest formants (F1, F2) that contain roughly most of the phonetic information related to vowels, F3 and higher formants are probably more speaker-specific.

In conclusion, the suggested approach on using VQ in an alternative way, i.e. creating multiple codebooks instead of a single codebook for each one of the speaker models, makes it possible to retrieve notable information about what makes a difference in the perception of speaker size and/or gender. In each case, the rectangles may vary in terms of where they are located but there is a convergence of their positons indicating that the area of interest is located at, approximately, the first 10ms of the auditory image.

As a result, the information included in this smaller time interval supports the hypothesis that perceptual differences can depend on lower - dimensional auditory features.

4.4 Discussion

Speaker recognition is a challenging task that depends on a number of parameters. In this chapter, we have described an identification system that uses a biologically inspired auditory model and learns a matching between known and unknown speakers in quiet conditions.

At first, a series of experiments was performed to estimate the performance of the proposed system and compare it to the results obtained by the MFCC-based system. The experiments were conducted for a small and a larger speech corpus. The results from this experimental set support our initial hypotheses that the new front - end, which mimics characteristics of the human auditory system, can provide an effective representation for speaker identification that has similar performance to the MFCCs.

Furthermore, we assumed that it may be possible to characterize the filterbank size of the AIM, for the task at hand, in order to achieve the best possible SID accuracy. Based on the identification results of the data set of 180 speakers (since the results of the 30-speaker data set indicate successful recognition for all cases), it appears that the performance of the auditory features is better for the case of 64 frequency channels. As a result, our assumption can be supported and this finding will be combined with the results of the second part of this set of experiments. Thus, after having obtained satisfying results for the behavior of the auditory model with regard to an established method for feature extraction, the next stage is about working on potential improvement of the proposed front - end.

The second part of this experimental set consisted of proposing an alternative way of using VQ for obtaining more speaker – specific information from the SAI. The latter was based on the hypothesis that it may be possible to choose a lower - dimensional SAI representation after taking into account these informative regions.

After specifying the discerning patterns of the image, it is noticeable that their positions cover specific parts of the SAI. In terms of the time – interval dimension, the boxes converge to the area up to, roughly, 10 ms. Moreover, it appears that, above 10 ms, there are not any rectangles that indicate distinctive patterns for speakers. This is an interesting observation since it appears that, after the first pitch ridge, the SAI contains a repetition of previous patterns, especially in the case of stable sounds. As a result, the information included in that region is redundant. With regard to the frequency dimension, the location of the rectangles depends on the frequency resolution. For a smaller filterbank, the boxes are taller covering the complete frequency range. For a larger number of frequency channels, the boxes correspond to more local patterns in the middle and high frequency spans. Thus, the general impression is that the boxes cover structures that are relatable to characteristics of the speech signal, such as the glottal pulse and the formant frequencies.

Consequently, it appears that our hypothesis to choose a subset of the extracted auditory features with lower dimensions is supported. This has also been suggested by Walters (2011), who stated that there are substantial redundancies in the SAI and if it is to be used as the basis of a recognition system, it would be advisable to try and find a compact feature representation that summarises a signal.

Moreover, those findings helped us attain further insight about the effect of the filterbank size. For this task, it seems that a very high number of frequency bands (i.e. 96) may result in highly correlated features because this size gives too much spectral resolution (that can also produce spurious formants).

Inversely, a low number of frequency bands (i.e. 32) may result in uncorrelated features. This can cause information loss and consequently, not clearly recognizable patterns (as shown in figures 4.7 and 4.10).

Additionally, from the results of table 4.7, it appears that the choice of 64 filters does not provide recognition rates that are significantly different from those for 32 or 96 filters. However, the 64 frequency bands have produced the highest mean SID accuracy. After taking into consideration all of the above, it seems that the bank of 64 filters is a sensible choice for having enough frequency resolution for reliable patterns.

In conclusion, the results of testing these hypotheses helped us gain insight about the advantages of using the auditory model. Since the proposed system has only been tested in quiet conditions, it is interesting to incorporate challenging conditions that associate with real-world environments. After taking into account the achieved knowledge from these experiments, we will develop the initial version of the SAI-based system and use it to evaluate its performance in the following chapter.

Chapter 5

Noise-Robust Speaker Identification

5.1 Introduction

In chapter 4, the findings of the experiments have shown that the proposed system consisting of the AIM-based front-end can have satisfying performance in quiet conditions. Additionally, the informative SAI regions have been specified and this knowledge is used for the improvement of the system configuration in this chapter. Furthermore, the aim of this experimental set is to use the SID system that consists of the developed feature extractor in the presence of distortions. Generally, in everyday environments, interfering sounds pose considerable challenges to speaker recognition systems. A lot of research has been devoted to deal with such difficulties and it is very interesting to study the behavior of the auditory features with regard to such conditions.

At first, our hypothesis is that the feature representation from the SAI will produce better SID accuracy levels compared to the system that uses the MFCC parametrization. The motivation behind it is that the auditory image retains the fine timing information and contains the relative magnitudes of all the frequency bands whilst including the locations of the high-magnitude frequency regions (resonances). This characteristic may prove to be robust to noise-corrupted speech, since, perceptually, more noise can be tolerated around spectral peaks. This hypothesis will be tested for two speaker data sets since the size of the speaker population affects the outcome of the identification process (Rose, 2002).

Secondly, we hypothesize that the length of the speech material used for the training session affects the levels of SID accuracy. Larger amounts of training data are expected to result in better recognition rates. However, it is not always possible to have adequate training material (Rose, 2002). For that reason, we are motivated to investigate how the performance of the proposed system is affected for various durations of the training speech utterances.

Furthermore, our third hypothesis is that the length of the test data can also influence the accuracy of the identification process. Longer test speech segments are expected to produce higher accuracy levels since more information can be extracted to represent the target speakers. Nevertheless, in real-world situations such as forensics, it is not always feasible to have sufficient test speech material. Therefore, it is interesting to examine how the outcome of the identification process is influenced by the variation in the test speech length.

The rest of this chapter is organized as follows. Section 5.2 describes the system overview as it has been implemented for the developed AIM-based system. The MFCC-based system that will be used in this experimental set is the same that has been described in section 4.2. Section 5.3 contains the results of all the experiments that have been conducted in order to test the hypotheses that have been described above. Finally, in section 5.4, the outcomes of the experiments are discussed.

5.2 System Overview

In this section, the modules of the proposed system are reviewed in detail. As mentioned in the previous chapter, the findings of the first part of experiments will help us in improving the AIM-based system. The developed system is more computationally efficient, which is a consequence of modifications that are explained as follows.

5.2.1 AIM-based system

Firstly, the front-end consists of the AIM, the box-cutting and the downsampling procedures. The system design is shown in figure 5.1 and it consists of the same modules as the one used in the previous chapter.

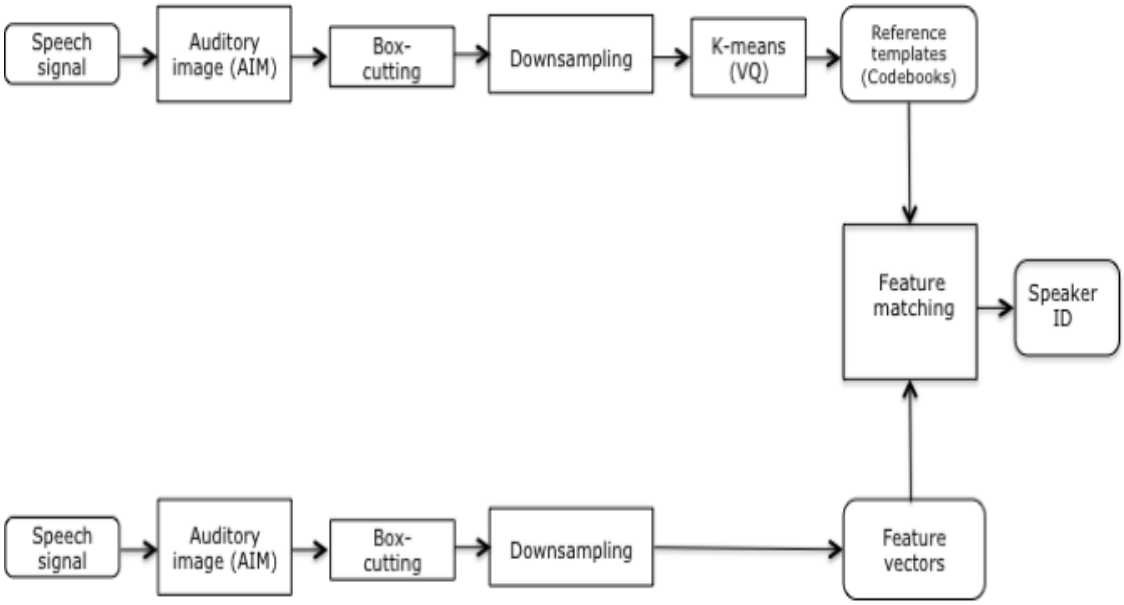


Figure 5.1: Schematic design of the SID system using AIM as a front end

As before, the BMM module simulates the auditory spectral analysis with a gammatone filterbank. Based on the results for the discriminative boxes, we have chosen to use 64 channels for the filterbank that cover the frequency range from 80 Hz to 6 KHz for all of the experiments that follow. This choice provides a good compromise between adequate frequency resolution for clear patterns and controlled feature dimensionality since the filterbank size is directly associated with the box-cutting procedure.

Furthermore, the neural activity pattern is obtained as before after the three sequential operations of the NAP module of AIM: the half-wave rectification of the filterbank output, the logarithmic compression that simulates the cochlea compression and the low-pass filtering corresponding to the loss of phase-locking. Then, the strobe-finding process determines the position of important peaks at the output of each channel. The strobe points initiate the temporal integration processes for all of the channels as described before. This process adds the dimension that represents the time interval from the strobe points that lasts for 35 ms. The end result is the 35 ms SAI, which is a sequence of two-dimensional frames of real-valued data. All of these frames are the input for the box-cutting step.

According to the results of chapter 4, the boxes in the figures vary in size based on the frequency resolution of the filterbank. For the SAI with 64 cochlear channels, most of them are intermediate in height and narrow in width including a variety of localized features. Nonetheless, it is noticeable that most of the boxes are placed alongside each other and if they are added up altogether, they cover the image structure on a larger scale as well. As mentioned previously, the positions of these rectangles converge in around the first 10 ms of the time interval dimension. This leads us into assuming that it would be more functional to segregate this specific region of the SAI where the boxes are located. This will reduce the redundancy of data, which is a result of using the whole auditory image, and improve the computational efficiency of the clustering algorithm.

In consequence, we have chosen to modify the box-cutting process based on these results. After taking into account the spatial convergence of the notable boxes, the SAI is cut into one rectangle that includes all of them. This larger box that includes the best subset of boxes covers the whole of the filterbank and stops at 12.8 ms (which corresponds to 256 time samples).

The choice of this box size has been made because, throughout the initial box-cutting procedure, one of the parameters was that the dimensions of the rectangles were equal to powers of 2. These dimensions were, then, converted into the smallest pair of dimensions, i.e. 32 x 16 pixels, which are powers of 2 as well. For computational convenience, we choose the box that covers up to 256 samples (which is a power of 2) instead of 200 time samples (which corresponds to 10 ms). Additionally, in terms of the frequency dimension, the choice of covering all of the frequency bands is based on trying to maintain the spectral formation and not disassemble structures caused by resonances. The final outcome of the box–cutting module is one rectangle covering 64 frequency channels by 256 time samples and it is shown in figure 5.2.

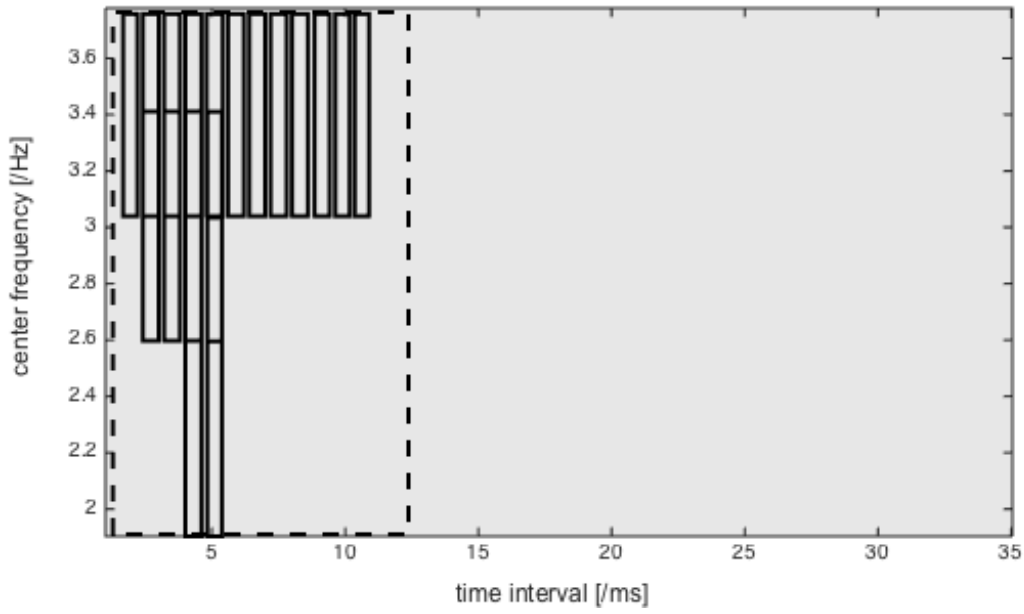


Figure 5.2: Selection of the discriminative areas of the SAI as determined by the VQ process of the first experimental set

After cutting this specific box size and shape for the total number of frames, all of these rectangles are downsampled to boxes with the smallest pair of dimensions (32 x 16 pixels). Then, all of them are reduced to 48 values each, after averaging the elements on both dimensions and concatenating them in one feature vector. The end result is a number of 48-element feature vectors equal to the number of SAI frames.

This modification in the way of extracting features is a salient contribution for the reduction of feature dimensionality and, at the same time, it provides a robust representation that can capture complex structures in the SAI data. In addition, this change makes the system more computationally efficient.

After completing the process of converting the extracted data from the SAI frames into feature vectors, the speaker modeling step follows. In this set of experiments, the dimensionality is reduced, compared to the initial system design, but it is still important to represent the data with dimensions that are further decreased. To achieve that, the K-means clustering algorithm is used but, it is now applied only in the context of feature matching.

In practice, during the enrolment session of a speaker, one codebook is created for the one box that is cut over the complete number of SAI frames. This means that the input of the clustering algorithm is the features contained in that one rectangle (over all frames). Therefore, the end result is a single codebook for each speaker instead of the multiple codebooks included in each reference template that was used before.

The first step of this process is to concatenate all the 48 – element feature vectors that represent each box over all of the training frames. Given the average durations of the different speech corpora consisting of 30 and 180 speakers (that will be described in the following section) and that the frame length is 10 ms, the average numbers of frames is estimated by dividing the duration with the frame length.

For the 30-speaker corpus, the average number of frames is 1420, which results in 1420 feature vectors on average. Considering the same parameters for the corpus of 180 talkers, the average number of feature vectors is 2070. These numbers reflect the feature dimensions before the vector quantization procedure. After concatenating all of them for the total number of frames, we have the representation of the entire speech signal.

Afterwards, the clustering is achieved through the K-means algorithm using 64 codewords per codebook. In general, it is expected that larger codebook sizes give better recognition rates. However, if the codebook size is too high, it learns the training samples but not the general data distribution (this is called overfitting) (Manning et al., 2009). Usually, an overfit model of data may exaggerate minor fluctuations in data which results in incorrect generalization.

According to existing literature, the range for minimum number of codewords is between 16 and 64, and it depends on the feature dimensionality and the amount of training data. As a result, the choice of 64 is in between sizes of codebooks that are considered as very small or very large. Additionally, an advantage is that it results in a reasonable amount of computational time for the codebook generation. The size of the codebooks remains constant throughout all of the experiments. This has been done in order to focus our attention on the feature sets and how they affect the SID accuracy of the system.

The outcome of the VQ process is a number of speaker templates equal to the number of trained speakers. Each template is a single codebook with a size equal to 64 x 48 elements. Also, this is a similarity between the SAI-based and the MFCC-based systems. The latter uses the single codebook approach for the speaker templates, which are equal to the number of trained speakers and consist of 64 x 40 elements as described before. Figure 5.3 shows the steps for creating the reference template of one speaker from the SAI features that are extracted from one’s speech. The process is repeated as many times as the number of trained speakers.

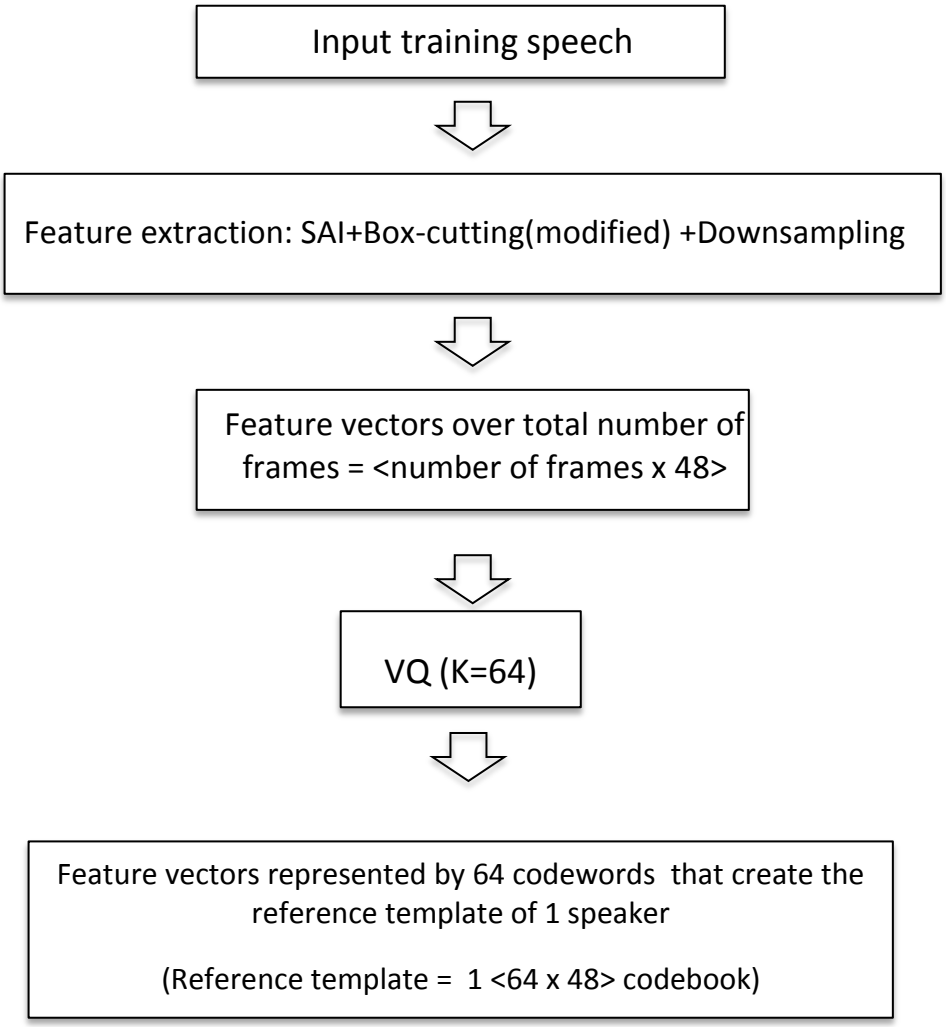


Figure 5.3: The steps of the enrolment (training) session for creating the template of 1 speaker that contains only 1 codebook. The same process is repeated for the total number of speakers in the database.

Then, during speaker testing, the feature extraction stage follows the same steps as the ones in the training session. For each one of the frames of the test speech segments, one box is cut covering the filterbank and the first 12.8 ms as before. For every box of every frame, the downsampling step results in a vector with 48 elements and finally, all of these vectors are concatenated in order to obtain the complete feature representation of the signal.

For the speaker matching stage, the principal idea is to see how good the reference template of each speaker is for encoding the features of the target speaker. To obtain that, the values that reconstruct every frame of the test speaker using each one of the trained speaker models are estimated. For every frame, the Euclidean distance is computed between every centroid in the codebook and the current feature vector.

The result is a matrix of all the distances between the complete number of frames and the centroids. For every frame, the minimum of these distances is the reconstruction value for that frame using that codebook. Then, those minimum distances are estimated for the total number of frames (the whole speech utterance) and finally, their average is computed. This value is the mean reconstruction value and the speaker, which is most likely to be the target speaker, is the one corresponding to the smallest average reconstruction value. After completing this step, it is possible to estimate the SID accuracy as the number of correctly identified speakers over the entire speaker population used for the testing phase. The steps of the procedure for speaker matching is explained analytically in figure 5.4.

In conclusion, through the described process, we managed to create a recognition system that makes a good comparison with the baseline system, since we have chosen a low-dimensional representation from the SAI that has similar size to the MFCCs. Our subsequent processing is based on this type of SAI representation. In section 5.3, the performance of the developed system is evaluated through comparing it with the system that adopts the MFCC parameterization. The design of the baseline system is the same as it was explained in section 4.2.

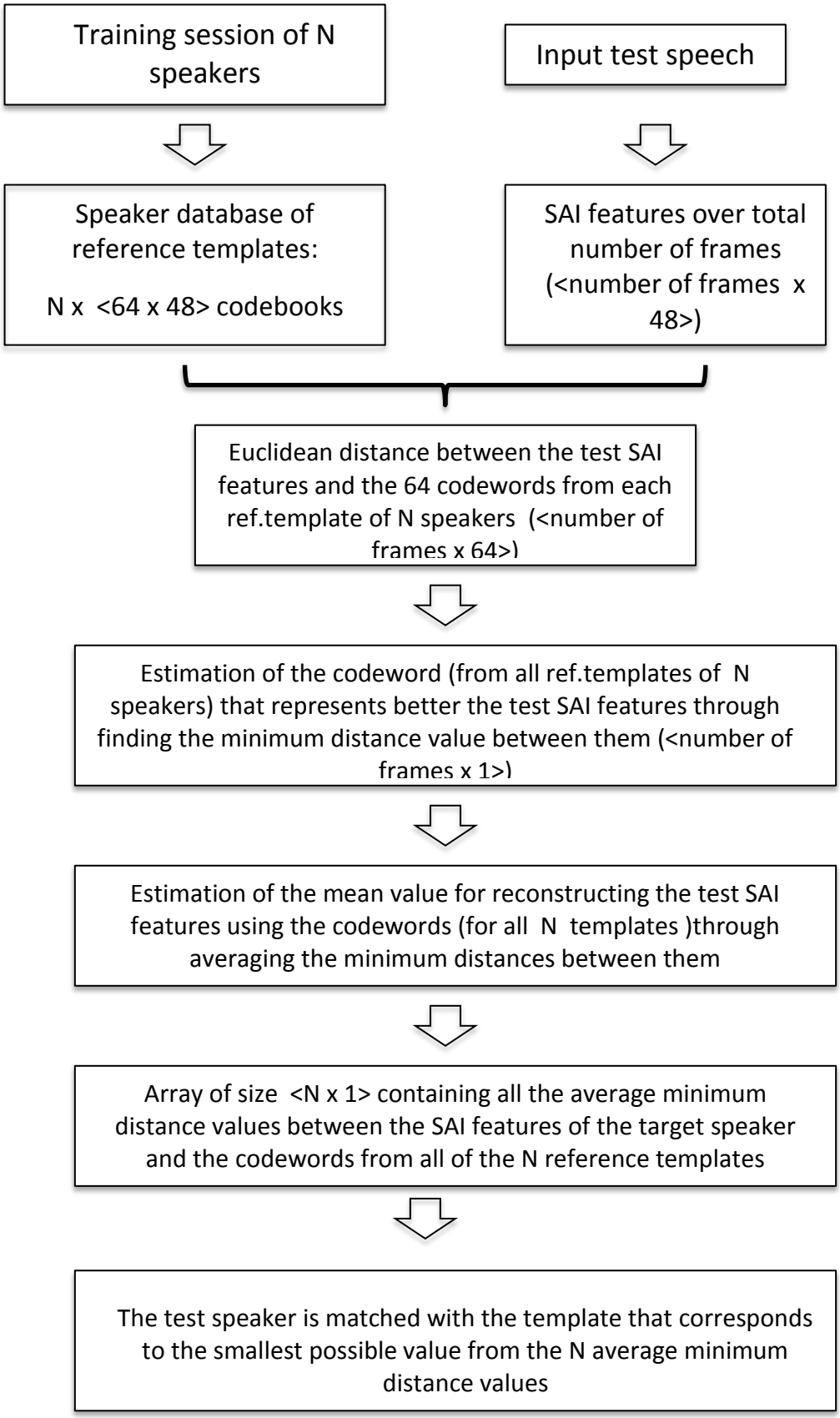


Figure 5.4: The steps of the speaker matching stage for 1 test (target) speaker and N trained (enrolled) speakers (i.e. N reference templates and N codebooks). The same process is repeated for the total number of test speakers.

5.3 Evaluation

5.3.1 Performance on the EUROM1 corpus of 30 speakers

In this section, the auditory features are evaluated in the speaker identification task using a small set of the speech corpus and our first hypothesis is tested. In what follows, the dataset and the experimental results are described.

5.3.1.1 The Data Set

The speech corpus that will be used in this work is the same multilingual corpus, named EUROM1, that has been used in the first set of experiments. As previously, we randomize the system in terms of the groups of speakers so that we estimate the variability of the accuracy when there is disparity in the speaker population. The 30-speaker speech corpus consists of 3 groups of 10 speakers. The speech material that has been chosen for each speaker is in the form of passages consisting of 5 sentences. The speech signals have been pre-processed for pause removal and the proposed SID system performs text-independent recognition.

To study how the proposed system performs for interfering sounds, the test speech utterances are mixed with babble noise of 8 talkers, which is non-stationary. The audio file of the noise was selected from a database of sounds used in our laboratory. The RMS level of the clean test speech was used as the reference level to establish the SNR. The choice of the speech babble was based on the fact that it is one of the most challenging conditions and at the same time, it reflects a realistic environment (Krishnamurthy et al., 2009). A summary of the corpus of 30 speakers that has been employed and the experimental set is given in table 5.1.

Corpus of 30 speakers	
Language	English
Speakers	30 (12 F+ 18 M)
Speech type	Read speech
Sampling frequency	20 KHz
Training speech type	Clean
Training speech duration (avg)	14.2 sec
Test speech type	Noisy
Type of noise	Multi-talker babble
Test speech duration(avg)	15 sec

Table 5.1: Summary of the speech corpus consisting of 30 speakers

5.3.1.2 Results

For this set of experiments, our hypothesis that the proposed system will be more robust in the presence of noise than the baseline system is tested. The training sessions were performed using clean speech. Each test utterance is mixed with noise at various SNR levels from −5 dB to 10 dB, at 5 dB intervals, so that we can test the effect of noise level in the system performance. Figure 5.5 plots the SID accuracy against the SNR levels. The error of the identification score is estimated as the standard error of the mean among the 3 subsets of 10 speakers that make up the 30-speaker corpus.

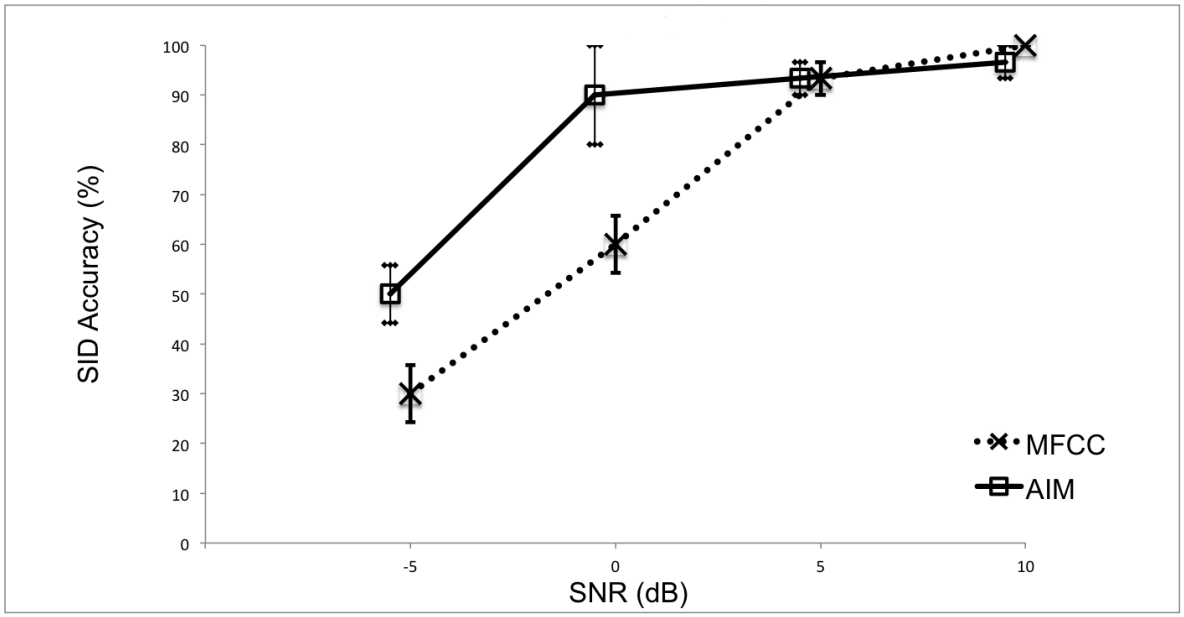


Figure 5.5: Speaker identification (SID) accuracy (%) of the SAI-based and MFCC-based systems for the corpus of 30 speakers using multi-talker babble noise. The error bars represent the standard error of the mean (estimated as the error among the levels of SID accuracy of the 3 subsets of 10 speakers)

In general, the results in the figure above indicate that the SAI-based system produces satisfying accuracy levels for all of the noise levels. This shows that the proposed feature extractor, with the modified box-cutting process, works well as expected. The auditory features are significantly better than the MFCCs for -5 and 0 dB SNR. This has been determined by repeated measures ANOVA (where the factors are SNR and type of method (MFCC or AIM) with p-value = 0.0124) and post hoc analysis using t-tests at -5 dB SNR (p-value = 0.0352) and 0 dB SNR (p-value = 0.0301).

Specifically, the auditory features perform better than the MFCCs for the lower SNRs. For 0 dB SNR, this configuration achieved on average 90% identification (with a 10% error), while the baseline system achieved 60% recognition (with a 5.77% error). Also, SAI features achieve better accuracy at -5 dB SNR that is equal to 50%, compared to 30% obtained by the MFCCs (with a 5.77% error for both cases). Although the average improvement in identification is lower for this level of noise, it is obvious that the auditory features are better in accomplishing the recognition task for demanding circumstances where noise is dominating.

Interestingly, as the noise level decreases, performance seems to be reaching a point of saturation. For 5 dB SNR, the two systems have equal identification accuracy of 93.33.% with an error of 3.33%. Maximum accuracy is attained for 10 dB SNR, where the scores are quite similar. MFCCs achieve 100% identification whilst the SAI reaches 96.66% with an error of 3.33%.

5.3.1.3 Case study

From this experimental set, there are some notable observations regarding systematic differences between speakers. First, for 0 dB SNR, the first group of 10 speakers reaches 70% correct identification. Table 5.2 shows the misidentification results of this specific group and SNR.

Actual Speaker Index	**Hypothesized Speaker Index**
3	10
6	10
8	10

Table 5.2: Speaker mismatches of the 1st group of 10 speakers for SNR= 0 dB

From the results in table 5.2, it is interesting that speakers 3, 6 and 8 have been consistently misidentified for the same person. The speakers' characteristics are described in table 5.3. From the information provided by the appendix of the corpus, it is known that all of them are females.

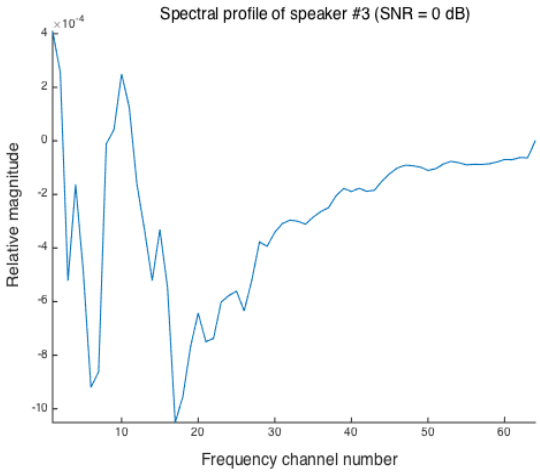
Speaker Information	**Speaker Index**			
	3	**6**	**8**	**10**
DoB	1966	1939	1961	1965
Height	1.80	1.68	1.57	1.62
Nationality	English/German	English	English	English/German
Smoker	Yes	No	No	No
Pathologies	Pneumonia/Tracheostomy	No	No	No

Table 5.3: Description of the misdentified speakers 3, 6, 8 and 10

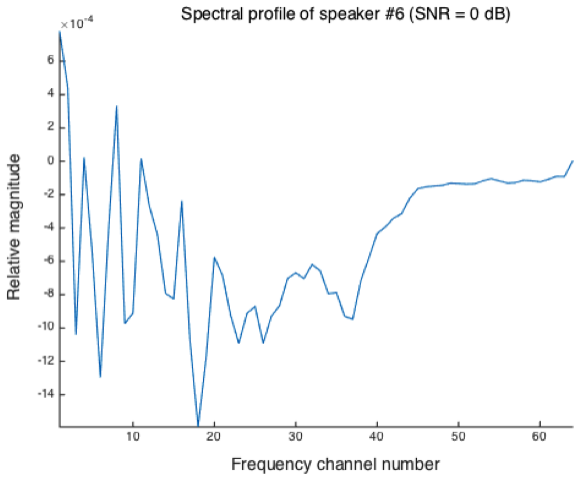
Firstly, it is worth noticing that speaker 10 has a height, which is the average of the heights of speakers 6 and 8. Also, the accent of speaker 10 may be a parameter for the mismatch. Moreover, speaker 3 is much taller than speaker 10. However, this inconsistency is quite reasonable since the person had a tracheostomy.

In the case of a tracheostomy, when the tube is inserted, most of the air bypasses the vocal cords and goes out through the tube. Depending on the fitting of the tube inside the trachea, some amount of air may leak up to the vocal cords but it may not be enough to cause the vocal cords to vibrate or it may allow enough force for very short utterances. Usually, the aim is to use the smallest trach tube possible for the patient. The latter may explain why speaker 3 with 1.80m height and consequently, a longer vocal tract, is hypothesized to be speaker 10 with a shorter vocal tract, given that her height is 1.62m.

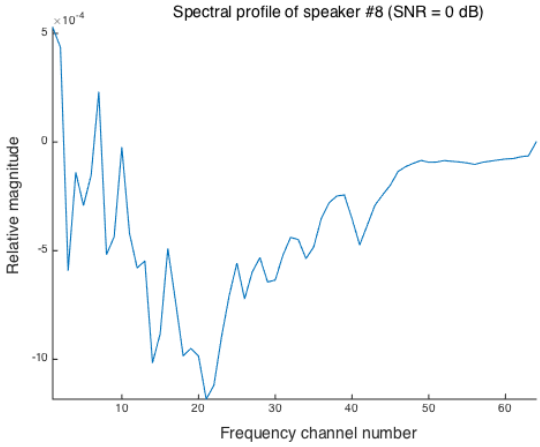
The spectral profiles of these speakers can also provide interesting findings. In general, the calculation of spectral profiles is achieved through averaging in each channel over time. Nonetheless, when they are estimated for the total number of frames, the final result will contain much broader peaks. For that reason, they are computed using a shorter speech segment with 2 sec duration, which helps in obtaining more clear spectral information about the utterances. Figures 5.6 ((a)-(d)) and 5.7((a)-(d)) show the spectral profiles and spectrograms of speakers 3, 6, 8 and 10 respectively.



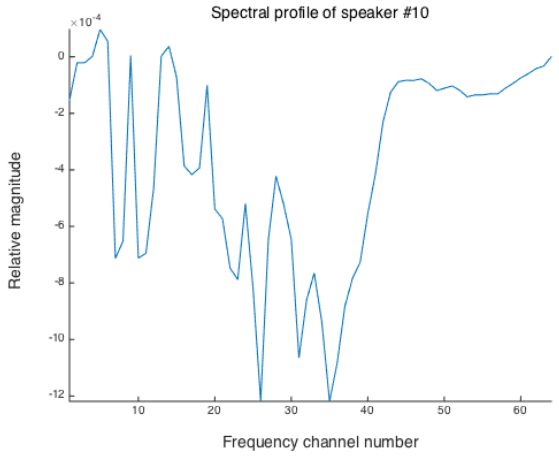
(a)



(b)



(c)



(d)

Figure 5.6: Spectral profiles of (a) speaker 3 using speech mixed with noise at 0 dB SNR, (b) speaker 6 using speech mixed with noise at 0 dB SNR, (c) speaker 8 using speech mixed with noise at 0 dB SNR and (d) speaker 10 using clean speech.

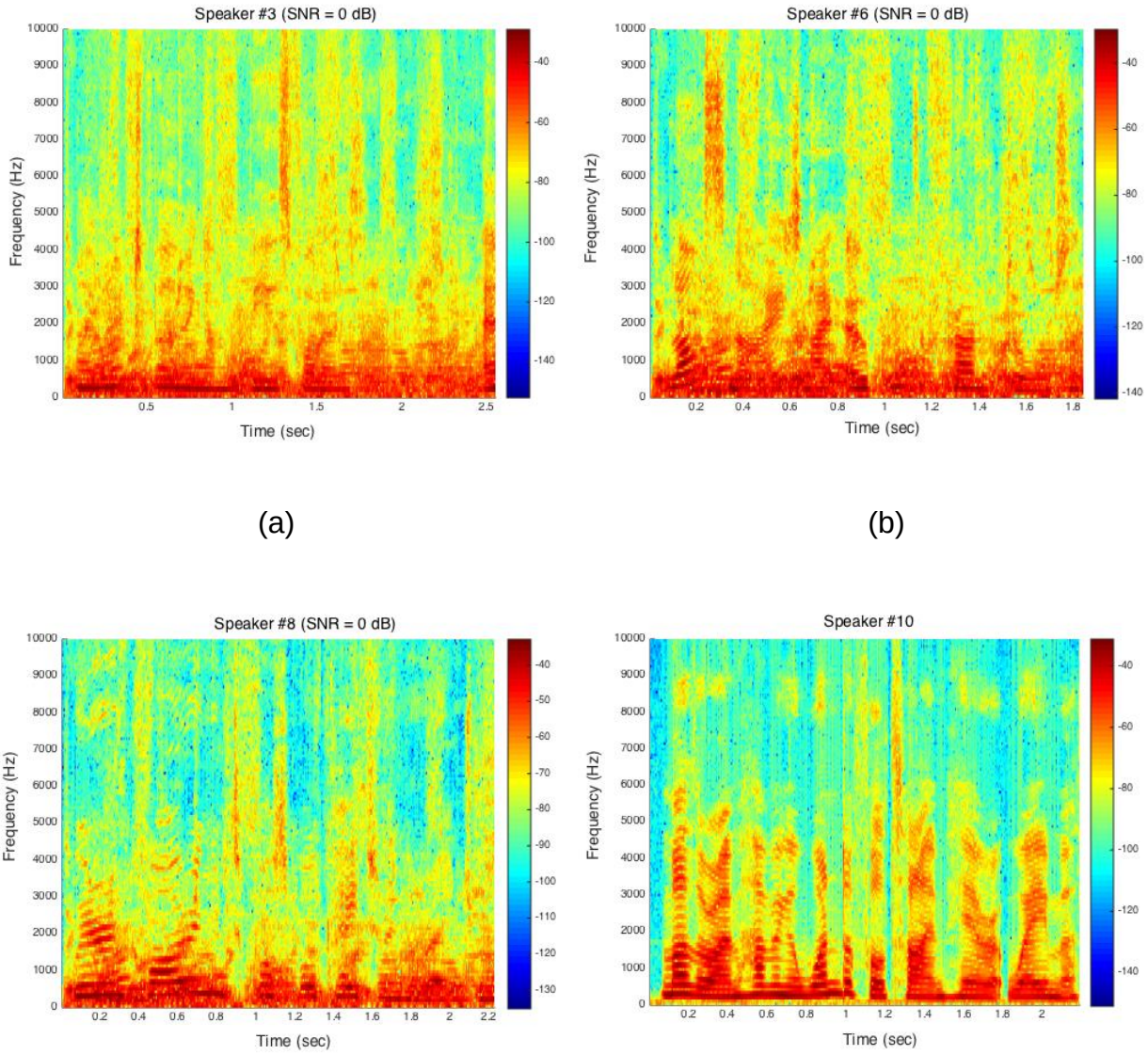


Figure 5.7: Spectrograms of (a) speaker 3 using speech mixed with noise at 0 dB SNR, (b) speaker 6 using speech mixed with noise at 0 dB SNR, (c) speaker 8 using speech mixed with noise at 0 dB SNR and (d) speaker 10 using clean speech

From the spectral profiles of speakers 3, 6 and 8, it is obvious that there are many spectral peaks below the 30th frequency channel, which corresponds to approximately 1KHz. This is also apparent in the spectrograms where there is a lot of acoustic energy concentrated below 1KHz. For all of these speakers, the peaks seem to be located around the same frequency regions. More specifically, the frequencies that correspond to them are around 126 Hz (frequency channel 4), 200 Hz (frequency channel 8), 242 - 265 Hz (frequency channel 10 and 11), 367 Hz (frequency channel 15), 524 Hz (frequency channel 20) and 724 Hz (frequency channel 25).

Provided that formants occur at, approximately, 1000 Hz intervals, it is expected to see one in each 1000 Hz band. In this case, the formant F1 is expected to be found in the area up to 30th channel. However, the presence of babble noise makes the tracking of it difficult because of the creation of spurious formants.

Considering that there is a darker area in all of the spectrograms (indicated by the dark red color) around 250 Hz, it is possible that F1 associates to the peaks located around channels 10 and 11, but it cannot be guaranteed. Further, speakers 6 and 8 seem to have peaks between the 30th and 40th frequency channel, which correspond, rougly, to the interval between 1 and 2 KHz. In their spectrograms, this region seems to have more energy present as well. Lastly, all three of them seem to have the same magnitude level in the frequency area above 2 KHz (above the 44th frequency channel) and up to 6 KHz. This region relates to their higher formants.

As for speaker 10, her spectral profile consists of spectral peaks around the frequency channels 5 (144 Hz), 10 (242 Hz), 15 (367 Hz), 19 (490 Hz), 24 (680Hz), 30 (1KHz) and 33 (1160Hz). Additionally, her spectrogram shows more acoustic energy concentrated up to 2 KHz approximately (indicated by the dark red color). Since the speech is clean, this is an indication of the locations of F1 and F2. Also, the higher frequency region above 2 KHz seems to have the same level of magnitude as the other 3 speakers.

In conclusion, these 4 speakers have similiarities in terms of spectral content. These are caused by the presence of noise, which results in spurious spectral peaks that are misinterpreted to be formants, and they affect the associated patterns as well. Thus, this can explain the consistency in the false identification between them. Importantly, this speaker mismatch denotes the importance of formants for the SID task. Since the VTL is directly associated with the resonances, it also points out the influence of the vocal tract for achieving correct recognition.

Lastly, the following spectral profiles in figure 5.8 ((a) - (c)) correspond to randomly chosen speakers (1,5 and 9) that were not confused with the target speakers 3, 6 and 8. From the figures, it appears that the overall shape of their spectral patterns differs. More specifically, from the spectral profiles of speakers 1, 5 and 9, it is obvious that there is one spectral peak below the 10^{th} frequency channel (242 Hz) and another one between the 10^{th} and the 20^{th} channel (242 - 524 Hz). Moreover, another notable fact is that their spectral shapes above the 20^{th} channel are very different compared to those corresponding to speakers 3, 6 and 8. As a result, it seems that the outline of the pattern is the reason that these speakers have not been confused for the target ones and it shows that the similarity between the spectral profiles of the mismatched speakers has not happened by chance.

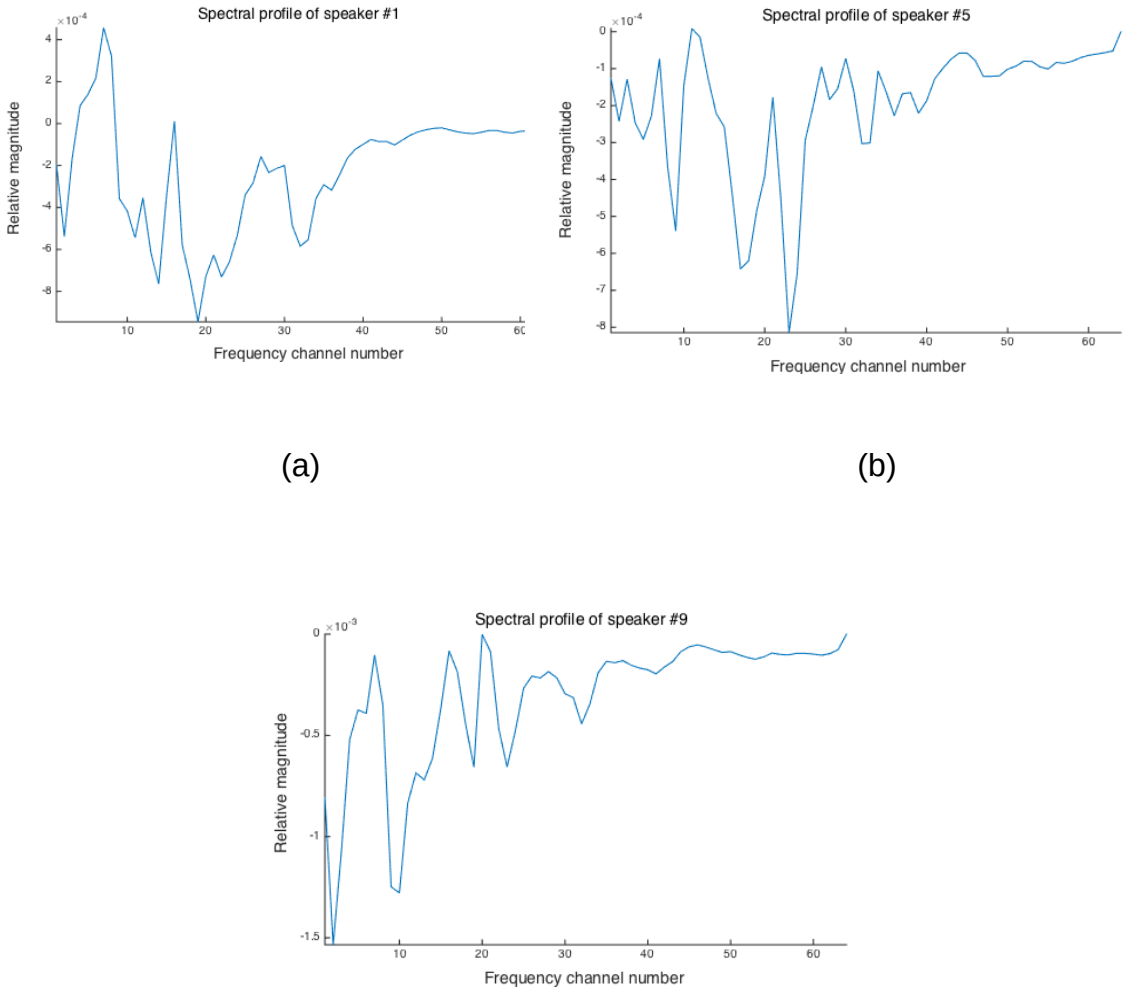


Figure 5.8: Spectral profiles of (a) speaker 1 using clean speech, (b) speaker 5 using clean speech and (c) speaker 9 using clean speech. These speakers belong in the same group of 10 speakers with speakers 3, 6 ,8 and 10. The test speakers 3,6 and 8 are confused with the reference template of speaker 10 (for 0 dB SNR) but they are never confused with the templates of speakers 1, 5 and 9 (for all SNRs)

5.3.2 Performance on the EUROM1 corpus of 180 speakers

In this section, the auditory representation is assessed for speaker identification using an extended speech corpus. Similarly to the previous section, our first hypothesis is tested for a larger speaker population. This happens so that we can examine the behaviour of the proposed system since the task becomes more challenging as the number of speakers increases. The data set and the experimental results are described as follows.

5.3.2.1 The Data Set

For this series of experiments, the speech material is chosen from the EUROM1 database as previously. In this case, the system is used for an enlarged population of 180 speakers. In order to test the variability of the system performance, the population is divided into 3 different groups that consist of 60 subjects each. As before, the speech material for each speaker is in the form of passages consisting of 5 sentences. The speech signals have been pre-processed for pause removal and the proposed SID system conducts text-independent recognition so the training and test speech utterances differ.

The aim is to test the system for truthful conditions that make the identification task as demanding as possible. For that reason, clean speech is used to train the speakers while the testing is done using utterances in the presence of babble noise created by 8 talkers. The audio file of the noise is the same that was used in the previous experiment. The SNR was established through using the RMS level of the clean test speech as the reference level. The details of the speech corpus and the experiment are outlined in table 5.4.

Corpus of 180 speakers	
Language	English, French, Swedish
Speakers	180 (90 F+ 90 M)
Speech type	Read speech
Sampling frequency	20 KHz
Training speech type	Clean
Training speech duration (avg)	20.7 sec
Test speech type	Noisy
Type of noise	Multi-talker babble
Test speech duration(avg)	20.6 sec

Table 5.4: Summary of the speech corpus consisting of 180 speakers

5.3.2.2 Results

In the previous experimental set, the hypothesis was satisfied for a small speaker population. In this experiment, the same hypothesis is tested for groups of people that are 6 times larger. Figure 5.9 plots the speaker identification accuracy against the SNR levels (ranging from -5 to 10 dB, at 5 dB intervals) for both the SAI and the MFCC features. Again, the error of the identification accuracy is estimated as the standard error of the mean among the 3 subsets of 60 speakers that compose the 180-speaker corpus.

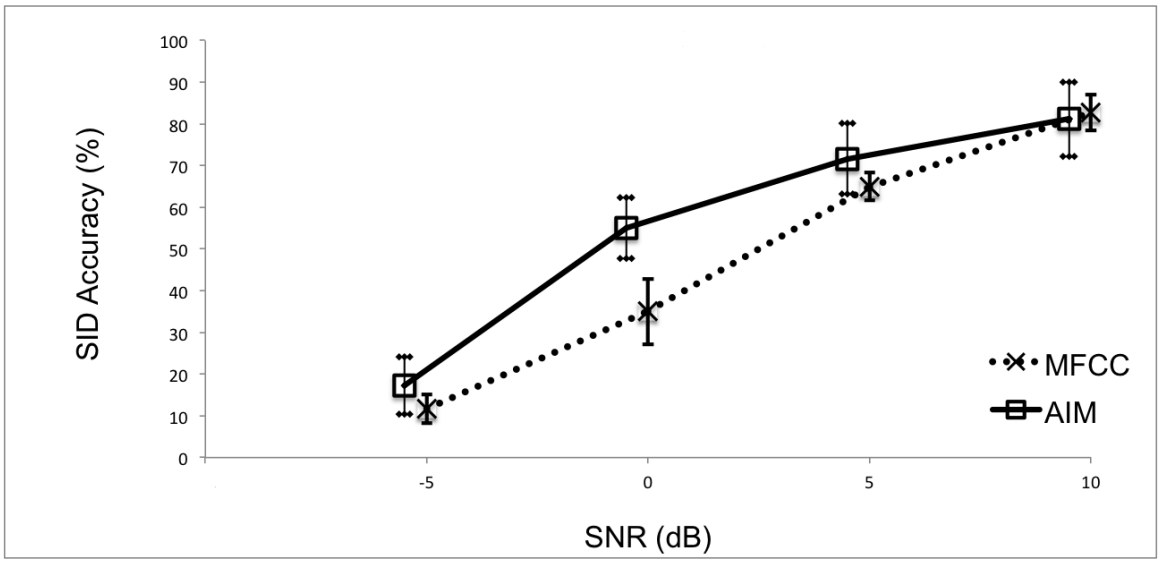


Figure 5.9: Speaker identification (SID) accuracy (%) of the SAI-based and MFCC-based systems for the corpus of 180 speakers using multi-talker babble noise. The error bars represent the standard error of the mean(estimated as the error among the levels of SID accuracy of the 3 subsets of 60 speakers)

From the results shown in the figure, it is clear that our hypothesis for the auditory features is justified despite the change in the number of speakers. Given the expansion of the speaker population, it was anticipated that the recognition rates would decrease to a certain extent. Moreover, the SAI features may not be significantly better than the MFCCs (as determined by repeated measures ANOVA) but still achieve higher SID accuracy especially for low SNRs.

In particular, the result for the accuracy that is achieved for 0 dB SNR is notable since the recommended configuration reaches on average 55% recognition (with a 7.26% error) compared to 35% (with a 7.8% error) obtained by the MFCCs. For even more noisy conditions, reflected by -5 dB SNR, the average identification is almost 5.5% higher for the auditory features (17.22% with 6.82% error). Still, the system shows consistency in achieving better recognition rates for demanding conditions.

Furthermore, as the SNR increases, there seems to be a saturation of performance with comparison to the baseline system. Yet, for 5 dB SNR, there is still better identification score (71.66% with 8.55% error) whilst for 10 dB SNR, there is convergence of the outcomes of both systems, i.e. 81.11% with 8.94% error for the SAI and 82.77% with 4.33% error for the MFCCs.

5.3.2.3 Case study

Overall, the auditory features seem to perform very well in terms of recognizing people of the same gender. This type of speaker matching is considered to be more challenging compared to the trials between men and women. The reason behind that is the fact that, in general, the VTL difference between males and females makes their formant structures differ a lot and this characteristic makes it less difficult to identify them correctly.

Yet, for the experiments in the 2nd set of 60 French talkers, there is a specific mismatch between a female (speaker 40) and a male speaker (speaker 1). This incorrect identification is interesting because it is the only one that happened between 2 people of different gender and it was consistent for all SNRs (-5, 0, 5 and 15 dB). As a result, it is worth trying to gain insight about the possible causes of this mismatch. The characteristics of the 2 speakers are summarized in table 5.5.

Speaker Information	**Speaker Index**	
	40	**1**
Gender	Female	Male
Age	51	46
Height	1.61	1.70
Accent	Std French	Std French
Smoker	No	Stopped 10 yrs ago

Table 5.5: Description of the misidentified speakers 40 and 1

Additionally, figures 5.10((a) - (d)) and 5.11((a) - (d)) show the spectral profiles and the spectrograms of speaker 40 for the noisy conditions of -5, 0 and 5 dB SNR and for speaker 1 in quiet. The speech segments were made shorter, i.e. around 2 sec, in order to see better the spectral differences between them.

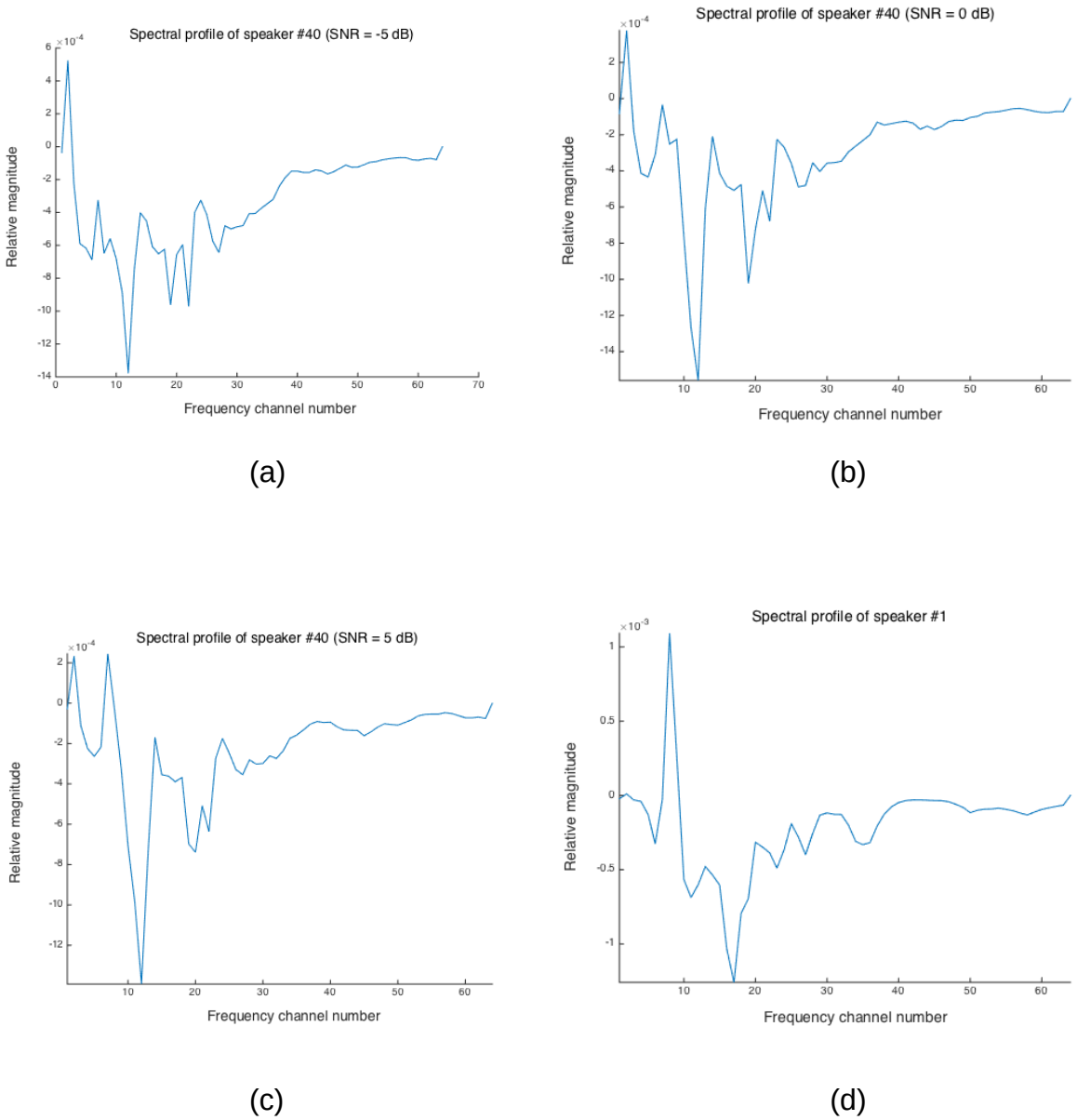


Figure 5.10: Spectral profiles of (a) speaker 40 using speech mixed with noise at -5 dB SNR, (b) speaker 40 using speech mixed with noise at 0 dB SNR, (c) speaker 40 using speech mixed with noise at 5 dB SNR and (d) speaker 1 using clean speech.

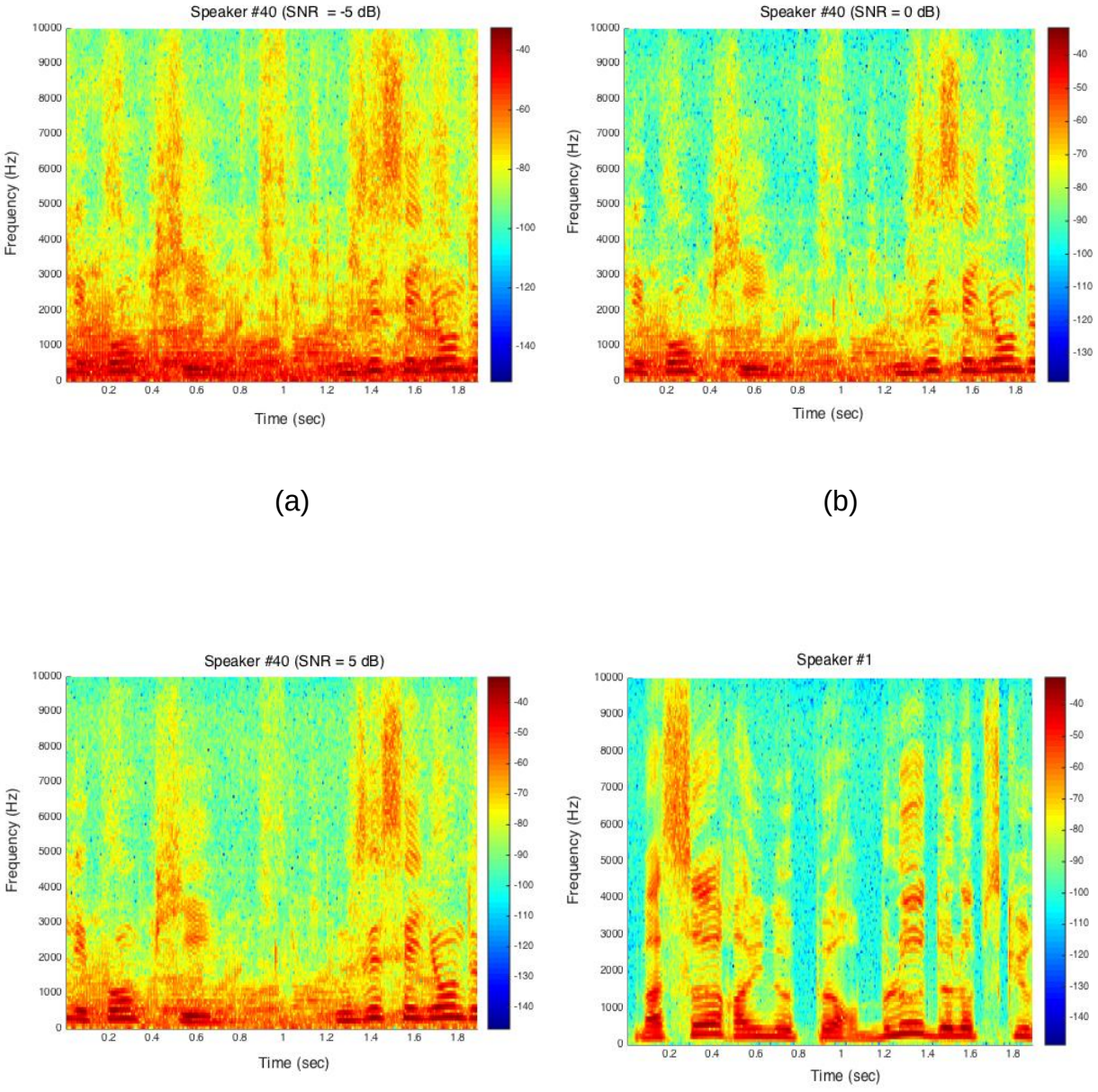


Figure 5.11: Spectrograms of (a) speaker 40 using speech mixed with noise at -5 dB SNR, (b) speaker 40 using speech mixed with noise at 0 dB SNR, (c) speaker 40 using speech mixed with noise at 5 dB SNR and (d) speaker 1 using clean speech

From the spectral profile of speaker 1, it is noticeable that there is a sharp spectral peak located around 200 Hz (frequency channel 8), which corresponds to formant F1. In the corresponding spectrogram, there seems to be a lot of concentrated energy around this frequency. In the same frequency area, there are spectral peaks in the spectral profiles of speaker 40, which are also obvious in her spectrogram.

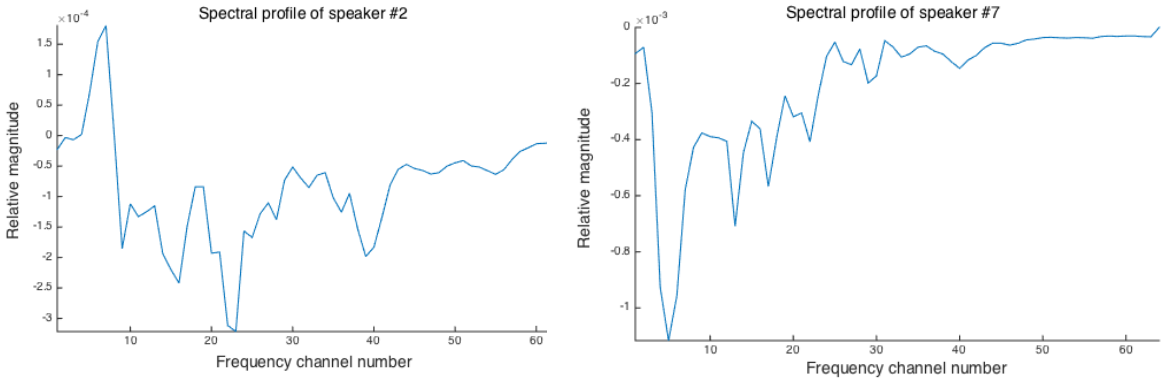
Moreover, for all spectral profiles, it appears that there is a spectral trough that occurs between the 10^{th} and 20^{th} frequency channels. For the spectral profiles of the female, the spectral dip occurs around 265 Hz compared to the male's, which corresponds to approximately 400 Hz. Nevertheless, it remains an important element of the created pattern around that frequency area.

Further, in the case of speaker 40, there are a few peaks, after the trough, placed around the frequency channels 15 (367 Hz), 21 (560 Hz) and 25 (724 Hz). These peaks also exist in the profile of speaker 1 with the difference that the one in channel 15 is before the spectral dip. As mentioned before, this does not necessarily create a salient difference since the patterns are seen at a coarser resolution after the process of downsampling.

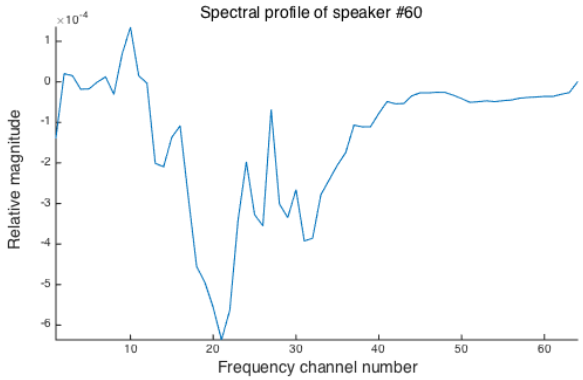
Overall, it seems that these 2 speakers have similiarities in their spectral content up to 1KHz (frequency channel 30). Besides, there seems to be another reason that explains this mismatch between a male and a female. Given that females have, generally, lower height, which means shorter VTL, it is expected to see resonances more in the high frequency range of speech. In the spectrograms of speaker 40 for the 3 SNRs, it is clear that there is energy above 3 KHz. Interestingly, this also happens in the case of speaker 1, where there is activity, in his spectrogram, in frequencies from 3 KHz up to 10 KHz. This can be related to resonances of his vocal tract and give a reason for matching him falsely with a woman. Additionally, it manifests the importance of specifying higher formants since they may be useful for distinguishing speakers of the same gender or explaining mismatches between people of different gender.

Lastly, figure 5.12 ((a) - (c)) shows the spectral profiles of randomly chosen speakers (2, 7 and 60) from the French population that were not confused with the test speaker (40) of this case study. From the figures, it seems reasonable that they have not been misidentified for speaker 40 since their spectral patterns differ.

In particular, from the spectral profile of speaker 2, it is apparent that there are 2 spectral peaks up to the 20^{th} frequency channel (524 Hz). Above that frequency, the peaks are broad and they are distributed in different frequencies. Also, in the case of speaker 7, there is a spectral dip in the 5^{th} frequency channel (144 Hz) and above that, the spectral shape is contrasting to that of speaker 40 for that frequency area. Finally, speaker 60 has a profile that contains one broad peak up to channel 20. Above that frequency, there is a trough and a sharp peak corresponding to the 28^{th} frequency channel (868 Hz). Consequently, the elements of these patterns seem to be quite dissimilar of those of speaker 40 which explains why they have not been mismatched.



(a) (b)



(c)

Figure 5.12: Spectral profiles of (a) speaker 2 using clean speech, (b) speaker 7 using clean speech and (c) speaker 60 using clean speech. These speakers belong in the subset of 60 French talkers. The test speaker 40 is confused with the reference template of speaker 1 (for all SNRs) but she is never confused with the templates of speakers 2, 7 and 60 (for all SNRs)

5.3.3 Evaluation of the system in terms of speaker identification error

In sections 5.3.1 and 5.3.2, the first hypothesis was tested for 2 speaker populations of different size. After obtaining the results from these experiments, it is interesting to see the outcome for the error of the identification score for the proposed system. Figure 5.13 plots the speaker identification (SID) error (%) (i.e. standard error of the mean) of the SAI-based system against the SNR levels (ranging from -5 to 10 dB, at 5 dB intervals) for both speaker populations.

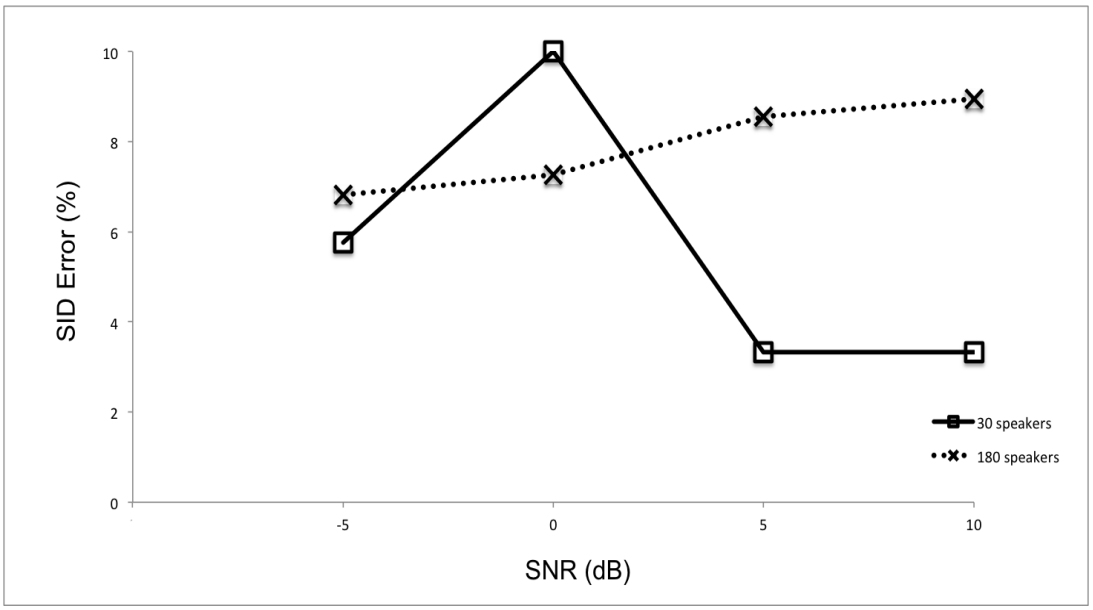


Figure 5.13: Speaker identification (SID) error (%) (standard error of the mean) of the SAI-based system for the corpora of 30 and 180 speakers

As previously mentioned, one of the sources of error in a SID system is the presence of noise. Additionally, as the number of speakers increases, the identification accuracy degrades and it is more possible that errors will occur. From the figure above, it is clear that the size of the speaker data set affects the outcome for the error rate. This is obvious for the SNRs that are equal to -5, 5 and 10 dB where the error is smaller for the 30-speaker database compared to that for the corpus of 180 speakers.

However, for 0 dB SNR, an interesting observation is that the error that corresponds to the corpus of 30 speakers is larger compared to that of the data set of 180 speakers. In this case, the amount of error may have been influenced by various parameters that are involved in the design of a recognition system.

Firstly, for the database of 30 speakers, the number of females and males is not equal. More specifically, the 1st group of 10 speakers consists of more women while the other 2 groups consist of more men. On the other hand, the number of men and women is equal for the database of 180 speakers and for each one of the 3 groups of 60 talkers as well. As explained before, the matching of people of the same gender is more difficult compared to that between men and women. For that reason, if there is not an equal distribution of males and females in a speaker population, it seems that it is more likely that there may be variability in the outcome of the speaker matching and consequently, the SID accuracy. As a result, this may increase the amount of error of the identification score. Additionally, the larger SID error may be explained by the fact that it has been estimated for only 3 subsets of 10 speakers and may decrease if the number of subsets increases.

In order to test the hypothesis that the number of males and females in a database affects the SID performance, we generate 2 databases with balanced and unbalanced distributions of them. Both data sets consist of the same 6 subsets of 10 speakers (60 speakers in total with 30 men and 30 women). In the case of the balanced distribution, each group consists of 5 males and 5 females. In the opposite case, the number of men and women is different for every subset. Table 5.6 shows the results for the SID error between the balanced and unbalanced databases that have been trained using clean speech and tested using noisy data for the specific case of 0 dB SNR.

Database configuration	Balanced distribution of males and females	Unbalanced distribution of males and females
SID error (%)	4.47	6.32

Table 5.6: SID error for 2 databases of the same 60 speakers (6 subsets of 10 speakers) that differ in the way the number of males and females are distributed in each of the 6 groups.

From the results in table 5.6, it appears that the number of subsets of speakers affects the outcome since the identification error has reduced (from 10% to 6.32%) when the number of speaker groups increased (from 3 to 6). Moreover, it seems that our hypothesis that the gender distribution in a database affects the SID performance can also be supported. When the men and women are equally distributed in the 6 different groups, the SID error is smaller compared to that when there is a variation in their distribution. Interestingly, in terms of the results obtained for the 3 groups that consisted of more females, it appears that their accuracy levels remained equal or improved when the number of women decreased and the number of men increased in order to be equally distributed. The opposite result occurred for the other 3 groups that contained more males (their accuracies decreased or remained the same when the number of women increased and the number of men decreased).

Furthermore, another factor is the parameters involved in the classification procedure such as the size of the codebooks. With regard to the codebook size, it is generally considered that the recognition error decreases as the codebook size increases (Kinnunen, 2004). In these experiments, efforts were made to keep the number of codewords constant (equal to 64) in order to focus on the feature sets. Additionally, since the VQ (that has been used in this study) is a non-parametric modeling approach, minimal assumptions are made about the feature distribution. Nevertheless, the

classifier is interrelated with the features and this may have also influenced the variability in the accuracy levels for the case of 0 dB SNR.

Lastly, text-independent speaker identification is an additional source of intra-speaker variability due to the differences in the training and the test speech utterances. As a result, this may have also affected the accuracy results of the 3 groups of 10 talkers.

In conclusion, there are several factors that may impact the SID accuracy levels for different groups of speakers. Since there are various parameters involved in the design of a SID system, it is possible that some or all of them may affect the outcome of the identification process for each group and consequently, the error of the identification score.

5.3.4 Performance for varying parameters of the speech material

In this section, other factors that can influence the SID accuracy of the proposed system are examined. These variables relate to the length of the speech utterances that are used for both the training and testing phases. This set of experiments consists of two parts. Both of them have been conducted using a corpus of 60 talkers (from the EUROM1 database) that consists of 6 different groups of 10 speakers.

5.3.4.1 Performance for varying duration of training speech

In this section, the hypothesis that larger amounts of training data will result in better recognition rates is tested. To achieve that, we vary the duration of the training speech material, through doubling it for each trial, while the test speech length remains constant (Grimaldi, 2008).

The speech length for training varies from 1 to 16 seconds, which is the maximum average duration of this speech corpus. The enrolment session is carried out in quiet conditions and the test speech is mixed with babble noise at 0 dB SNR. Speakers are modeled using the K-means clustering algorithm using 64 centroids. For the speaker matching, the extracted test features are matched against the speaker models in the database with the use of Euclidean distance. Table 5.7 summarizes the details of the corpus and the experiment. The results for the SID accuracy of the SAI – based system against the length of the training speech material are shown in figure 5.14. The SID error is estimated as the standard error of the mean among the 6 subsets of 10 talkers that compose the 60-speaker corpus.

Corpus of 60 speakers	
Language	English
Speakers	60 (30 F+ 30 M)
Speech type	Read speech
Sampling frequency	20 KHz
Training speech type	Clean
Training speech duration (avg)	1, 2, 4, 8, 16 sec
Test speech type	Noisy
Type of noise	Multi-talker babble
Test speech duration(avg)	16 sec

Table 5.7: Summary of the 60-speaker corpus for the experiment with the varying training speech duration

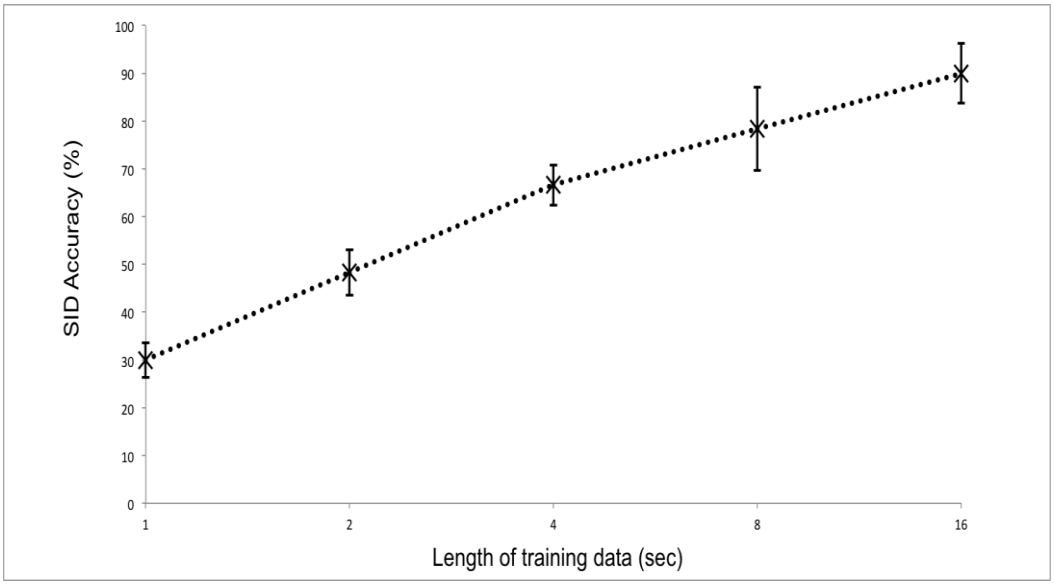


Figure 5.14: Speaker identification (SID) accuracy (%) of the SAI-based system for varying training speech duration. The error bars represent the standard error of the mean (estimated as the error among the levels of SID accuracy of the 6 subsets of 10 speakers)

As expected, it is obvious from the figure above that the identification accuracy gets better as the length of the training speech increases. This happens because, in general, more training data results in obtaining more reliable estimates of the speaker models (Kinnunen, 2004). Despite the fact that the codebook size (64 codewords) is not very large, it is remarkable that the accuracy level for 1 sec of training data is 30% (with 3.65% error), 48.3% (with 3.65% error) for 2 sec and 66.6% (with 4.22% error) for 4 sec. Additionally, it appears that for these cases, there is a constant relationship between the two variables since the accuracy improves up to 18.3% for every doubling of the speech duration. This observation is also valid for the speech durations of 8 and 16 sec, where the recognition rate is 78.3% (with 8.7% error) and 90% (with 6.3% error) correspondingly. For both cases, the performance improves up to 11.7%, which is less than the previous upgrade, but still remains the same.

5.3.4.2 Performance for varying duration of test speech

In this experimental part, the hypothesis about the effect of the length of the test speech material on the identification score is tested. As mentioned previously, more test data are expected to produce better accuracy levels since more information can be extracted to represent the target speakers. In order to test this hypothesis, the training speech duration remains constant whilst the test speech segments range from 1 to 16 seconds. For each trial, their duration is doubled. As before, the test utterances are different from the training ones. In terms of the surrounding conditions, the enrolment session is completed in quiet and the speaker testing is done in the presence of babble noise at 0 dB SNR. Table 5.8 summarizes the attributes of the used corpus and the experiment. The results for the SID accuracy against the test speech duration are plotted in figure 5.15. The SID error is estimated as the standard error of the mean among the 6 subsets of 10 talkers as before.

%	**Corpus of 60 speakers**
Language	English
Speakers	60 (30 F+ 30 M)
Speech type	Read speech
Sampling frequency	20 KHz
Training speech type	Clean
Training speech duration (avg)	16 sec
Test speech type	Noisy
Type of noise	Multi-talker babble
Test speech duration(avg)	1, 2, 4, 8, 16 sec

Table 5.8: Summary of the 60-speaker corpus for the experiment with the varying test speech duration

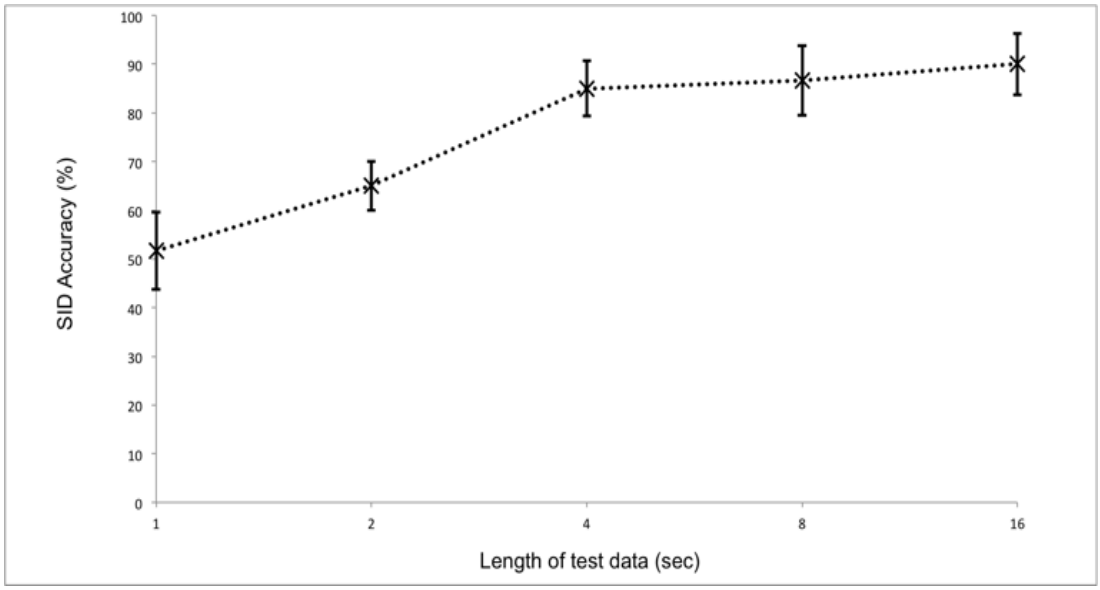


Figure 5.15: Speaker identification (SID) accuracy (%) of the SAI-based system for varying test speech duration. The error bars represent the standard error of the mean (estimated as the error among the levels of SID accuracy of the 6 subsets of 10 speakers)

From the results in the figure above, it is apparent that our hypothesis can be supported. Firstly, it is obvious that the performance becomes better as the test speech length increases. More specifically, when the test speech lasts for 1 and 2 sec, the average identification accuracy is 51.6% (with 7.9% error) and 65% (with 5% error) respectively. The biggest improvement of performance occurs for the speech length equal to 4 sec, which is 20% on average, and the SID accuracy reaches 85% (with 5.6% error). For the length of 8 sec, the accuracy is almost similar to that of 4 sec and equal to 86.6% (with 7% error) while it slightly improves up to 90% (with 6.3% error) for the maximum speech duration of 16 sec. Lastly, it is worth mentioning that after the duration of 4 sec, there seems to be a saturation of the recognition score.

In conclusion, there is evidence that the proposed system seems to be achieving satisfying recognition rates even for a small duration of speech such as 4 sec. The latter is an important finding given that, for specific applications, it is not easy to obtain long test speech segments.

5.4 Discussion

In this chapter, a speaker identification (SID) system has been developed and deployed for a recognition task in a real-world environment. The goal of this study has been to demonstrate the potential utility of the proposed system, which uses features extracted from the SAI, in situations where there is some type of distortion.

Firstly, the AIM-based system has been developed from its initial version that was used for the experiments in chapter 4. Based on the findings of the work in the previous chapter, the feature extraction module was improved through focusing on the informative regions of the auditory image.

More specifically, the discriminative areas, which provide more speaker-specific information, cover the whole filterbank and extend up to a certain point (12.8ms) on the time-interval dimension. Furthermore, the optimal filterbank size, has been chosen to be equal to 64 frequency channels since it helps in obtaining clear patterns for the SID task. After taking into account these details, the box-cutting stage has been modified. Coupled with VQ, for the speaker modeling module, this part is based on a single codebook, which is different from the multi-codebook approach that was used in the initial version of the system.

Accordingly, this change results in a representation that can capture complex structures while avoiding redundancy of data. Hence, the dimensionality of the new SAI feature representation of the signal is significantly reduced and has similar size to the MFCC feature dimensionality. At the same time, it provides computational efficiency given that running times were longer for the initial version of box-cutting used in chapter 4. In this chapter, the improved system achieves satisfactory results with only a fraction of the time needed for extracting features in both the training and testing sessions.

Subsequently, this study investigates how the auditory features perform in a realistic environment. To simulate that, specific parameters were combined, such as text-independent identification and mismatched training and test conditions, in order to make the SID task as demanding as possible. Also, babble noise was chosen since it is known to be one of the most challenging distortions because of its speaker/speech like characteristics (Krishnamurthy et al., 2009). Then, the system was assessed on how it works on different speech corpora created by the same EUROM1 database. This was done because the number of speakers is another parameter that influences the system performance. Additionally, an interesting fact is that the EUROM1 database has not been used before for speaker recognition. Thus, there is not any existing literature about it as for other dabases (TIMIT, NIST SRE) where there is evidence of their performance on different system configurations.

At first, the hypothesis of the performance of the SAI features was tested for a small speaker population (30 speakers) and afterwards, for a larger one (180 speakers). In the first case, the results suggest that the extracted auditory feature vectors lead to much better performance, i.e. higher SID accuracy, compared to the MFCC-based recognition system, especially for low SNRs. For the larger database, the performance profile of the proposed system is similar, i.e. better accuracy levels compared to MFCC features. Though, the general impression is that the SID accuracy decreases as the number of talkers increases, which was expected in the first place. As a conclusion, the hypothesis of perceptual differences relying on lower-dimensional auditory features, which has also been discussed in chapter 4, is supported from the results of these experiments as well.

Furthermore, some interesting mismatches between speakers have been analyzed. In the case of the misidentification between female speakers, the importance of vocal tract and its correlation to speaker size is emphasized.

Given that a wide range of different vocal tract lengths exist in the population, it seems that it can define how humans distinguish or misunderstand speaker size from it. Moreover, the incorrect matching between a male and a female speaker points out the important role of high frequencies and higher formants. This supports what Besacier et al. (2000) have previously suggested about high frequencies being as important, for speaker recognition, as low frequencies. Also, Zhou et al. (2011) have concluded that the high frequency range of speech should not be overlooked because it contains meaningful speaker information.

Moreover, the error of the identification score has been examined in terms of the 2 speaker populations that have been used for the experiments. In the case of 0 dB SNR, the error rate for the 30-speaker data set is higher compared to that of the 180-speaker population which is a noteworthy fact.

From the results of the experiment that has been conducted using 6 groups of 10 talkers (with various and same numbers of males and females), it appears that the dispersion of the speakers in terms of gender can infuence the variability of the identification score and the error rate. Other factors that may possibly have an impact is the parameters involved in the classifier such as the codebook size. Also, the choice of text-independent identification makes the task more difficult since it enhances the intra-speaker variability because of the different training and test speech material. Consequently, it appears that there are several parameters in a SID system design that may affect the end result.

Lastly, in the final set of experiments, the performance of the SAI features was tested for varying parameters related to the duration of the speech material. At first, the training speech length was varied while the test speech remained contant. As expected, the SID accuracy gets better as the size of the training data increases. In the inverse case, where the test speech length is varied, the perfomance improves as the test speech duration increases, but saturates after a specific point, i.e. 4 seconds of

speech. This is a salient asset of the proposed system provided that, in real-world applications, it may not always be feasible to obtain enough test data to identify a specific person.

Besides, it is notable that for the minimum duration (i.e. 1 sec) of training and test speech data, the corresponding SID accuracy levels are on average 30% (with 3.65% error) and 51.6% (with 7.9% error). These recognition rates are very satisfying considering that the minimum probability for correct identification, in both cases, is on average 10% (1 out of 10 speakers for each of the 6 groups).

In addition, in recent literature, there has only been one study by Grimaldi et al. (2008), where they compared different features while varying the training and test speech durations. The best accuracy score was 94% (with 1% error) and it was achieved for 60 sec of training speech and 10 sec of test speech. Yet, these durations are much longer, compared to the results of the SAI features, and obtained for a much smaller corpus of 16 speakers.

Also, another case of high accuracy levels was in a study by Atal (1974) where 98% correct identification (with 2% error) was achieved for 0.5 sec of test speech. Nevertheless, the system conducted text-dependent identification which makes the recognition task less difficult. In the same study, the text-independent task obtained 93% accuracy which is a very satisfying result. Still, the experiments were conducted for a speaker population that consisted of only 10 people.

In conclusion, it seems that the auditory features are very promising to be used for speaker recognition in challenging conditions that could exist in a realistic situation. This has been proven by a number of experiments that show the SAI's robustness in noisy conditions where there is variation of some of the parameters involved in the design of a SID system.

Chapter 6

Conclusions

Feature extraction is a widely used term that describes the conversion of a digital signal into a sequence of numerical descriptors, called feature vectors. In the case of speaker recognition, it can be considered as a process that attempts to capture the essential characteristics of the speaker from one's speech signal with a reduced data rate.

In this thesis, I have investigated if it would be a good choice to use the scheme behind the operation of the human auditory system in understanding sounds for designing a speaker identification (SID) system. Over the course of this work, I have examined aspects of the auditory image model for creating a robust SID system and have obtained a number of results, which suggest that the inspiration from the auditory system is a rational choice to make as well as technically achievable.

6.1 Speaker identification in quiet conditions

Firstly, this study addressed the issue of robust speaker identification from the perspective of using the stabilized auditory image (generated by the AIM) to extract features. The SAI is a model for the early - stage representation of a signal entering the brain.

In order to extract the auditory features from the image, the box-cutting process by Lyon et al. (2010) was used as part of the feature extraction module, i.e. the SAI was broken into overlapping rectangles of different scales. Then, the content of each box contributed independently for speaker matching through creating codebooks for all of the boxes. The purpose was to inspect the relationship between the perception patterns contained in the boxes of the test speakers and the codebooks for those boxes. In order to achieve that, inter- and intra-speaker Euclidean distances were used and confusion matrices among the speakers were created. Through this procedure, the perception of speaker identity was examined and a decision was made about correct or incorrect identification.

Based on this system configuration, our initial hypothesis was that the SAI features can achieve high accuracy levels. Additionally, the system performance was compared with that on MFCCs and we hypothesized that the proposed system would produce similar or better results. Furthermore, we assumed that, based on the identification results, it may be possible to characterize the best possible size of the filterbank for this specific recognition task since the auditory model may not always be optimal and this characterization may further improve the outcome of the system. This may happen because the filterbank size of the SAI is an impactful parameter due to its effect on the created patterns of the image. Also, it is directly related to box–cutting and feature dimensionality which is another primary issue in such systems.

In the case of the small speaker data set (i.e. 30 speakers), the system achieved 100% correct identification for all filterbank sizes. This accuracy level is equal or better compared to results in existing studies by Pruzansky (1963), Glenn et al. (1967), Atal (1974), Reynolds (1994), Grimaldi et al. (2008), Shirali- Shahreza et al. (2011) and Kumar et al. (2011).

Nevertheless, it is not viable to make a comparison with these studies given that the speaker data sets as well as the speaker modeling techiniques (in some of the cases mentioned above) are different from the ones used in the AIM-based system. These parameters affect the SID performance so given their impact on the system design, it is difficult to only focus on comparing the different feature extractors.

Furthermore, for the larger speaker database (i.e. 180 speakers), there was a decrease in performance, which was expected given the increase of the number of speakers. With regard to the filterbank size, the best mean accuracy level was achieved for 64 filters and it was equal to 89.4% (with 2% error). This identification score was a little lower relative to that with the MFCC features (90.5% with 3.1% error). Nevertheless, our hypotheses were supported in terms of the high recognition scores of the SAI features and their similarity with the ones obtained by the baseline system.

In the second part of this experimental set, the hypothesis that the combination of the methods used for the front-end and the classifier can provide us with salient speaker-specific information, which can also help us improve the existing configuration, was investigated. This work was based on the incorporation of the new training strategy in the system for vector quantization that was described in chapter 4, in which individual codebooks were learnt and they were fitted to represent the boxes at specific positions in the auditory image. Except for its use for feature matching in the first experimental part, this approach to the problem of feature extraction allowed us to analyze different parts of the SAI independently. Each codebook was learnt from the data of the complete set of rectangles through the K-means clustering algorithm. This way of yielding codebooks resulted in specifying the most informative regions of the SAI that indicate features that are more speaker-specific.

After the specification of the most discriminative areas of the auditory image, it seems that the boxes converge to the area up to, approximately, 10 ms in terms of the time – interval dimension. With regard to the frequency dimension, the rectangles may cover the whole filterbank or parts of it in the middle and high frequency spans depending on the frequency resolution.

From these results, it seems that the pitch of a speaker is one source of individuality since the first pitch ridge lies in that SAI region. Also, the boxes contain part of the structures that have been created as a result of the resonances of the vocal tract, which is a very important characteristic of the anatomy of a person. Lastly, the first glottal pulse is usually included in that time span and the shape of it can affect the speaker's voice quality. As a result, it appears that this area of SAI can provide information about the characteristics of a speech signal that are speaker-dependent. Additionally, this procedure is important since there were substantial redundancies in the SAI and it was essential to try and find a more dense representation of the signal with reduced data dimensionality. Given that the discerning patterns of the image converge in this specific location, it is possible to select this subset of the extracted features and create a lower-dimensional representation which also supports our hypothesis about it.

Overall, it seems that the benefit of the SAI approach is that the signal is converted into a two-dimensional representation that makes it possible to segregate the glottal pulse rate from the resonance structure of the vocal tract in the SAI. This allowed us to specify the characteristics that make a speaker more discriminative compared to others such as the pitch, lower and higher formant frequencies as well as the shape of the glottal pulse.

On the other hand, one of the limitations is that the auditory model may not be always optimal and it was necessary to characterize certain parameters such as the filterbank size. This needed to be done because the frequency resolution of the image can affect the created patterns by either overly smoothing spectral details or inversely, creating spurious

spectral events.

In addition, the issue of feature dimensionality was another difficulty that occurred in the experimental set with the initial version of the box-cutting used in chapter 4. As the number of speakers increased, the running times became much longer. This resulted in a lack of computational efficiency, which is one of the major topics that SID systems deal with in the case of large speaker databases and long speech utterances. For that reason, it was essential to investigate the possibility of achieving a lower-dimensional SAI feature representation and develop the existing system configuration that was used in the second experimental part.

6.2 Noise-robust speaker identification

In the second experimental set of this study, the aspects of the previous work in this thesis were drawn together in order to improve the existing system configuration. Firstly, the box-cutting was modified so that it includes the informative SAI features that were specified in chapter 4. Particularly, the process was developed in a way that the image is cut into one rectangle that covers the whole filterbank and extends up to a certain point (12.8ms) on the time-interval dimension. For this task, the optimal number of filters was chosen to be equal to 64. The outcome has been a lower–dimensional SAI representation, which retains as much of the interesting information as possible, whilst creating an adequately compact feature vector that is also useful. Also, after the speaker modeling part, each speaker template consisted of a single codebook, which is different from the multi-codebook approach that was used in the initial version of the system. As a result, the new representation makes a good comparison with the MFCC features since their dimensionality is similar and it is computationally efficient given that the features are extracted in much less time for both the training and testing phases.

Furthermore, it was hypothesized that the SAI and the developed version of box–cutting is an effective combination for creating a noise–robust SID system. The improved system was tested in demanding conditions, which simulate a real-world situation, and for 2 different speaker databases. For both cases, the results suggest that the auditory feature vectors lead to much better performance, i.e. higher SID accuracy, compared to the MFCC-based system especially for low SNRs.

In the case of the 30-speaker corpus, the average SID accuracy was 50% (with 5.77% error) for -5 dB SNR, which is much better compared to the results in the study by Pullela et al. (2008) for 31 speakers and the same SNR. Also, in the case of 5 dB SNR, the proposed system achieved 93.33% accuracy (with 3.33% error) while the results by Shao et al. (2007, 2008) for 34 speakers are slightly better. Nevertheless, even though the number of speakers and the noise levels are similar to the ones in our work, it is not feasible to make a legitimate comparison of the used features since the speaker databases, the type of noise as well as the modeling techiniques vary and can influence the outcome for the accuracy levels.

For the database of 180 talkers, it is not simple to make a comparison with existing studies that have been previously described in this work. This happens because there are many parameters involved in a SID system evaluated in noisy conditions, i.e. speaker corpus, size of speaker population, noise level or type of noise (stationary/non-stationary). For instance, for 0 dB SNR, the SAI-based system achieved on average 55% correct identification whilst 16.36% was obtained in the study by Zhao et al. (2013) using factory noise (which simulates another type of real-world environment). Yet, Zhao et al. (2013) conducted the experiments for 330 speakers (from a different corpus) so it is reasonable to obtain a reduced level of SID accuracy. A similar case is the study by Ming et al. (2007) that reached 26.43% accuracy for 10 dB SNR using engine noise while the auditory features obtained on average 81.11% correct identification. Still, Ming et al. (2007) used 630 speakers (from another corpus) which cannot be compared to our speaker data set. The same observation is valid for the studies by Wang et al. (2010) as well.

Overall, one characteristic of the SAI that is key to noise robustness is the representation type of the auditory image, which has the benefit of combining different types of information to a certain extent. The first kind of information is the use of the temporal fine structure at the output of the filterbank. This results in the SAI preserving the fine timing information whereas the MFCCs retain the spectral envelope.

Another important element of the auditory model is that the image contains the relative magnitudes of all frequency bands. At the same time, it includes the specific positions of the frequency areas with high magnitudes that associate to resonances of the vocal tract. This trait of the SAI is one of the reasons behind its robustness for distorted speech, since more noise can be tolerated around the spectral peaks. Consequently, even in the presence of multiple sounds, it is possible that some of the features, which correspond to SAI regions dominated by a specific sound, will often still create an identifiable pattern. This is also related to the stabilization mechanism of the model in which the STI makes the neural patterns that are associated with consecutive cycles of periodic sounds to reinforce each other in the SAI (Patterson, 1992; Walters, 2011).

Moreover, the error rate of the SID accuracy was inspected in terms of the 2 speaker data sets that have been used for the experiments. In the case of the 30-speaker speech corpus, the increased amount of error for 0 dB SNR led us to a noteworthy observation about how speakers are distributed in a speaker population in terms of their gender (more males or females) may have an impact on the outcome of the SID score. Apart from the other parameters that may affect the identification error (i.e. classification parameters, text-independent/ text-dependent identification), this detail should be considered with regard to creating or selecting speaker data sets for future experiments since it could become a limiting factor for achieving satisfying accuracy levels with small error rates.

In the final part of the experiments, we tested two hypotheses about the durations of the training and test speech segments influencing the SID accuracy levels. As expected, the system performance improved as the training speech length increased (while the test speech remained constant). In the opposite case, as the test speech duration increased (and the training speech remained constant) the performance improved as well but reached a point of saturation for 4 seconds of speech.

However, it is notable that the proposed system achieved very satisfying recognition scores for relatively short training and test speech utterances, i.e. 30% and 51.6% average SID accuracy for only 1 second of training and test speech data respectively. This type of hypothesis has been previously studied by Grimaldi et al. (2008) (94% achieved for 60 sec of training speech and 10 sec of test speech) and Atal (1974) (93% for 0.5 sec of test speech). However, their results were obtained for much smaller speaker databases. Also, their experiments were conducted in quiet conditions, which makes the task less challenging, compared to our work which was conducted using noisy test data.

6.3 Conclusions

Based on these experiments, the first conclusion is that, apart from the smoothed spectral envelope, a significant amount of speaker information is included in the spectral details. The voice production system is a complex time-varying system and it is important to capture those fast varying details. This was indicated indirectly by the fact that the filterbank size influenced the created patterns and that affected the extracted features inside the boxes and the correct feature matching. Also, the informative SAI regions contained information about voice source features such as the glottal pulse. Moreover, the noise robustness of the time – interval axis of the SAI shows that the fine structure of speech is the element that carries a large amount of information about speaker individuality.

Furthermore, it appears that the speaker-dependent information may not be distributed only in the low-frequency bands. This was shown by the results from the specification of the discriminative SAI features as well as the mismatches between speakers that have been studied. In particular, the high-frequency bands (over 2KHz) may contain useful information about speaker characteristics, such as the structure of the vocal tract and particularly the VTL.

As a result, the detection of formants in the higher frequency region of speech may work in a complementary way with the ones in the lower frequencies in order to distinguish people of the same gender or alternatively, predict a possible mismatch between people of different gender that have similarities in their formants.

Lastly, another conclusion is that correct identification among females is more challenging compared to that among males. This was indicated by the result of the speaker matching for the 3 groups that consisted of more females in which their SID scores remained equal or improved when the number of women decreased (and the number of men increased in order to be equally distributed). Additionally, the other 3 groups of the population that consisted of more males achieved equal or worse accuracy levels when the number of women increased and the number of men decreased.

This is mainly due to the relatively shorter vocal tract in females and the resulting higher formant frequencies in speech. Usually, these formants are not captured properly because of the low resolution of the used filterbanks in the high frequency region.

6.4 Future work

Since this study is the first attempt of applying the AIM for a speaker recognition task, there is scope for further research in various directions.

Firstly, the SID system that has been described above currently uses features from short windows of speech signals. However, it would be interesting to deal with the longer-term temporal structure of sounds as well. In this manner, it will be possible to see if there is possibility for identifying patterns in a more temporal context.

Furthermore, the study presented here has demonstrated a potential benefit in the use of formants located in the high frequency region of speech. In general, the auditory model architecture is based on the ERB scale, which provides a finer frequency resolution at low frequencies than at high frequencies. In order to capture more spectral details in the high frequency area, it is worth studying whether a linear frequency scale is more advantageous over the ERB scale for the SID task.

Additionally, throughout the experiments of this study, vector quantization (VQ) has been used for the speaker modeling stage. Our main goal was not the optimization of the classifier, but the evaluation of the feature sets. However, it is known that these two stages are interrelated which means that specific modeling techniques may be more suitable for certain features but not appropriate for others. Consequently, future studies should investigate how the auditory features, which have been evaluated in this project, perform when they are combined with another type of classification method. An interesting case is the Gaussian Mixture Model (GMM) method, since it is considered to be the state-of-the-art technique for speaker modeling. Our assumption is that the results that have been obtained with the VQ approach can be generalized to the GMM approach.

Another possible consideration is the testing of the system using different types of conditions. Robustness against noise and distortions is a major issue in speaker recognition research. In this project, the concept was to simulate a real-world situation. For that reason, babble noise of multiple speakers has been used. An alternative idea may include the simulation of other real-life conditions, such as factory, car engine or aircraft noise.

Lastly, the findings of this study show that there is prospective for further research into auditory features related to the voice source such as the glottal pulse. Also, future work in this area should extend to problems in a larger scale. For the SID task, an example is to incorporate larger speaker databases or use longer training and test speech segments than the ones that were used in this study and examine the system performance.

A final thing that should be mentioned is that the parameters that have been found here for the SID task are not guaranteed to be optimal for other recognition tasks. Nevertheless, the feature extraction module that has been developed in this work provides a fine basis for the future study of recognition systems.

List of References

Ashour, G. and Gath I.,(1999). Characterization of speech during imitation. In Proc. 6th European Conference on Speech Communication and Technology (Eurospeech 1999), pp. 1187–1190.

Atal, B. (1972). Automatic speaker recognition based on pitch contours. Journal of the Acoustic Society of America, 52(6), 1687–1697.

Atal B.,(1974). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. Journal of the Acoustic Society of America, 55(6) , 1304–1312.

Besacier, L., Bonastre, J., and Fredouille, C. (June 2000). Localization and selection of speaker-specific information with statistical modeling. Speech Communication, 31,89–106.

Besacier, L. and Bonastre, J.F.(2000). Subband architecture for automatic speaker recognition. Signal Processing 80, 1245–1259.

theBiometricConsortium.http://www.biometrics. org/.

Bleeck, S., Ives, T. & Patterson, R.D. (2004). Aim-mat: The auditory image model in matlab. *Acta Acustica*, **90**, 781–787.

Campbell, J. (1997). Speaker recognition: a tutorial. Proceedings of the IEEE, 85(9), 1437–1462.

Campbell, W. M., Sturim, D. E., and Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. IEEE Signal Processing Letters, 13(5), 308-311.

Carey, M., Parris, E., Lloyd-Thomas, H., and Bennett, S. Robust prosodic features for speaker identification. In Proc. Int. Conf. on Spoken Language Process- ing (ICSLP 1996) (Philadelphia, Pennsylvania, USA, 1996), pp. 1800–1803.

Chan D., Fourcin A., D. Gibbon, B. Granstrom, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld & J. Zeiliger, "EUROM- A Spoken Language Resource for the EU", in Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Speech Technology. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 867-870.

Damper, R., Higgins, J. (2003). Improving speaker identification in noise by subband processing and decision fusion. Pattern Recognition Letters, 24, 2167–2173.

Doddington, G. (1985). Speaker recognition - identifying people by their voices. Proceedings of the IEEE 73, 11,1651–116.

Doddington, G. Speaker recognition based on idiolectal differences between speakers (2001). In Proc. 7th European Conferenceon Speech Communication and Technology (Eurospeech 2001)(Aalborg, Denmark, September 2001), pp. 2521–2524.

Duda, R., Hart, P., and Stork, D.(2000). Pattern Classification, second ed. Wiley Interscience, New York.

Dudley, H. (1939). Remaking speech. J. Acoust. Soc. Am., **11**, 169–177.

Eskelinen-Roenkae,P.Niemi-Laitinen T.(1999). Testing voice quality parameters in speaker recognition. In Proc. The 14th Int. Congress on Phonetic Sciences (ICPhS 1999) (San Francisco, California, USA, 1999), pp. 149–152.

Espy-Wilson, C., Manocha, S., and Vishnubhotla, S. (2006). A new set of features for text-independent speaker identification. In Proc. Interspeech 2006 (ICSLP) (Pittsburgh, Pennsylvania, USA, September 2006), pp. 1475–1478.

Fant, G. (1960). Acoustic Theory of Speech Production. The Hague, Mouton.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. IEEE Transactions on Acoustics, Speech and Signal Processing 29(2), 254–272.

Ganchev T., Fakotakis N., and Kokkinakis G. (2005). "Comparative evaluation of various MFCC implementations on the speaker verification task," Proceedings of the SPECOM-2005, Patras, Greece. Vol. 1, pp. 191–194.

Genoud, D., and Chollet, G. (1998). Speech pre-processing against intentional im- posture in speaker recognition. In Proc. Int. Conf. on Spoken Language Processing (ICSLP 1998).

Gersho, A. & Gray, R.M. (1992). Vector quantization and signal compression. Kluwer Academic Publishers, Norwell, MA, USA.

Glasberg B. R., Moore B. C. J. (1990). Derivation of auditory filter shapes from notched noise data. Hear. Res. **47,**103–138.

Glasberg B.R., Moore B.C.J. (2002). A model of loudness applicable to time-varying sounds. J. Audio Eng. Soc. **50**,331–342.

Glenn J.W., and Kleiner N. (1967). Speaker identification based on nasal phonation, J. Acoust. Soc. Amer. , vol. 43(2), pp. 368-72.

Godfrey J., Graff D., Martin A. (1994) Public Databases for Speaker Recognition and Verification, ISCA Archive, pp.39–42.

González, J. (2004). Formant frequencies and body size of speaker: A weak relationship in adult humans, J. Phonetics **32**, pp. 277–287.

Hermansky, H. (1990). Perceptual linear prediction (PLP) analysis for speech. Journal of the Acoustic Society of America, 87, pp.1738–1752.

Hermansky, H. (1994). RASTA processing of speech. IEEE Trans. on Speech and Audio Processing, 2(4), 578–589.

https://www.nlm.nih.gov/medlineplus/ency/article/002955.htm

http://www.phon.ox.ac.uk/~jcoleman/phonation.html

Huang, X., Acero, A., and Hon, H.-W. (2001). Spoken Language Processing: a Guide to Theory, Algorithm, and System Development. Prentice-Hall, New Jersey.

Irino, T., Patterson, R.D. and Method, I. (1996). A time-domain, level-dependent auditory filter: The gammachirp. Acoustical Society of America, 101(1), pp.412–419.

Irino, T. & Patterson, R.D. (2006°). A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE transactions on audio, speech, and language processing, 14(6), pp.2222–2232.

Irino, T. & Patterson, R.D. (2006b). A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE transactions on audio, speech, and language processing, 14(6), pp.2222–2232.

Irino, T. & Patterson, R.D. (2002). Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. Speech Communication, 36(3-4), pp.181–203.

Ives, D.T., Smith, D.R.R. & Patterson, R.D. (2005). Discrimination of speaker size from syllable phrases. J. Acoust. Soc. Am., **118**, 3816–3822.

Jain, A., Duin, R., and Mao, J. (2000). Statistical pattern recognition: A review. IEEE Trans. on Pattern Analysis and Machine Intelligence, 22(1), 4–37.

Kekre, H.B. & Sarode, T.K. (2009). Vector Quantized Codebook Optimization using K-Means. International Journal on Computer Science and Engineering, 1(3), pp.283–290.

Kesarkar, M. P. (2003). Feature extraction for speech recogniton, M.Tech. Credit Seminar Report, Electronic Systems Group, EE. Dept, IIT Bombay.

Kinnunen, T. (2004). Spectral Features for Automatic Text- Independent Speaker Recognition. Licentiate's thesis, Universityof Joensuu, Department of Computer Science, Joensuu,Finland.

Kinnunen, T., and Gonzalez-Hautamaki, R. (2005). Long-term f0 modeling for text-independent speaker recognition. In Proc. 10th International Conference Speech and Computer(SPECOM'2005) (Patras, Greece, October 2005), pp. 567–570.

Kinnunen, T., Koh, C., Wang, L., Li, H., and Chng, E. (2006). Temporal discrete cosine transform: Towards longer term temporal features for speaker verification. In Proc. 5th Int. Symposium on Chinese Spoken Language Processing (ISCSLP 2006) (Singapore, December 2006), pp. 547–558.

Kinnunen, T., Saastamoinen, J., Hautamaki, V., Vinni, M., and Franti, P. (2009). Comparative evaluation of maximum a Posteriori vector quantization and Gaussian mixture models in speaker verification. Pattern Recognition Letters 30, 4 (March 2009).

Kittler, J., and Nixon, M. (2003). Lecture Notes in Computer Science, Eds. 4th Internation Conference on Audio and Video Based Biometric Person Authentication (AVBPA 2003). Springer-Verlag, Berlin.

Krishnamurthy N. and Hansen J.H. L. (2009). Babble Noise: Modeling, Analysis, and Applications, IEEE transactions on audio, speech, and language processing, 17(7), pp.1394-1406.

Krumbholz, K., Patterson, R. D., & Pressnitzer, D. (2000). The lower limit of pitch as determined by rate discrimination. Journal of the Acoustical Society of America, 108, 1170–1180.

Kumar Ch. S. and Rao P. (2011). Design Of An Automatic Speaker Recognition System Using MFCC , Vector Quantization And LBG Algorithm. , International Journal on Computer Science and Engineering (IJCSE), 3(8), pp.2942–2954.

Kunzel, H. (1989). How well does average fundamental frequency correlate with speaker height and weight?, Phonetica 46, 117–125.

Laver, J. (1994). Principles of Phonetics. Cambridge University Press, Cambridge.

Lin L., and. Wang S. (2006). A Kernel method for speaker recognition with little data, in Int. Conf. signal Process. Budapest, Hungary 2006.

Linde, Y., Buzo, A., and Gray, R. (1980). An algorithm for vector quantizer design. IEEE Transactions on Communications, 28(1), 84–95.

Linguistic Data Consortium: https://catalog.ldc.upenn.edu/LDC93S1

Lu, X., and Dang, J. (2007). An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech Communication, 50(4), 312–322.

Lyon, R.F., Rehn, M., Bengio, S., Walters, T.C. & Chechik, G. (2010). Sound retrieval and ranking using sparse auditory representations. Neural computation, 22(9), pp.2390–416.

Lyon, R.F., Ponte, J. and Chechik, G. (2011). Sparse coding of auditory features for machine hearing in interference. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5876–5879.

Lyon, R.F., Shera, C. and Olson, E.S. (2011). A Pole–Zero Filter Cascade Provides Good Fits to Human Masking Data and to Basilar Membrane and Neural Data, 224, pp.224–230.

Madikeri, S.R. & Murthy, H. (2011). Mel Filter Bank energy-based Slope feature and its application to speaker recognition. 2011 National Conference on Communications (NCC), pp.1–4.

Maesa, A. (2012). Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models. Journal of Information Security, 03(04), pp.335–340.

Manning C., Raghavan P., Schutze H.,(2009). An Introduction to Information Retrieval,Online edition, Cambridge.

Martin, A., and Przybocki, M. Speaker recognition in a multi-speaker environment.(2001). In Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001) (Aalborg, Denmark, 2001), pp. 787–790.

Mary L., Rao K.S., Gangashetty S.V., and Yegnanarayana B.,(2004). Neural network models for capturing duration and intonation knowledge for language and speaker identification, in Proc. Int. Conf. Cognitive Neural Systems, Boston, Massachusetts.

May, T., van de Par, S., and Kohlrausch, A. (2012b). Noise-robust speaker recognition combining missing data techniques and universal background modeling. IEEE Transactions on Audio, Speech, and Language Processing, 20(1), 108-121.

Meddis R., Hewitt M. J. (1991) .Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I. Pitch identification. J. Acoust. Soc. Am. 89, 2866–2882.

Meddis, R. & O'Mard, L., (1997). A unitary model of pitch perception. The Journal of the Acoustical Society of America, 102(3), pp.1811–20.

Ming, J., Hazen, T. J., Glass, J. R., and Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. IEEE Transactions on Audio, Speech, and Language Processing, 15(5), 1711-1723.

Mokhtari, P.(1998). An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy. PhD thesis, School of Computer Science, University of New South Wales, Canberra, Australia.

Monaghan, J.J., Feldbauer, C., Walters, T.C. & Patterson, R.D. (2008). Low-dimensional, auditory feature vectors that improve vocal-tract-length normaliz- ation in automatic speech recognition. J. Acoust. Soc. Am., 123, 3066.

Moore, B.C.J. (2003). An Introduction to the psychology of hearing. Academic Press, San Diego, 5th edn.

Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Journal of Computing, ,*2*(3), 138–143.

Myers L., (2004). "An Exploration of Voice Biometrics".

O'Shaughnessy D. (1987). Speech communication: human and machine. Addison-Wesley. p. 150. ISBN 978-0-201-16520-3.

Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C. & Allerhand, M. (1992). Complex sounds and auditory images. In Y.C.L. De- many & K. Horner, eds., Auditory Physiology and P erception, 429–446, Pergamon Press, Oxford.

Patterson, R.D. (1994b). The sound of a sinusoid: Time-interval models. J. Acoust. Soc. Am., 96, 1419–1428.

Patterson R. D.. Allerhand. (1995). Time-domain modelling of peripheral auditory processing. Acoustical Society of America, 98(4).

Patterson R. D., J. Holdsworth, (1995).A functional model of neural activity patterns and auditory images. – In: Advances in Speech, Hearing and Language Processing. W. A. Ainsworth (ed.). JAI, London,Vol. 3, Part B, 554– 562.

Patterson, R.D. & Irino, T. (1998). Modeling temporal asymmetry in the auditory system. J. Acoust. Soc. Am., 104, 2967–2979.

Patterson, R.D., Walters, T.C., Monaghan, J., Feldbauer, C. & Irino, T. (2010). Auditory speech processing for scale-shift covariance and its evaluation in automatic speech recognition. In IEEE International Symposium on Circuits and Systems.

Peskin B., J. Navratil, J. Abramson, D. Jones, D. Klusacek, D.A. Reynolds, and B. Xiang,(2003). Using prosodic and conversational features for high-performance speaker recognition, in Int. Conf. Acoust., Speech, Signal Process. , vol. IV, pp. 784-7.

Plumpe, M., Quatieri, T., and Reynolds, D. (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Trans. on Speech and Audio Processing, 7(5), 569–586.

Prasanna S.R.M., Gupta C.S., and Yegnanarayana B.,(2006). Extraction of speaker-specific excitation information from linear prediction residual of speech, Speech Communication , vol. 48, pp. 1243-61.

Pressnitzer D., Patterson R. D., Krumbholz K. (2001). The lower limit of melodic pitch. J. Acoust. Soc. Am., 109 ,2074–2084.

Proakis, J., and Manolakis, D.(1992). Digital Signal Prosessing. Principles, Algo- rithms and Applications, second ed. Macmillan Publishing Company, New York.

Pruzansky S., (1963). Pattern matching procedure for automatic talker recognition", J. Acoust. Soc. Am., 35 (3),pp.354-8.

Pullella, D., Kuhne, M., and Togneri, R. (2008). Robust speaker identification using combined feature selection and missing data recognition. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4833- 4836.

Rabiner, L., and Juang, B.H.(1993). Fundamentals of Speech Recognition. Prentice Hall, Englewood Cliffs, New Jersey.

Rehn, M., Lyon, R.F., Bengio, S., Walters, T.C. & Chechik, G. (2009). Sound ranking using auditory sparse-code representations. In International Conference on Machine L earning 2009,Workshop: Sparse Methods for Music Audio, Montréal, Canada.

Rendall, D., Vokey, J. R., Nemeth, C., and Ney, C. (2005). Reliable but weak voice-formant cues to body size in men but not women, J. Acoust. Soc. Am., 117, 2372.

Reynolds, D. a., Quatieri, T.F. & Dunn, R.B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3), pp.19–41.

Reynolds, D., and Rose, R. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans. on Speech and Audio Processing 3 (January 1995), 72–83.

Reynolds, D. (2002). An overview of automatic speaker recognition technology. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2002) (Orlando, Florida, USA, 2002), pp. 4072–4075.

Reynolds, D.A. (1995). Automatic Speaker Recognition Using Gaussian Mixture Speaker Models, The Lincoln Laboratory Journal, 8 (2).

Reynolds, D. (1995). Speaker identification and verification using Gaussian mixture speaker models. Speech Communication 17(August 1995), 91–108.

Reynolds D.A.,(1994). Experimental evaluation of features for robust speaker identification, IEEE Trans. Speech Audio Process, 2(4), pp. 639-43.

Rhode, W.S. & Robles, L. (1974). Evidence from Mössbauer experiments for non- linear vibration in the cochlea. *J. Acoust. Soc. Am.*, 55, 588–596.

Rose, P. (2002). Forensic Speaker Identification. Taylor & Francis, London.

Rosenberg A.E., and Sambur M.R.(1975). New techniques for automatic speaker verification,IEEE Trans. Acoust., Speech, Signal Process. , vol. ASSP-23(2), pp. 169-76.

Sahidullah, M. and Saha, G., (2012). Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition. Speech Communication, 54(4), pp.543–565.

Shackleton, T.M. and Carlyon, R.P., (1994). The role of resolved and unresolved harmonics in pitch perception and frequency modulation discrimination. The Journal of the Acoustical Society of America, 95(6), pp.3529–40.

Shao, Y., Srinivasan, S., and Wang, D. (2007). Incorporating auditory feature uncertainties in robust speaker identification. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. IV- 277-IV-280.

Shao, Y., and Wang, D. (2008). Robust speaker identification using auditory features and computational auditory scene analysis. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1589-1592.

Shirali-Shahreza, M.H., and Shirali-Shahreza, S.,(2011). Effect of MFCC normalization on Vector Quantization based speaker identification. Proceedings of 10th IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2010), Luxor, Egypt, 15-18, pp. 250-253.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. (2005). Modeling prosodic feature sequences for speaker recognition. Speech Communication, 46 (3-4), 455–472.

Singh, G., Panda, A., Bhattacharyya, S., and Srikanthan, T.(2003). Vector quantization techniques for GMM based speaker verification. In Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2003) (Hong Kong, 2003).

Smith, D.R.R., Patterson, R.D., Turner, R.E., Kawahara, H. & Irino, T. (2005). The processing and perception of size information in speech sounds. J. Acoust. Soc. Am., 117, 305–318.

Smith, D.R.R., Patterson, R.D.(2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age.J. Acoust. Soc. Am., 118, 3177–3186.

Smith, D.R.R., Walters, T.C. & Patterson, R.D. (2007). Discrimination of speaker sex and size when glottal-pulse rate and vocal-tract length are controlled. J. Acoust. Soc. Am., 122, 3628–3639.

Soong, F. K., Rosenberg, A. E., Juang, B. H., and Rabiner, L. R. (1987). Report: A vector quantization approach to speaker recognition. AT&T Technical Journal, 66(2), 14-26.

Stevens, S., Volkman J., and Newman E. (1937). A scale for the measurement of the psychological magnitude pitch. Journal of the Acoustical Society of America, 8 (3), 185–190.

Story, B. (2002). An overview of the physiology, physics and modeling of the sound source for vowels. Acoustical Science and Technology, 23(4),195–206.

Theodoridis S., Koutroumbas K. (2009). Pattern Recognition, 4th Edition, Academic Press, Elsevier.

Thevenaz P., and Hugli H. (1995). Usefulness of the LPC-residue in text-independent speaker verification, Speech Communication , vol. 17, pp. 145-57.

Turner, R.E., Walters, T.C. & Patterson, R.D. (2004). Estimating vocal tract length from formant frequency data using a physical model and a latent variable factor analysis. In British Society of Audiology Short P apers M eeting on Experimental S tudies of H earing and Deafness, 61, UCL London.

Vimala, S. (2011). Convergence Analysis of Codebook Generation Techniques for Vector Quantization using K-Means Clustering Technique., International Journal of Computer Applications, 21(8), pp.16–23.

Vuuren, S. Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch.(1996). In Proc. Int. Conf. on Spoken Language Processing (ICSLP 1996) (Philadelphia, Pennsylvania, USA, 1996), pp. 1788–1791.

Walters T.C. (2011). Auditory-based processing of communication sounds, University of Cambridge (PhD thesis)

Wang D.L. and Brown G.J. (2006). Ed., Computational Auditory Scene Analysis: Principles, Algorithms, and Applications. Hoboken, NJ: Wiley-IEEE Press.

Wang L., Minami K., Yamamoto K., Nakagawa S. (2010). Speaker Recognition by Combining MFCC and Phase Information in Noisy Conditions, IEICE TRANS. INF. & SYST., 93(9) .

http://www.internationalphoneticassociation.org

Yegnanarayana B., and Kishore S.P.(2002). AANN: An alternative to GMM for pattern recognition, Neural Networks , vol. 15, pp. 459-69.

Zhao, X., Shao, Y., and Wang, D. (2011). Robust speaker identification using a CASA front-end. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 5468-5471.

Zhao, X., Shao, Y., and Wang, D. (2012). CASA-based robust speaker identification. IEEE Transactions on Audio, Speech, and Language Processing, 20(5), 1608- 1616.

Zheng, F., Zhang, G. & Song, Z. (2001). Comparison of different implementations of MFCC. Journal of Computer Science and Technology, 16(6), pp.582–589.

Zhou, X., Garcia-Romero, D., and Duraiswami, R., Espy-Wilson, C., and Shamma, S. (2011). Linear versus mel-frequency cepstral coefficients for speaker recognition. In: Proceedings of IEEE Workshop on ASRU, pp. 559-5