Open-set Speaker Identification

Rawande Karadaghi

University of Hertfordshire

A thesis submitted in partial fulfilment of the requirement of the University of Hertfordshire for the degree of Doctor of Philosophy

The Program of research was carried out in the school of Engineering and Technology,
University of Hertfordshire, Hatfield, Hertfordshire AL10 9AB, United Kingdom

August 2017

Abstract

This study is motivated by the growing need for effective extraction of intelligence and evidence from audio recordings in the fight against crime, a need made ever more apparent with the recent expansion of criminal and terrorist organisations. The main focus is to enhance open-set speaker identification process within the speaker identification systems, which are affected by noisy audio data obtained under uncontrolled environments such as in the street, in restaurants or other places of businesses. Consequently, two investigations are initially carried out including the effects of environmental noise on the accuracy of open-set speaker recognition, which thoroughly cover relevant conditions in the considered application areas, such as variable training data length, background noise and real world noise, and the effects of short and varied duration reference data in open-set speaker recognition.

The investigations led to a novel method termed "vowel boosting" to enhance the reliability in speaker identification when operating with varied duration speech data under uncontrolled conditions. Vowels naturally contain more speaker specific information. Therefore, by emphasising this natural phenomenon in speech data, it enables better identification performance. The traditional state-of-the-art GMM-UBMs and i-vectors are used to evaluate "vowel boosting". The proposed approach boosts the impact of the vowels on the speaker scores, which improves the recognition accuracy for the specific case of open-set identification with short and varied duration of speech material.

Keywords: Speaker recognition; vowel boosting; open-set speaker identification;

Acknowledgments

Firstly, I would like to express my sincere gratitude to my principle supervisor, Prof. Aladdin Ariyaeeinia, for his continuous support, and for his ongoing enthusiasm, passion and determination to my research topic, which has acted as a paramount catalyst to my learning. Similarly, thanks to Prof. Ariyaeeinia for his guidance, wisdom and support.

Furthermore, a special thanks to Dr. Heinz Hertlein, my second supervisor and dear friend, for his immense patience, invaluable support, guidance and encouragement throughout my studies. Another special thanks goes to my equally respected supervisor Dr. Zoe Jeffrey, for her generous determination to assist me throughout my thesis journey. Thank you very much.

Lastly, but by no means the least, my immense gratitude goes to my family for their endless help, support and love during the course of this work, particularly my father, who has always motivated, supported and encouraged me to do better.

Table of Contents

List of figures

List of tables

List of abbreviation

OSTI-SI	Open-set text independent speaker identification
VB	Vowel Boosting
OS-SI	Open-set speaker identification
OS-IE	Open-set identification errors
SNR	Signal to noise ratios
GMM	Gaussian mixture model
UBM	Universal background model
VAD	Voice active detection
SCD	Speaker change detection
T-norm	Test normalisation
Z-norm	Zero normalisation
SR	Speaker recognition
ICCST	International carnahan conference
ML	Miss labelling
FR	False rejection
SVM	Super vector machines
MMI	Maximum mutual information
MAP	Maximum a posterior
AER	Accumulated error rate
AN	Auditory nerve
MFCC	Mel-frequency cepstrum coefficients
FDLP	Frequency domain linear prediction

GFCC	Frequency cepstral coefficients
PCA	Principal component analysis
HLDA	Heteroscedastic linear discriminant analysis
LDA	Linear discriminant analysis
WCCN	Within class covariance normalisation
VTS	Vector taylor service
JFA	Joint factor analysis
GPLDA	Gaussian probability linear discriminant analysis
LPC	linear predictive coding
LSP	line spectral pairs
DFT	Discrete fourier transform
DCT	Discrete cosine transform
SAD	Speech active detection
DD	Decision direct
LPCC	Linear predictive coding-based cepstrum
FFT	Fast fourier transform
LP	Linear prediction
L-D	Levinson-durbin
DTW	Dynamic time wrapping
VQ	Vector quantization
HMM	Hidden markov model
PDF	Probability density functions
MM	Mixture model

ML	Maximum likelihood
EM	Expectation maximisation
NIST	National institute of standards and technology
JFA	Joint factor analysis
CSS	Cosine similarity score
TV	Total variability
TVM	Total variability matrix
FA	False acceptance
CSS	Cosin similarity score
SRE	Speaker recognition evaluation
WBS	Weighted bilateral scoring
i-vector	Total variability
MLP	Multi layer perception
PLP	Perceptual linear predictive

CHAPTER ONE -- Introduction

1. Introduction

Individuals have access to a wide range of mediums to communicate with one another from different parts of the world. One of the most prominent mediums of communication is via the voice: telecommunications and video are common mediums of communication that are now available everywhere, with the proliferation of cell phones and tablets which use telecommunications-based internet signals. Anyone can post coded communications activities on the internet for exposure to global audiences as well. The interest here is to examine the identification of individuals through their voice data, using the science and techniques of speaker recognition biometrics.

To be used reliably for identification, any behavioural characteristic used in the speaker identification process require that some extended observation of a subject occur, or that the subject be an otherwise known quantity to the observer and the system [1]. The latter implies the subject consents to create training material for the voice biometric system (controlled situation). The Voice biometrics is a term that is used to describe several technologies which can look for, identify, or authenticate unique speech patterns belonging to an individual (See Appendix A). Unfortunately, not all subjects for which speaker recognition biometrics will be applied will be able to, or may be willing to voluntarily provide training utterances. This situation is considered uncontrolled, and is typical of the conditions under which surveillance activities are carried out. In all cases of controlled or uncontrolled conditions, the voice biometric identification system compares the subject to the many records in its database, to first attempt to determine a set of likely matches, then narrow those possible cases down to a few, so that it may then refine and process the data to come to a decision regarding the identity of the subject. This decision may be positive, in the case where the voice biometric sample can be matched to a subject utterance held in the database, where the processed data lies within the bounds of a statistically relevant match threshold, or the decision may be negative, if the statistical bounds cannot be met by comparison with existing database records, or if the system determines positively that no possible matches exist. In order to find that statistically relevant threshold, three things must occur [2];

- First, a reference database must be built. Reference models must be generated, processed, categorised, and stored in the model database.
  - For example: this could be accomplished by using a test group to provide utterances in various kinds of controlled sample environments.
  - An example of this would be how the voice and speech recognition databases underlying home digital assistants like Siri, Amazon Echo, OK Google, and Cortana were developed. A broad, diverse group of speakers was given a set of phrases to use with the system, and encouraged to vary emotional content, volume, and pitch, over a period of a month, on a daily basis, to 'train' the algorithm.
- Second, a subset of the reference models must be chosen to be compared against the larger set of reference models to test the algorithm for accuracy, precision, and robustness.
  - These will generate some real responses in some cases, and imposter responses in others.
  - Once enough of both are obtained, they are used to calculate the threshold for the database, above which the identification is accepted as reliable and accurate.
- Third, after a threshold boundary is calculated, the voice biometric system must be tested rigorously against selected model templates in the database to verify and quantify accuracy.

Whatever method to establish identity is used, the overall performance of a given voice biometric system is measured in terms of its accuracy, speed, and database capabilities. Beyond this, cost and ease-of-use are critical factors which impact the systems utility for recognition, authentication, or both.

Whichever set of traits is used to build the system, the selection should be such that the combination of traits to be analysed considers a number of factors [3]. Are the chosen traits:

- Universal? Every person should possess the trait of interest.
- Unique? The variation of the trait from one individual to the next should be distinctive enough that the voice biometric system can tie it to one individual.
- Permanent? Does the trait vary a lot or a little over time? This makes a difference in how well it can be used to tag an individual.
  - If there is a high degree of variability, is there some way this variability can be minimized or removed from the system so that identification can be made?
- Measureable? How easily and how well can the trait(s) and characteristics be acquired? Can they be extracted and processed? How sound are the results obtained from the processing?
- Easily Processed? How well do the characteristics work for identification or authentication, and is the processing method accurate, fast, and robust?
- Acceptable? Is the identifying technology accepted well enough that analysts and the public both are willing to let it capture and assess their identity?
- Precise? Are the results that the system produces repeatable so that the same person is identified every time their unique set of traits is enrolled in the system?
- Circumventable? Can someone imitate this trait and get around the system?

o If someone does imitate the trait of interest, is the system sensitive enough to catch it with a high degree of probability?

For example, voice biometric systems are used to achieve the following goals:

- To identify, verify, or authenticate a person,
- To protect a system from unethical or fraudulent handling,
- To prevent identity theft or other crime,
- To control access to sensitive information or areas,
- To conduct surveillance

1.1. Speaker recognition and biometric authentication

Speaker recognition combines both physiological and behavioural modalities to identify and categorize a subject, to identify who is speaking rather than identify what is being said. The goal of speaker recognition is to identify whomever is speaking; this may or may not mean and include speaker authentication (is there positive identification of subject X?), speech recognition, or the recognition and identification of multiple speakers (recognition and/or authentication of whomever is speaking now), and less often, identification of their emotional state [4]. Often, in practice, speaker identification precedes speaker verification. The physiological components of speech recognition may include the shape, size and health of a person's vocal cords, as well as the physiological characteristics and contours of their lips and teeth, as well as nasal and mouth cavities. The behavioural components of speech recognition may include tone, timbre, accent, pitch, loudness, pace of talking, the subject's emotional state, and noisiness (rattling, whistling, non-vocal breath sounds, body, or environmental sounds) while speaking. The techniques on which speaker recognition are based originated, partly, in the field of psychoacoustic analysis (sound perception) for the study of both sound and sound perception in both music and speech[4].

Speaker recognition has been an active area of study and development for decades [5]. For instance, Davis, Balashek, and Biddulph built a system at Bell Laboratories in 1952 for the recognition of isolated digits for single speakers [6]; while Forgie and Forgie devised a speaker-independent system to recognize 10 vowels in a /b/-vowel-/t/ format in 1959 [7]. Since the 1950s, the technology has increasingly improved in precision and sophistication with time, to achieve the end-goal of being able to identify or authenticate individual identity using voice biometrics information. Identification and authentication are tandem activities, where in practice it is customary to first identify, then authenticate a speaker. Identification permits the narrowing of the number of likely candidates, whereas further analysis to complete authentication completes a positive match. Many of the same criteria are used for each: what differs is the number and precision of the variables used. In the end, voice biometrics is best defined as a group of measurable physiological or behavioural characteristics that can be used to verify the identity of an individual based on a sample utterance. Therefore, speaker recognition may be generally defined as the identification of individuals of concern using their statistically unique voice biometric data. This technology has already been widely applied for authentication, personal security, financial transactions (banking, for example), restricted access to secure locations and information, and as a means of protecting personal information and assets. For instance, speaker recognition has been used in the penal system to control and monitor phone privileges for inmates, as well as for the identification and verification of juvenile inmates, parolees, and persons under house arrest [8]. However, the potential for wider application of this technology in domestic and international security is an area that has been less explored.

Voice communication technology plays a vital role in the day to day endeavours of today's society, and offers a variety of mediums with which people may communicate all over the world. Voice communications occur via phone, internet, in person, or through video. The near-universal accessibility of these mediums has led to their increasing use by criminal organizations for the purposes of organising groups or cells to carry out illegal activities, for recruiting new members, holding rallies and meetings, passing along instructions, gathering funds to support their activities, or planning various kinds of attacks [9]. These potentially criminal voice communications may be intercepted and analysed, thus speaker recognition can serve as a significant tool to use to identify criminal individuals, and their associates through their voice data, hence increasing the likelihood of protecting societies from such individuals [9, 10]. As stated earlier, this area has been less explored, primarily because previous research on speaker recognition has depended on individuals being cooperative – participating voluntarily in research to provide voice samples with which novel voice data methods can be compared. In contrast, a distinctive feature of the security scenario under which surveillance of criminal activities might occur, is that one may assume participants will not be cooperative in providing usable voice samples for analysis and incorporation into a speaker recognition database, therefore, any voice data retrieved will of necessity be highly variable in length and noise levels, both phonetically and acoustically. Should these challenges be overcome, the potential for speaker recognition for practical application in domestic and international security efforts is considerable.

1.2. Open-set speaker identification definition

Speaker identification is one main subclass of the larger field of speaker recognition, and may be described as determining the correct identity of a specific speaker in a selected test utterance, obtained from a pre-registered population. When the identification process includes the option of declaring that the analysed test utterance does not belong to any of the registered speakers, then it is specifically referred to as open-set speaker identification (OS-SI). Moreover, if the utterances used for training and testing are not constrained to be of the same linguistic content, the process is called open-set, text-independent speaker identification. This is the most challenging subclass of speaker recognition analysis [11], but is one which has a wide range of applications in areas such as audio indexation, surveillance, and screening.

As mentioned earlier, the research into speaker identification over the past several years has resulted in considerable advances in the field and the establishment of well-defined approaches which may be further expanded and improved. These approaches are based on firm, well-established pattern matching principles, and incorporate capabilities for dealing with variation in speech characteristics such as the ones mentioned in the last section 2 [3, 12]. However, to date, there has been limited attention to the challenging problems posed by operating under uncontrolled conditions. A major issue under such conditions is the earlier-mentioned lack of voluntary user cooperation leading to uncontrolled conditions for analysis of voice samples. Attempting to perform speaker identification without the user's cooperation presents many challenges which will be discussed in the next section.

1.3. Challenges

The main challenges to consider with speaker recognition are efficacy and accuracy. Both challenges positively depend heavily upon the length and quality of the audio files that are obtained. While in an ideal world training samples for speaker recognition could be obtained voluntarily, in real situations, the speaker recognition analyst sometimes can have little to no control over where, when, how long, or with what clarity a subject speaks, or nor can s/he control or specify the range and/or duration of the speaker's emotional state. Hence, when considering speech recognition for security applications such as covert surveillance, these challenges are exacerbated because, in most cases, voice data are nearly always obtained without the user's cooperation or permission, and the voice data that are obtained will most certainly be highly variable, or compromised in other ways.

When obtaining data under field conditions, for example, where security surveillance is most likely to be carried out, the most obvious challenge is the issue of background noise. As voice data are recorded in uncontrolled conditions, one must assume there may be significant noise and disturbances captured in the recordings obtained, which will make the process of parsing speaker vocalization from the background noise much more challenging. It is worth mentioning that extensive research has been undertaken around solving the issue of contaminated audio documents. Many effective methods have been developed in reducing noisy background effects on recognition performance [13, 14]. However, a comparative investigation of different signal to noise ratios (SNR) that are representative of realistic scenarios, as proposed in Chapter 3 have not yet been explored thoroughly in the literature. It may be a realistic assumption, also, that there is a lack of control over the duration consistency of reference speech data obtained with a lack of user cooperation. Another difficulty arising from operating in conditions where user cooperation is absent, is that the phonetic contents of a given test utterance may not serve as a reasonable reference model for the speaker's true speech patterns, as based on short training speech. Additionally, in cases where variable duration training data have been collected for the registered subjects, a test utterance spoken by an enrolled speaker could possibly achieve a better match score against the voice model for another speaker who provided a fuller representation of the phonetic elements in the test utterance.

1.4. Aims and scope of project

1.4.1. Aim

The aim of this research is to develop an effective, novel method to enhance voice recognition performance of current classifiers when the audio data are obtained under uncontrolled environments. An ability to classify and successfully recognize speech in suboptimal conditions, such as those encountered in the street, in restaurants or other places of businesses, or even under combat field conditions or where there are multiple speakers and line-of-sight visual observation is not possible to support identification, is critical to the scope of surveillance activities under threat conditions.

The main focus of this study is to reduce open-set identification errors (OS-IE), which are unrecoverable errors, which can occur during the first stage of open-set speaker identification systems. OS-IE can have severe consequences when being used in law enforcement applications, when identification data could be used to motivate or support prosecution or further investigation in a criminal case. In order to complete the study, the work requires a systematic literature survey of existing current speaker recognition classifiers, including audio preparation procedures and classification techniques. This provides an in-depth understanding of the various methods and aspects of previous work.

As discussed previously, the open-set text independent speaker identification system is the most challenging sub-class of speaker recognition and it is believed that the challenges proposed in this study further complicate the decision made by the system to identify a speaker. Therefore, a main part of this thesis is to develop methods for reducing the OS-IE, by emphasising and optimizing areas within the audio document that are richer in speaker characteristics vs. the rest of the audio document. Thus, an in-depth investigation is conducted to identify the effect of data obtained in uncontrolled environments on the recognition performance of the baseline and the current state of the art recognition system. In addition, the realistic scenarios proposed for study is investigated, and new methods are proposed in enhancing the recognition performance. Finally, a novel method of analysis is proposed and implemented against the current classifiers.

1.4.2. Motivation

Major development has been achieved in the field of speaker recognition but not significant emphasis on realistic conditions with no user cooperation was considered, which leads to variation in duration as well as quality of reference material. For this reason, this study is mainly concerned with investigating the current classifiers performances under realistic conditions and finding solutions in improving recognition accuracy.

1.4.3. Contribution to knowledge

Open-Set Speaker Identification

As part of this study, investigation was conducted to identify the effect of environmental noise and variation in reference materials duration, on the current baseline and state of the art classifiers recognition performance. The following contributions were achieved and followed by a novel approach which improved all classifiers identification performances by relative improvement of 6% in some cases.

The effect of environmental noise on recognition performance

- The current state of the art and baseline classifiers recognition performance are very similar under extremely contaminated reference material
- White noise is not a good representation of environmental noise as it contaminates all components of the data, while realistic noise (coloured noise) has a more random effect on the audio data components.
- Normalisation techniques offer a significant improvement for the baseline and the state of the art classifier

The effect of varied duration reference material on recognition performance of the baseline, an extension to the baseline and current state of the art classifiers

- In the case of short duration reference material performances of all classifiers drop dramatically. There is no improvement witnessed between the performances of the current state of the art to the baseline system, with normalisation techniques.
- In the case of varied duration reference material, the current state of the art outperforms the baseline and extended baseline classifiers.

- In the case of varied reference and test data all classifiers performance dropped and were similar in performance. Furthermore, the normalisation technique was to a disadvantage when it comes to the current state of the art.

The novel approach

The novel approach was implemented to the baseline, extension of the baseline and current state of the art classifiers, under four different training conditions (Long, Medium, short and mixed)

- Under considered conditions of long, medium and short reference material, the novel approach improved performance of all classifiers.
- Under the more realistic condition of mixed reference material (varied duration of reference material) the novel approach improved recognition performance for all considered classifiers and were more beneficial to the current state of the art.

1.5. Publications

Chapter 3 was published by the author under the title of: Effectiveness in Open-Set Speaker Identification, Security Technology (ICCST), 2014 International Carnahan Conference

Chapter 4 was published by the author under the title of: Open-set speaker identification with diverse-duration speech data SPIE 9457, Biometric and Surveillance Technology for Human and Activity Identification XII, 94570G (15 May 2015)

Chapters 5 have been published by the author under the title of: Open-set speaker identification with mixed-duration reference data, Journal of IET Biometrics, awaiting approval

1.6. Organisation of thesis

Chapter 2: literature review

This chapter will outline the human speech system, followed by a description of the pre-processing of the audio material which will be undertaken. This pre-processing will include the analogue-to-digital process, pre-emphasis, windowing techniques, and will feature extraction techniques.

This chapter will also thoroughly describe existing current speaker recognition systems and methods, such as the baseline Gaussian mixture model (GMM) and Universal background model (UBM) system and the current state-of-the-art i-vectors. This is followed by examination of pre-systems, which the audio material is passed through before it is used by the speaker recognition systems. These pre-systems include voice active detection (VAD) and speaker change detection (SCD). Further investigation is then conducted into the developed methods of normalisation: techniques used to overcome some problems posed by the challenges identified in this study; finally the adopted measurement method is explained.

Chapter 3: The Effects of Environmental Noise on the Accuracy of Open-Set Speaker Recognition

This chapter further investigates contaminated data and its effect on the performance of speech recognition systems. A comparison study using experimental investigation is undertaken to identify the performance of each recognition classifiers under realistic conditions for contaminated data. In addition normalisation techniques such as the test normalisation (T-norm), zero normalisation (Z-norm) and TZ-norm, are applied to enhance the recognition performance of the current baseline Speaker recognition (SR) system.

Chapter 4: Open-set Speaker Identification with Diverse-Duration Speech Data

Open-Set Speaker Identification

This chapter thoroughly investigates the challenge proposed, which involves operating with short and varied duration reference speech. The study presents investigations into the adverse effects of operating conditions on the accuracy of open-set speaker identification, based on both GMM-UBM and i-vector approaches. Furthermore, an experimental investigation is conducted and the WBS is adopted to further enhance the GMM-UBM with the assistance of normalisation techniques such as TZ-norm.

Chapter 5: Vowel Boosting: A Novel Approach to Enhance the Reliability in Speaker Identification

This chapter proposes a novel approach to speaker recognition: vowel boosting. This approach is focused on enhancing the recognition performance of speaker recognition, under the conditions being considered in this study. Investigation into the current phonetic-based speaker recognition methods is undertaken as well, followed by the introduction of the” Vowel Boosting” approach.

In this chapter further shows the experimental study that is conducted on the proposed “Vowel Boosting” approach. The new method is applied to the current classifiers (baseline GMM-UBM, and current state of the art i-vector) under the condition of short and varied training material and the results assessed and discussed to provide a thorough analysis of the results obtained.

Chapter 6: Summary and future work

This final chapter summarizes the work and suggests a number of ways the project may be advanced.

CHAPTER TWO

2. Literature review

2.1. Speaker recognition overview

Speaker Recognition is the process of recognizing an individual based on his/her voice. It can be classified into two types, speaker verification and speaker identification. Speaker verification is the process of verifying a speaker's claimed identity based on his/her already registered voice whereas speaker identification involves identifying whether a speaker's voice matches or not with any member of several registered voices [15]. Speaker verification is therefore a one to one matching process whereas speaker identification typically involves performing one to many matches. Both of these can either be text-dependent or text-independent. In the former case, a fixed and pre-defined text string is provided to the speaker with which the voice patterns are compared to, while in the latter case the text is arbitrary and typically unknown [16].

Speaker identification can again be of two types: open-set and closed-set. In the closed-set case, it is assumed that the test voice pattern is already present in the database and simply needs to be identified. In the open-set case, it is not known from beforehand whether the test voice pattern is actually present in the database or not. Open-set matching is therefore more challenging [11] as it not only involves a comparison technique but also requires appropriate thresholds to prevent false matching of new voices with existing voices. As already mentioned earlier in the introduction, the focus of this work is in open-set speaker identification, which is reviewed further in Section 2.3. However, firstly, the following section will review speaker recognition process in detail, then followed by open-set speaker identification as this is part of speaker recognition process.

2.2. Speaker recognition process

In general, speaker recognition systems are used to identify an individual using their voice data. Prior to the speaker identification stage, the audio data are passed through several systems to format and process the audio material, an illustration of these processes is given in Figure 2.1

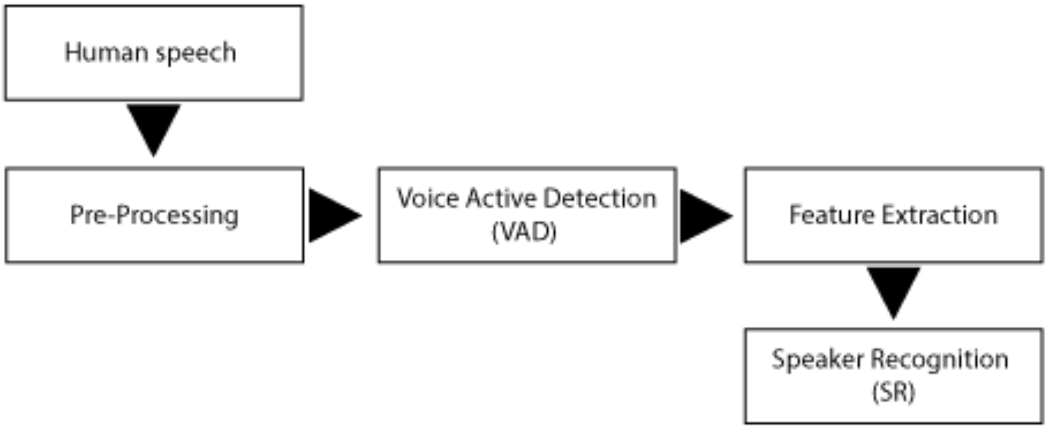

Figure 2.1 Illustration of speaker recognition processes

This part of the chapter presents a review of the literature for the methods concerned with the process of speaker recognition. An appreciation of the human speech production system as well as the phoneme classes that makeup a language will be discussed: these will be useful in understanding the nature of the differences observed between speakers' voices, or in the description of the human speech production system. This discussion is followed by a brief overview of speech analysis techniques considered appropriate for the speaker classification systems. Finally, prior systems through which the evaluation data have been passed are discussed; these include the voice active detection (VAD) [17] system, followed by the speaker change detection (SCD) system. The speaker recognition system, as well as the most popular calculation techniques that are currently used in speaker recognition, are also presented in the chapter.

2.2.1. Human speech production system

Speech is formed through the natural acoustic pressure that originates from the voluntary movement of several anatomical and physiological structures which comprise the human vocal system. These anatomical and physiological structures include the different structures as given in Figure 2.2 [18]: the lungs, wind pipe (trachea and larynx), throat (pharyngeal cavity), oral or buccal cavity (mouth) and the nasal cavity (nose). Normally, the pharyngeal and the oral cavities are grouped into one unit called collectively, the oral tract. The nasal cavity is normally called the nasal tract. Together these comprise the vocal tract.

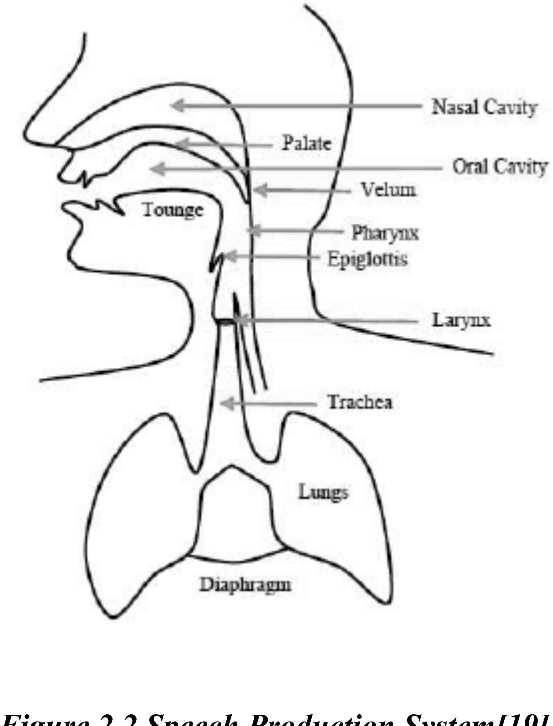

Figure 2.2 Speech Production System[19]

Muscle force is generated from the diaphragm, then the intercostal muscles apply pressure to the lungs, expiring air through the bronchi and trachea into the larynx, which houses the vocal cords (also known as vocal folds). The muscular vocal folds consist of twin infolding mucous membranes stretched horizontally across the larynx from back to front. Air movement and laryngeal muscle action push the vocal folds together, which lengthens or shortens them. Phonation occurs as the vocal folds are pushed together; thus, the air flow expelled from the lungs is modulated to produce sounds during phonation. The right amount of pressure and specific vocal fold positioning can result in vibration of these folds at different acoustic frequencies to modulate vocal pitch. Pressure generated from the lungs, acting below the folds, results in them being forced upwards and apart. The high air pressure moves high-speed air through the glottis, or space between the folds, resulting in suction, which draws the folds back together, assisted by the tension already present in the folds.

This results in a cycle of opening and closing of the glottis, assisted and enabled by the elasticity of the folds. Depending on the ratio of pressure to flow, or acoustic impedance, oscillation is achieved. Thus, the vocal fold vibration is a passive process. Although the muscles housed in the larynx do not play an active role in directly causing the vibrations, they do contribute to its control by determining by how much the folds are pulled apart or pushed together. The generated glottal signal, as illustrated in Figure 2.3, passes through and is filtered by the vocal tract, which is comprised of three separate cavities: the pharyngeal cavity, the oral cavity and the nasal cavity (cf. figure 2.2). This is where vocal harmonics are acquired, as are articulations through precise movement of the jaw, tongue, soft palate, and lips, which cause the natural resonance to occur at different frequencies. The sound produced by these combinations is called a speech stream.

Formally, the speech streams that fall within the amplitude of the human voice range are classified into two categories voiced and unvoiced sounds as shown in Figure 2.3, below.

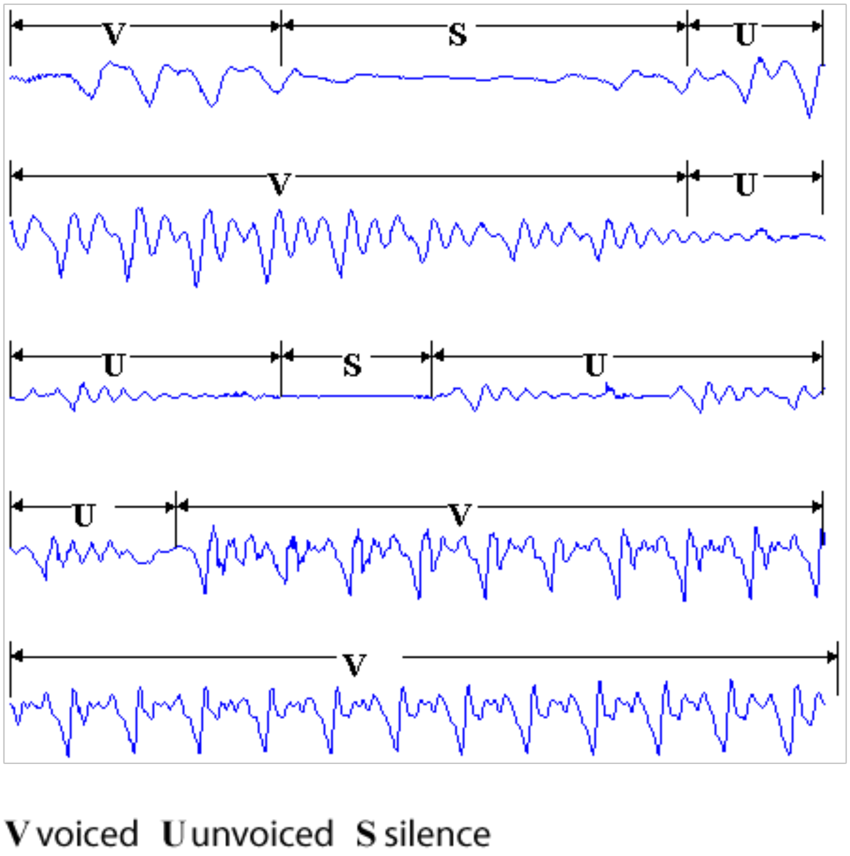

Figure 2.3 Voiced and unvoiced sound waves

During speech production, the sound category into which speech falls is dependent upon the state of the glottis, as well as the presence or absence of vocal-chord vibration during speech production, thus, speech may be voiced or unvoiced. Voiced Speech results from the combination of air pressure and vocal cord vibrations during phonation; it shows regular oscillatory characteristics in vocal document graphs (see Figure 2.3). Unvoiced Speech does not involve the use of vocal cords: it is produced by air flow moving through constricted regions of the vocal tract. In unvoiced speech, the vocal cords are open and separated, so that the voice stream appears more random and noisy. In fact, it closely resembles background noise. Phones produced as a result of unvoiced sounds include plosive sounds like /t/, /tch/, /ch/, /p/, or /k/. Voiced speech is easier to hear than unvoiced speech; unvoiced speech can be missed in the presence of a noisy background or in a group of multiple speakers. In unvoiced speech, vocal chord vibrations do not occur because of the rapid reduction of trans-glottal pressure, and separation of the vocal folds. Therefore, the voice stream is characterized as unvoiced, if a sound stream is produced only by the passing of air through an open glottis, without vocal chord vibration. Therefore, in the context of speaker recognition, the voiced sound will be considered the most important parameter in this study, and that is where the weight of this section will lie.

The process of phonation which produces voiced sounds finds its most common interpretation based on the neo-plastic aerodynamic principle [20]. The power supplied by the lungs is controlled by the diaphragm, resulting in expansion and contraction. The vocal folds (vocal chords) act like an oscillator, chopping of the storm of air pressure pushed by the lungs and thereby creating a sound. In practice, the nerve impulses transmitted from the brain to the larynx muscles constrain the vocal folds; in turn this provides the necessary conditions for vibration to occur. The last stage in this process is the movement of the air from the lungs and restricted areas of the trachea and sub-glottic space through the glottis, into a bigger space, which causes a sudden drop in pressure. The sudden drop of pressure occurs when the vocal folds are drawn together resulting in a complex tone, which is the initial source of voiced sound in speech. For this reason, it has been named the voice source. Because of the inclusion of more vocal cord anatomy in the production of voiced data, voiced sound is considered more distinct and unique to an individual, because the sounds produced depend highly on the specific anatomy of the vocal chords, as well as the bone and musculature that control the vocal chords [21].

2.2.2. Phones

The sounds of language, or articulated sound streams, are called phonetics or phones. A phone is defined as the minimal unit of sound that has a semantic constant within a language. Phones determine the difference between words, for example the phoneme ‘p’ and ‘b’ are what determine the words ‘pat’ and ‘bat’ respectively. There are two major classes of phones: consonants and vowels. Consonants can be further divided into sub-classes, some of which have voiced and unvoiced phones. The sub-classes are defined per the method of production of the sounds. For instance, stops or plosive consonants are those which involve the obstruction of the speech stream either using the tongue or lips, or the rapid release of the obstruction [22]. These phones are unvoiced, and produce these pairs of sounds as follows p and b, t and d, k and g. Hissing sounds generated by constraining the speech stream using teeth and lips, are called fricative consonants. These phones produce voiced and unvoiced pairs; a few examples include F and V, Th and Dh (as in this and that respectively), as well as S and Z. There are also nasal consonants, which involve the movement of air through the nasal cavity by blocking passage through the oral cavity, producing M, N, NX phonemes. Affricative consonants are similar to stops, but are stops that are followed by a fricative sound, as in CH, for example [23]. There are also semi-vowel consonants, which are those consonants with vowel-like qualities, including W, Y, L, and R, and whisper consonants which produce the phoneme H.

Vowels, unlike consonants, are always voiced. The differences in vowel sounds depend primarily on the prominent resonances produced as a result of the position of the tongue and lips. Vowels generally remain unchanging over the duration for which they are produced [24]. These vowel sounds are called monophthongs, literally, “single sounds” where the tongue or other speech organs do not move once vocalization begins. When the tongue is positioned to the front of the oral cavity, depending on the height of the tongue, the phonemes IY (beat), IH (bit), EH (bat) and AE (bet) are produced. When the tongue is in mid-position the phonemes produced include AA (barb), ER (bird), AU (but) and AO (boat) [25].

There is a subclass of vowels wherein the vowel sounds do change over the duration of its production, the diphthong vowels. “Diphthong” comes from the Greek, translating, literally, as “two sounds” or “two tones” and may be known colloquially as “gliding vowels”. Diphthongs result when a vocalised phoneme begins with one vowel sound and ends with another. Examples include the vowel sound in the word buy, AY, which sounds like AA first and ends with IY, or the German ‘neu’ (new) or ‘auf’ (on). The specific movement and control of both the respiratory and articulatory components of the vocal anatomy produce voiced and unvoiced sounds which can be separated into phonemes, which together produce syllables and subsequently words, and constitute language. The digital processing of speech first requires the conversion of the pressure wave produced in the formation of speech into an electrical signal. This is achieved using appropriately chosen transducers. The electrical signal produced is then converted from an analogue signal to a digital one via an analogue- to-digital conversion process. Commonly, this process requires sampling the electrical signal as illustrated in Figure 2.6, at a rate ranging from 8000 to 16000 samples per second, then representing each sample as an 8-bit or 16-bit sequence. The aliasing problem can result in distortion of the signal, so to avoid this, the signal is put through a low pass filter to limit its bandwidth to the Nyquist range.

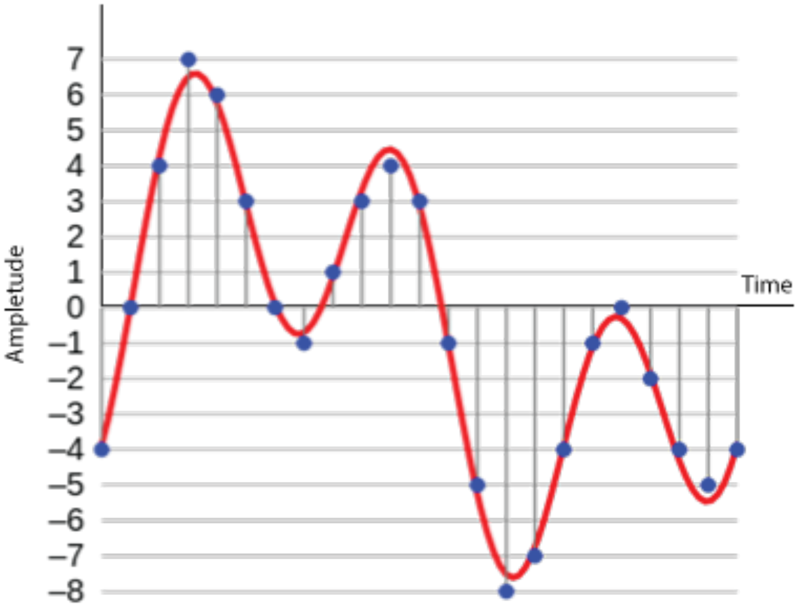

Figure 2.4 Vocal Signal Digitization result [29]

Due to the mechanism of production, speech is a gradually varying signal, meaning that, if examined in short time fragments, say, less than 100ms, the characteristics of the speech signal are nearly constant. However, if speech is analysed in longer fragments of time of over 200ms, for example, the characteristics may vary with time, from segment to segment. This is caused by variations in the vocal tract that occur as different sounds are produced. Resonances in the vocal tract may also change the frequency content of the signal as it passes through the vocal tract. Any changes in the resonance structure results in the formation of different sounds. This can be observed in Figure 2.7, where 500 ms of an utterance is shown.

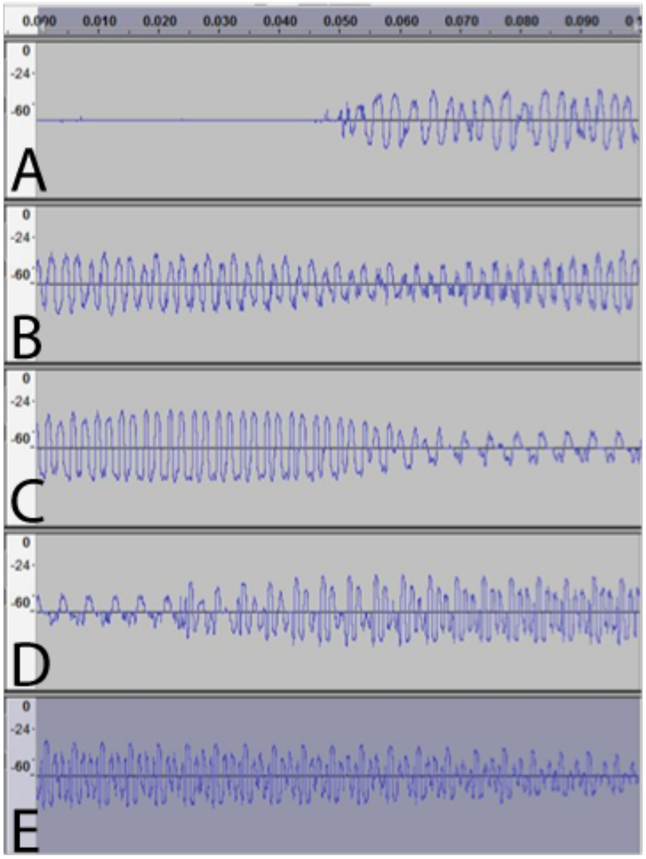

Figure 2.5 Sample utterance, 500ms [26]

Section A of Figure 2.5 shows the nature of slow time-variance in the signal. The first 100 ms of the waveform corresponds to background noise, then is followed by the start of the speech. The unvoiced part of the speech can be noted as the random area.

Acoustic speech waveforms are the two-dimensional representations of sounds generated by speech, wherein the vertical dimensions represent intensity of the sound, and the horizontal dimension represents time. Intensity also can be interpreted as sound pressure; both intensity and pressure are the physical measurements of sound amplitude. The peaks in the wave are called speech formants, and are caused by resonates that correspond to a specific configuration of the vocal tract. For each specific sound, the relative formant is located in a similar position within a speaker's spectrum, because it is the same sound being produced. However, close examination of corresponding formants for different speakers shows that they occur at slightly variable frequencies and intensities. This means individuals have unique frequencies and intensities in their voice, which result from the unique anatomical structure of their vocal tract. This is what most automatic speaker recognition systems rely upon for determining different speakers.

Considering the process of speech production, it can be said that the vocal chords expose a significant speaker-dependent or unique individual characteristic of speech signals known as the pitch [27]. The pitch of one's voice is a key feature which aids in distinguishing between different voices; however, with regards to the measurement of pitch, reliability issues can arise. In other words, reliable measurements are quite difficult to obtain, especially under conditions with a lot of noise [28]. Similarly, significant disadvantages arise in the use of speech for identifying speakers: speaker identification based on pitch is highly vulnerable to changes in speaker emotional state (the subject is excited, depressed, happy, mad, or sad), energy levels, or in response to non-physiological factors.

Another unique characteristic of the speech stream is the frequency component of the speech spectrum. As explained above, the unique anatomical variability of an individual's vocal anatomy result in speech being produced at varied frequencies which can occur randomly. In order to identify these variations in frequency, the speech signal may often be analysed in segments or windows of short duration, wherein the speech can be considered unchanging. This enables the analyst to make measurements on the short-term spectrum, a process that is most popular in speaker recognition techniques. The short-term spectrum signal itself consists of two parts: the first is called the spectral envelope, and is the characteristically slow-varying part of the speech signal, produced from speech system resonances. The second part is called fine structure, which, unlike the spectral envelope, is a quickly-varying signal that is produced from the vocal chord vibrations (Figure 2.6). Both the spectral envelope and fine structure may be used in automatic speaker recognition, but there is no agreement as to which gives the best spectral representation.

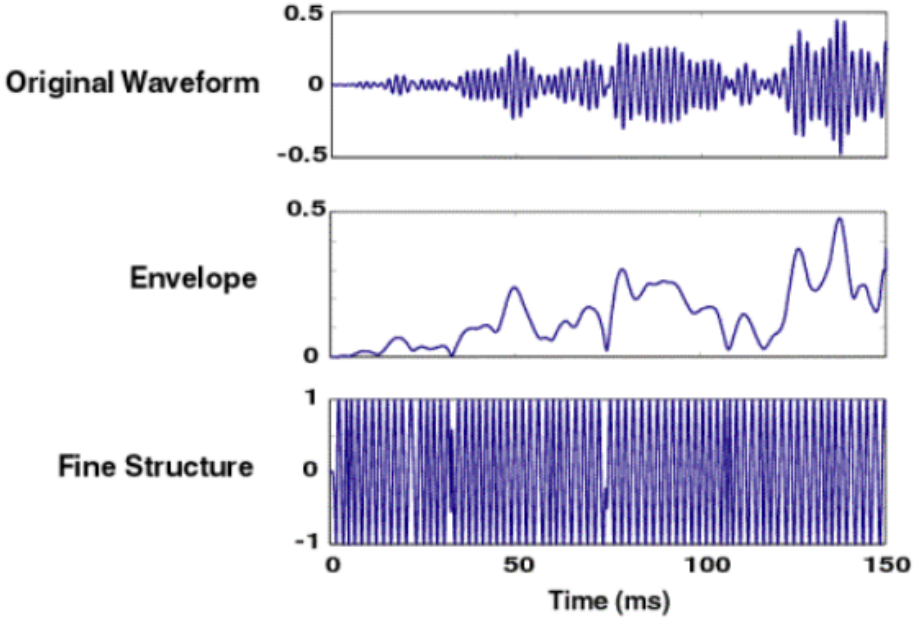

Figure 2.6 The original wave form, the spectral envelope and fine structure of an acoustical signal

One of the most common spectral representations used in automatic speaker recognition are linear predictive coding (LPC) and their many transformations [29], and the filter bank energies and their cepstral representations [30]. LPC analysis is based on an all-pole application of the speech signal produced as a result of nearly intervallic glottal pulses produced by vibrating vocal chords for voiced speech, or turbulent air flow through a constricted vocal tract, in unvoiced speech [29]. The LPC model assumes that each sample of the speech waveform is a linear combination of previous samples. Predictor coefficients, which are the coefficients used in this combination, are used to minimize anticipated mean-squared prediction error.

The suggested all-pole application of the speech signal is coherent with modelling the vocal tract as a continuous acoustic tube with variable cylindrical sections of roughly similar length, but with varied cross-sectional area. In this model, at the boundaries of the cross-sectional areas, a proportion of the sound waves are reflected. The percentage of reflected sound waves at these junctions is labelled reflection or PARCOR coefficients [31]. These coefficients can be deduced from the speech signal within the LPC analysis framework. This is further explored in (2.2.6).

As well as reflection coefficients, the all-pole spectral application can also yield line spectral pairs (LSP), which are coefficients that are the roots of two polynomials based on the reverse filter of the LPC model. These polynomials are the product of extending the variable acoustic vocal tube with an extra section that is either entirely closed (Area= 0) or entirely open (area=1).

Unlike all-pole applications, the alternative filter-bank analysis is based on mimicking the human perception of speech [32]. Studies have demonstrated that the perception of the pitch of a pure sine waveform produced by speech did not match up linearly with the actual observed frequency of the pure tone. A Mel scale was derived for the purpose of mapping real frequencies on to perceived frequencies [32], which shows the presence of a linear correspondence between real frequencies and perceived frequencies of up to 1kHz, as well as a logarithmic correspondence for higher frequencies.

An interesting phenomenon within the perception of tones is called masking, wherein the ability to hear one tone may be compromised by the presence of an adjacent tone [33].

The closer the frequencies of the adjacent tones are to one another, the greater the effect it has on the ability to hear them. This phenomenon results in the development of a critical band which defines the regions surrounding a frequency (the regions where masking is felt), resulting in the formation of the Bark scale. The Bark scale is a psycho-acoustical scale proposed by Eberhard Zwicker in 1961 [38, 39]. Before continuing, it might be reasonable to define what the word, 'psychoacoustic' means: it refers to the physical features of sound as related to audition, as well as with the physiology and psychology of sound receptor processes. In other words, it encompasses a field of study concerned with exploring the human perception of sound through physiology, psychology, and physics. The Bark psycho-acoustical scale, therefore, can be described as a frequency scale on which actual equal distances correspond with perceptually equal distances. Above ~500 Hz, this scale approximates a logarithmic frequency axis; below 500 Hz, it approximates a linear function. The Bark scale may be used, like the more popular Mel scale, as a representation of the frequency scale as a linear, perceptually meaningful scale [3].

The filter bank analysis method is the culmination of the two theories presented. It is a method wherein a set of filters, which cover the Nyquist range (also known as the folding frequency range) of the digitalised speech, are designed such that their centre frequencies are equally spaced in both the Mel and Bark scales [34]. Their bandwidths are deliberately chosen to be close to the critical bandwidths for the corresponding centre frequency. If explained a different way, essentially the upper and lower ends of the frequency filters are such that they lie in centre frequency range of adjacent filters [35]. The speech signal is analysed frame by frame in the time domain. The Discrete Fourier Transform (DFT) is used to transform each frame in the time-dependent vocal graph into the frequency domain. The logarithm of the sequence obtained is multiplied by the spectrum of each filter mentioned and its resulting sequence is summed. The Discrete Cosine Transform (DCT) is then used to transfer the result into the cepstrum domain. These parameters are known as the Mel-Frequency Cepstrum Coefficients (MFCC).

The relative performance of each of these techniques is highly dependent on the application. For example, the MFCC is based on the principle of homomorphic signal processing [35]. This transformation is useful in speech processing because it allows the separation of the two excitation and vocal tract signal components. Therefore, the MFCC provides much better representation of speech signals for speech recognition applications [36, 37]. The Linear Prediction Cepstrum (LPC) reflects the differences of biological structure in the human vocal track. Research has shown it performs best on text-independent identification applications [38, 39]. This is the main reason behind the decision to choose LPC to parameterise the audio for the purposes of this study.

2.2.3. Voice active detection

Considering the challenges around obtaining audio document in an uncontrolled environment, as proposed in this study, a realistic assumption to make, would be to assume that the data collected could be subject to silences, or to areas containing noise. In such cases, the Voice Active Detection (VAD) process could be used to identify areas within the audio document that contains speech. This is an important process, as it reduces computational time significantly, by preventing the analysis of time frames which hold no useful data [40]. Further, it assists in improving the recognition performance of speaker recognition classifiers [41].

Voice Active Detection (VAD), also referred to as Speech Activity Detection (SAD), is a fundamental processing task in almost all fields of speech processing. VAD uses the speech processing algorithms which analyse the audio signal and indicate speech segments in the vocal document. As stated previously, VAD is typically used to remove silence and noise segments in the vocal document [26, 42]. The kinds of non-noise segments in speech can be quite diverse: including silence or ambient noise such as paper shuffling, door knocks, or non-lexical noise such as breathing, coughing, and laughing. Therefore, highly variable energy levels can be observed in the non-speech parts of the signal.

In general, there are various approaches to SAD, such as feature extraction techniques [27] (energy, spectrum divergence between speech and background noise, and pitch estimation). These methods combined with a threshold-based decision, have proven to be relatively ineffective [28, 33, 43]. There are alternative model-based approaches which tend to have better accuracy than SAD. They rely on a two-class detector, with models pre-trained with external speech and non-speech data [26, 27]. Discriminant classifiers such as linear discriminant analysis (LDA) coupled with Mel frequency coefficients MFCCs [44] have also been used in the past. The main drawback of the model-based approaches is that they rely on external data to train the speech and non-speech models, which makes them less robust and less responsive to changes in acoustic conditions.

Hybrid approaches have been proposed as a potential solution to optimize speech recognition processing. In most cases, an energy-based detection is first applied in order to label a limited amount of speech and non-speech data for which there is high confidence in the classification. In a second step, the labelled data are used to train speech and non-speech models, which are subsequently used in a model-based detector to obtain the final speech/non-speech segmentation [27, 32, 45]. A good example of this is the study in [17] where higher detected performance was reported through the use of a Gaussian statistical model that was applied to the VAD process using decision-directed (DD) methods based parameter estimation.

This was further improved by the studies [46, 47] where Support Vector Machines (SVM) statistical modelling was applied, which further improved the performance. SVMs are supervised learning models that analyse data used for classification and regression analysis. They are non-probabilistic binary linear classifiers.

2.2.4. Speaker change detection

The final process step, prior to speaker recognition processing, is the speaker change detection (SCD) process. The audio document obtained under uncontrolled conditions is also likely to have been obtained under conditions where there is more than one speaker present. For this reason, it is necessary to pass the audio document through an SCD system. Because in many cases the audio document contains more than one speaker, it is essential to determine with a high degree of certainty, when a speaker change occurs in the audio document [48]. To identify the point of change in speaker, one task requires the segregating of parts of the document corresponding to homogenous speakers, which results in different segments classified as belonging to different speakers. This is an essential stage in speaker recognition and it is very crucial to correctly identify the sub-segment belonging to each speaker. Missed points around a speaker change or false detection of speaker change points where there has been none can adversely affect the performance of the system. Thus, Speaker change detection (SCD) is an essential stage in the speaker recognition process. More details of speaker change detection is given in appendix B

2.2.5. Pre-processing

Once a voice speech stream has been obtained, it must be prepared for analysis. This may involve a number of steps, depending upon the data quality as well as factors like speech stream volume, speaker pitch, and background noise [49]. The optimal approach to preparation, or utterance pre-processing, is chosen to optimize desired characteristics. What's more, where multiple speakers are present in uncontrolled environments, background conditions may be changing in random or in non-random ways at a given time: speakers may be moving around the space being monitored (introduces variability in volume and precision of captured speech, or the speaker(s) may be engaging in activities that can interfere with, or mask parts of the voice stream in unpredictable ways). Male and female speakers could present additional difficulties to the voice stream analysis as well, in terms of the emotional content and range of speech [49]. The problem of distinguishing the speech signals from non-speech signals is critical, and robust, thus reliable methods to parse them are needed.

Pre-processing is the first stage in analysing speech. Speech data obtained in the field is subjected to the first digital filtering as described in equation 2.1. This process is called *pre-emphasis*[50]. It is understood that audio signals, including speech signals, all tend to have lower energies at high frequencies; this is known as a *negative spectral slope* [51]. In the case of speech signals, this occurs due to the physiological characteristics of the speaker: to one's speech anatomy. For voiced sounds, this effect is highest where glottal signals can have a negative spectral slope of approximately 40 dB/decade [49] Although the radiation of the speech signal from the lips gives the spectrum a boost of about 20 dB/decade [51], the speech signal recorded at a distance via a microphone has a -20dB/decade slope when compared with the original signal of the vocal tract (also known as the true spectrum). The aim of the initial filter is to offset this low energy-high frequency effect, so that the measured spectrum has a comparable dynamic range across the entire frequency spectrum> Ensuring this helps to limit the effects that the vocal tract has on the glottal signal. In addition, this initial digital filter minimizes numerical instability during the LPC-based feature extraction process [65].

The pre-emphasis filtering procedure is accomplished by applying a high-pass FIR filter in the form of

$$\boldsymbol{H_{pre(z)} = 1 - \alpha z^{-1}} \qquad \textit{2. 1}$$

where 'α' determines the cut-off frequency of a single zero filter. The filter is a differentiator that flattens the speech spectrum. This counteracts spectral roll-off, thereby increasing the accuracy of speaker recognition [52]. Usually, α is a constant in the range of 0.4 -1.0 [45].

Unvoiced speech does not require compensation for spectral slope, because it does not occur via glottal contribution. Thus, unvoiced spectra do not demonstrate the same spectral trends as voiced speech. Knowing this, the application of pre-emphasis may be a negative process as it will result in the reinforcement of already large high frequency components [54].

Othman and Abdul Nasser (2003) have suggested a solution to alleviate this issue, where an optimum value of α, given by equation 2.1, may be used. In equation 2.2, R0 and R1 are autocorrelations of a segment of speech at lag zero and lag one respectively.

$$\boldsymbol{a_{opt} = \frac{R_1}{R_0}} \qquad \textit{2. 2}$$

For voiced segments of speech, it is expected that there is a high sample-to-sample correlation, meaning $R1 \approx R0$ or $\alpha opt \approx 1$. In unvoiced segments, there is little or no sample to sample correlation, therefore $\alpha opt \approx 0$. The determination of the optimal value is quite computationally expensive; it may be for this reason that in speech processing applications fixed values for α are preferred. For the purposes of this work, a value of 0.95 is used.

After pre-emphasis, the next step consists of grouping the speech samples into frames of approximately 20-30ms, in a process called frame-blocking. The time period is chosen to be very short for the frames of speech, because this reflects the stationary nature of speech at such durations. Frame-blocking is equivalent to multiplying the speech signal by a rectangular window which is zero during all periods except during the analysis period. This means discontinuities occur at the edges of the frames, which can distort the spectrum by adding false high frequency components.

A solution to this induced error is to multiply the signal by a tapered-type window, such as the Hamming window (analytically defined in equation 2.4), where the amplitude of a signal slowly tapers to a zero at both ends of the frame range, in a bell shape as demonstrated by Figure 2.8. Because of this feature, tapered-type windowing can mean speech events that are near the ends of the windows may be given a low weighting, meaning such samples will not be effectively included in the speech analysis. This issue is circumvented by overlapping the segments in a way that every section of the frame is covered by at least two overlapping windows. Often adjacent windows are overlapped by 50%. This means that the segments of speech in one window that are near the end and therefore receive a lower weighting, are near the centre of the adjacent window, wherein it will receive the highest weighting. The weighting function is:

$$w(n) = 0.54 - 0.46\left(\frac{2\pi n}{N-1}\right) 0 \leq n \leq N-1 \qquad \textbf{2. 3}$$

Where n represents the index of the sample and N is the total number of samples in one frame.

Figure 2.7 demonstrates how the frame blocking process effects a sequence of speech samples. After pre-processing, every speech frame is exposed to the feature extraction processes, further detail on which is given in the following sections.

A solution to this induced error is to multiply the signal by a tapered-type window, such as the Hamming window (analytically defined in equation 2.4 and discussed briefly in section 2.2.5.1), where the amplitude of a signal slowly tapers to a zero at both ends of the frame range.

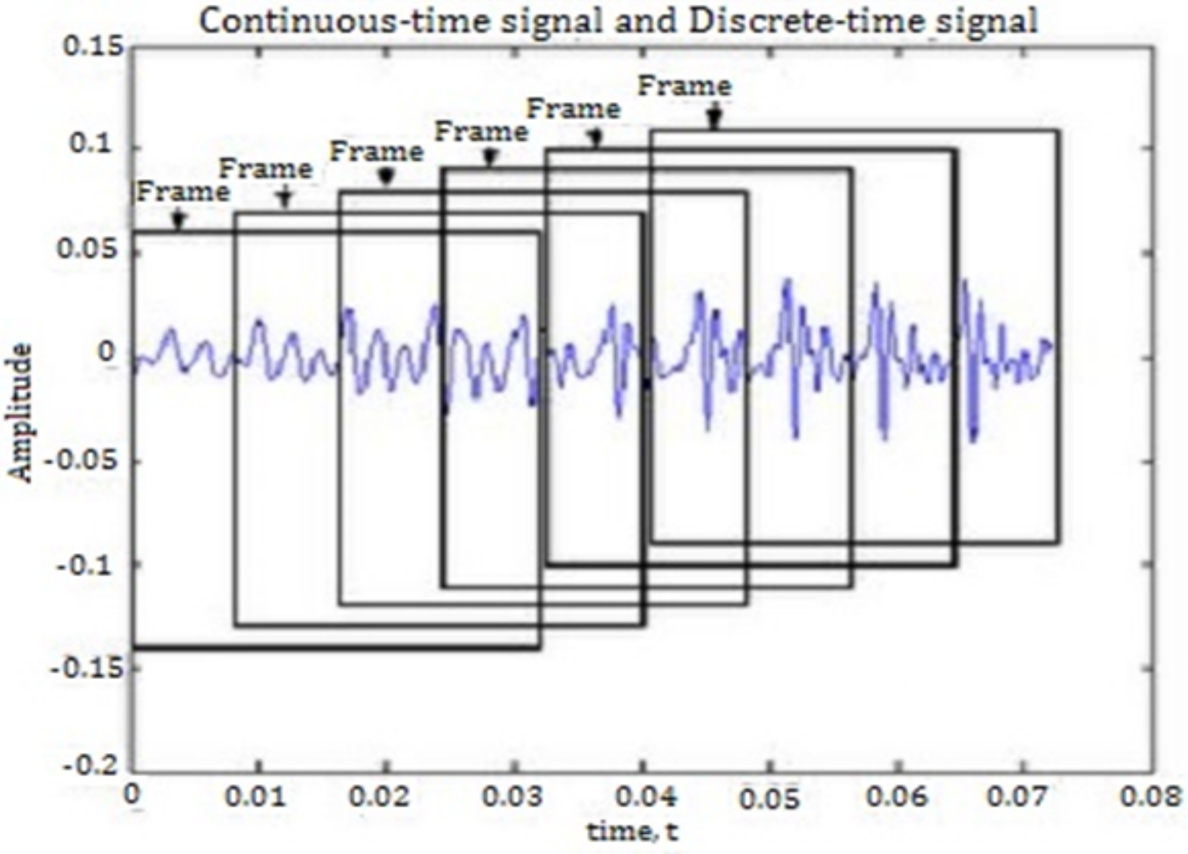

Figure 2.7 Frame-blocking schematic

2.2.5.1. Hamming window

The Hamming window is a cosine block windowing smoothing function that is named after the man who proposed it, Richard W. Hamming. The block windowing function is written:

$$\boldsymbol{w(n) = 0.54 - 0.46\, cos\left\{\frac{2\pi n}{N-1}\right\}},\ for\ \boldsymbol{0 \le n \le N-1} \qquad \textit{2. 4}$$

The Hamming block window is used to simplify complex functions (Figure 2.10) and is considered a natural choice to process real-time applications that require both windowed and non-windowed (rectangular windowed) transforms. The windowed transforms produced using the Hamming block window function to segment data can be derived efficiently by convolution from the non-windowed transforms. The function itself is symmetric and bell-shaped (figure 2.10) [53], so that features lying within the window near the centre are given the greatest weight, and as mentioned earlier, those which fall at the edges are given a lower weight. The function is easily processed by Discrete Fourier Transform (DFT) analysis, as shown on the right-hand side of figure 2.10.

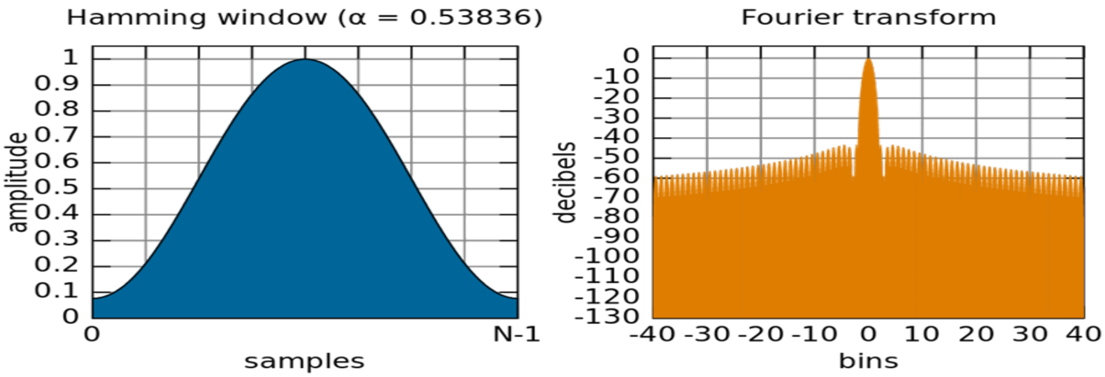

Figure 2.8 Example of a Hamming Window and its DFT. See Equation 3.4 for function

2.2.6. Voice stream feature selection and extraction

The speech samples which fall within a given block frame window may be subsequently encoded into a vector that represents the entire speech sample block which lies within the window. This has the effect of reducing the dimensionality of the data. This is possible because the number of feature coefficients contained within a window block is lower than the number of block windowing samples [48].

This method of discretising and simplifying a speech sample is called Feature Extraction. Feature extraction is an essential step in accomplishing speaker recognition; it is usually performed after the pre-processing step. It involves identifying components of the speech signal that can be used for identifying linguistic content such as power, pitch and vocal tract configuration, as well as the filtering out of other parts of the speech signal which are of no use, such as background noise, or emotional content. Because the shape of an individual's vocal tract determines the sounds produced, accurately determining the shape of the vocal tract of a given speaker facilitates the accurate identification of the phonemes which are being produced. Thus, the shape of the vocal tract manifests itself within the short-time power spectrum[54].

Schoeter and Sondhi discuss at length a number of techniques for solving this problem, stating "Mathematically, the estimation of the vocal tract shape from its output speech is a so-called inverse problem, where the direct problem is the synthesis of speech from a given time-varying geometry of the vocal tract and glottis". [55]

In the literature, there is no agreement as to what the best parametric representation may be, to use for speaker recognition applications. That said, in general spectral analysis methods are considered the core of the signal processing methodology when speech processing is involved. The most common spectral analysis methods are linear Mel Frequency Cepstrum Coefficient

(MFCC) analysis, Linear predictive coding (LPC) analysis, and Linear Predictive Coding-based Cepstrum (LPCC) analysis. The relative performance of each of these techniques is highly dependent on how it is applied, and where. A study by Antal demonstrates that LPCC performs best when applied to speaker verification, whereas MFCC provides a better representation of the speech stream signal for automatic speech recognition applications [55, 56]. The details of each feature extraction process are given in subsequent sections.

2.2.6.1. Mel frequency cepstrum coefficient

Mel Frequency Cepstrum Coefficient or MFCC analysis is one of the most common techniques used to extract features from a speech signal, based on the short term spectral representations. MFCC aims to accurately represent the short-term power spectrum of each vocal tract shape. A feature vector represents each frame. The concept behind the method is to process a speaker document in a way which approximates how the human ear hears: to simulate human perception of speech, in that the distinction of low frequency sound is better than that of high frequency sound. Perception of the sound frequency content for each signal does not follow a linear scale either, much like in human perception. The sound signal is filtered, to concentrate on certain regions of the speech signal, spaced non-uniformly on the frequency axes [57]:

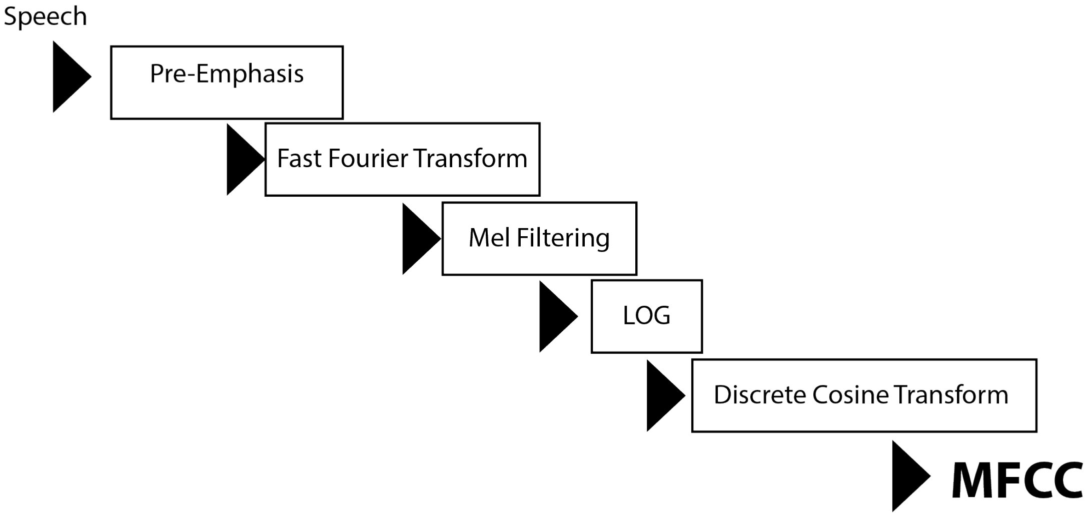

Figure 2.9 MFCC features vector creation steps [69]

The process of MFCC computation follows three steps, the first of which is called the periodogram estimate of the power spectrum. This is obtained by applying fast Fourier transform (FFT) to each short analysis window of the speech signal. Thus, each frame of N samples is transformed from a time domain into a frequency domain. The second step requires the periodogram estimates of the power spectrum to be mapped against the MEL scales using triangular overlapping windows as shown in Figure 2.10. As mentioned previously, the human perception of frequency constants in sound does not follow a linear scale. Thus, for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called Mel scale as defined by the following equation [58].

$$\boldsymbol{Mel(f) = 2595 log10\left(1 + \frac{f}{700}\right)} \qquad \textbf{2. 5}$$

Where 0≤f≤$\boldsymbol{f}$. $\boldsymbol{f}$ is defined as the frequency and the Mel(f) is the subjective pitch in Mels corresponding to the frequency in Hz. Mel filtering is the computation of a number of triangular filter outputs, applied to the power spectrum obtained from FFT, to smooth the spectrum. Window overlapping is used to compensate for data that might have been lost. As mentioned, the human ear distinguishes sounds of low frequency better than sounds of high frequency. That is why its bands are spaced linearly below frequencies of 1000 Hz, and logarithmic spacing is applied above 1000 Hz, as demonstrated in Figure. 2.10 [59] .

During the third step the log of the power of each of the filter outputs is taken, the resulting values of which are referred to as the Mel spectrum coefficients. This step is followed by the fourth step, wherein the discrete cosine transform (DCT) [60] is taken so that the Mel spectrum can be converted back into a time like or cepstral domain, resulting in MFCC features vectors. The mathematical framework of MFCC is demonstrated below:

The Mel spectrum is computed by multiplying the spectral coefficients with the filter coefficients. Triangular Mel weighted filter is summed up, and both results are integrated

This can be obtained using the formula below:

$$\boldsymbol{\breve{S}[i] = \sum_{k=0}^{N/2} S[k]M_i\,[k] 0 < i < l} \qquad \textbf{\textit{2. 6}}$$

Where S[k] is the magnitude spectral coefficients, N is the length of the FFT, l is the number of Triangular Mel weighting filters, $\boldsymbol{M_i}[k]$ is the filter coefficient of the $\boldsymbol{i}$ th triangular filter, and $\boldsymbol{\breve{S}[i]}$ is the out put of the Mel filter banks. The DCT is computed by [61]:

$$\boldsymbol{S[i] = log\,\breve{S}[i]} \qquad \textbf{\textit{2. 7}}$$

$$\boldsymbol{C(u) = \sum_{j=1}^{N}\left(S[i]\, cos\, [\frac{\pi i}{j}\ \ (j-0.5)]\right) \qquad (i = 1, 2 \ldots P)} \qquad \textit{2. 8}$$

$S[i]$ is the log of the filter output for the ith filter, N is the number of filters and P is the dimension of the MFCC

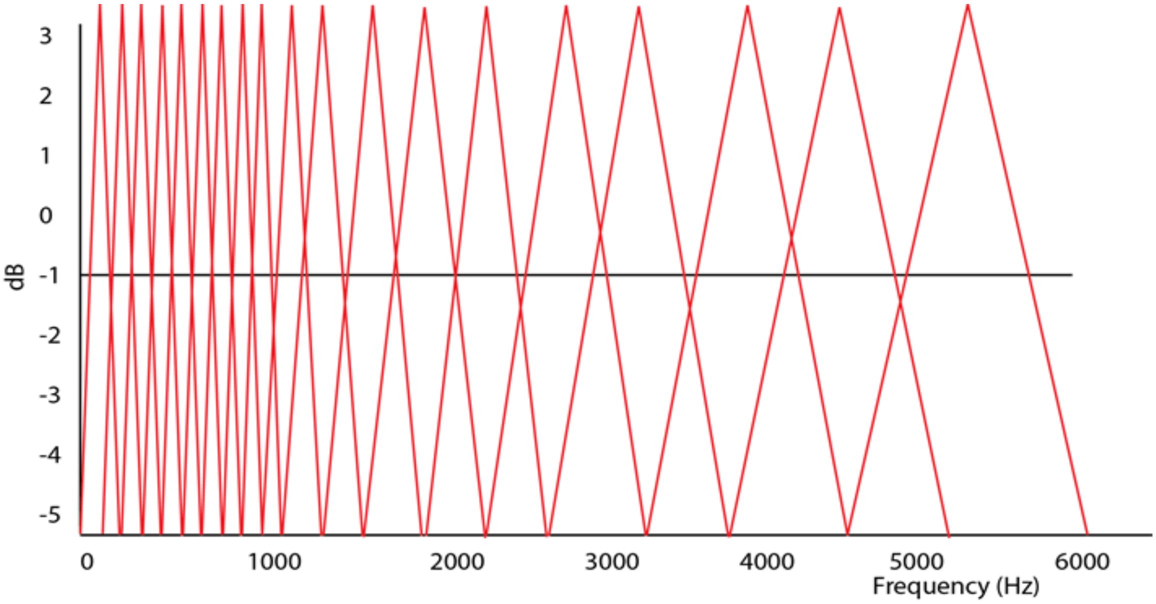

Figure 2.10 Illustrations of the Mel filter banks [71]

2.2.6.2. Linear predictive coding

Linear Predictive Coding (LPC) is a method of representing the spectral envelope of the converted digital signal in compressed form, produced after pre-processing of the raw digital voice stream, as described in section 2.2.5 via a linear predictive model. It is commonly used for speech analysis and re-synthesis, or speech compressions such as the type used for GSM communications by telecom companies, or for secure wireless communications. The process was developed as a result of research into automatic phoneme discrimination carried out in the 1960s at Nippon Telephone and Telegraph (Nippon Denshin Denwa Kabushiki-gaisha), leading to development of the methodology as it is used today, as first presented by Antal [56].

A visual overview of the Linear Predictive Coding methodology applied to a speech sample can be described as follows: the analyst assumes that the speech signal can be approximated as a buzzer at the end of a tube which produces voiced sounds, with occasional additions of hissing and popping, which approximates sibilants and plosives. This crude model is an effectively close approximation of the vocal tract. The buzz model approximates the functioning of the glottis, and sound originating from there is characterized in terms of its loudness and frequency or pitch. The tube model describes the vocal tract comprised of the throat and the mouth, and sounds originating from it are described in terms of resonances. Resonances give rise to formants: enhanced frequency bands in the produced sounds. Hisses and pops heard in the sounds originate from the actions of the tongue, lips, and throat during sibilants and plosives.

LPC analysis of the speech signal estimates the formants, removing the effects from the speech signal (inverse filtering), then estimates the intensity and frequency of the remaining buzz (the signal residue). The numbers to which the frequency and intensity of the signal components (buzz, formats, and residue signal) can now be transmitted, then reconstituted using LPC in reverse. The buzz and residue parameters can be combined to re-create a source signal. Formants are used to recreate a filter representing the tube in the original model. When the source signal is run through the tube filter, speech can be reconstituted. This process is carried out on short chunks of the speech signal to account for variances in the vocal signal with time by discretisation of the signal using block windowing to create frames. This process produces intelligible speech with good compression [62].

Digging deeper into the mechanics of LPC to examine the mathematical model, one finds that the primary idea behind it: that each speech sample s(n), can be approximated using a linear combination of the past $\boldsymbol{P}$ samples can be modelled using equation 2.9 [48, 59, 63]. The equation, shown below, demonstrates this linear combination:

$$\hat{s}(n) \approx a_1 s(n-1) + a_2 s(n-2) + \cdots + a_p s(n-P) \qquad \textit{2. 9}$$

Where the coefficients $\boldsymbol{a_1}$, $\boldsymbol{a_2}$……… $\boldsymbol{a_p}$ are assumed constant in the speech analysis frame and $\boldsymbol{P}$ is the linear prediction (LP) analysis order. It is assumed that during the whole duration of the speech signal frame, the speech signal remains unchanging [64, 65]; for example, the analyst might assume that the vocal tract has not moved into a new configuration within the boundaries of the frame.

Equation 2.9 represents an approximation only; therefore an error term is needed. This error term is the difference between the speech samples $\boldsymbol{s(n)}$and the estimate $\boldsymbol{s}$□$\boldsymbol{(n)}$ resulting from the right portion of equation 2.9 and is labelled the prediction error. The equation for this process is expressed in Equation 2.10 [66].

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^{p} a_k\, s(n-k) \qquad \textit{2. 10}$$

In the LPC model, voiced sounds can be described as an excitation source resulting from periodic glottal pulses, which can be modelled by an impulse train generator with adjustable period. For unvoiced sounds, because the sound production relies on turbulent airflow through the constrictions of the vocal tract, it can be modelled as a random noise generator, as expressed in Figure 2.11. Based on the model of 2.11, the equation 2.11 expresses the relation between the speech signal $\boldsymbol{s(n)}$,and the input excitation $\boldsymbol{u(n)}$, as affected by a gain term $\boldsymbol{G}$

$$\boldsymbol{s(n) = \sum_{k=1}^{p} a_k\, s(n-k)\ + Gu(n)} \qquad \textit{2. 11}$$

When comparing equations 2.10 and 2.11 it can be understood that $\boldsymbol{e(n)} = \boldsymbol{Gu(n)}$. Using the Z-transform to equation 2.12 yields the following:

$$\boldsymbol{s(z) = \sum_{k=1}^{p} a_k\, z^{-k} S(z)\ + Gu(z)} \qquad \textit{2. 12}$$

Reorganising equation 2.12 gives the all-poll filter transfer function $\boldsymbol{H(z)}$:

$$\boldsymbol{H(z) = \frac{S(z)}{Gu(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}}} \qquad \textit{2. 13}$$

Within each analysis frame, are a set of predictor coefficients $\boldsymbol{a_k}$, 1 ≤ k ≤ p. Consequentially the spectral properties of the all-pole filter (Figure2.11) are equivalent to the speech waveform within each analysis window. Computation of $\boldsymbol{a_k}$ is by minimising the mean-squared prediction error ε within the considered frame, as represented by the following:

$$\boldsymbol{\varepsilon} = \sum_{n=0}^{N-1+p} \boldsymbol{e}^2(\boldsymbol{n}) = \sum_{n=0}^{N-1+p} \left[\boldsymbol{s}(\boldsymbol{n}) - \sum_{k=1}^{p} \boldsymbol{a}_k \, \boldsymbol{s}(\boldsymbol{n}-\boldsymbol{k}) \right] \qquad \textit{2. 14}$$

Where N is the number of sample per frame.

The minimization of $\boldsymbol{\varepsilon}$ can be achieved by differentiating (2.14) with regards to each coefficient a_k and equal the result to zero.

A differentiating equation [67] can be used to limit error ε with regards to each coefficient $\boldsymbol{\alpha_k}$ and equating the results to zero:

$$\frac{\boldsymbol{\delta\varepsilon}}{\boldsymbol{\delta\alpha_k}} = \mathbf{0}\boldsymbol{k} = \mathbf{1}, \mathbf{2}\boldsymbol{m} \dots \boldsymbol{p} \qquad \textit{2. 15}$$

Equation 2.15 results in a system of equations [67]

$$\sum_{k=1}^{p} \boldsymbol{a}_k \sum_{n=0}^{N-1+p} \boldsymbol{s}(\boldsymbol{n}-\boldsymbol{k})\boldsymbol{s}(\boldsymbol{n}-\boldsymbol{i}) \qquad \textit{2. 16}$$

$$= \sum_{n=0}^{N-1+p} \boldsymbol{s}(\boldsymbol{n})\boldsymbol{s}(\boldsymbol{n}-\boldsymbol{i}), \qquad \mathbf{1} \le \boldsymbol{i} \le \boldsymbol{p}$$

When looking at the above equations, both the second and third extractions represent terms that are similar to short-term autocorrelation values $\boldsymbol{R}(\boldsymbol{k}-\boldsymbol{i})$ and $\boldsymbol{R}(\boldsymbol{i})$, of the speech signal $\boldsymbol{s}(\boldsymbol{n})$at lags $(\boldsymbol{k}-\boldsymbol{i})$ and $(\boldsymbol{i})$ respectively. This is an acceptable assertion as the samples outside of the analysis windows are assumed to be zero, hence they do not contribute to the autocorrelation values. If $\boldsymbol{R}(\boldsymbol{k}-\boldsymbol{i})$ is substituted with $\boldsymbol{R}(\boldsymbol{i})$, in equation 2.15 it produces the following:

$$\sum_{k=1}^{p} \boldsymbol{a_k R(m-k) = r(m), m = 1, \ldots.., p} \qquad \textit{2. 17}$$

Equation 2.17 can also be expressed in the matrix form, as demonstrated below:

$$\begin{bmatrix} R(0) & R(1)\ldots & R(p-2) & R(p-1) \\ R(1) & R(0)\ldots & . & R(p-2) \\ . & R(1)\ldots & . & . \\ R(p-2) & . & R(0) & R(1) \\ R(p-1) & R(p-2) & R(1) & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \cdot \\ \cdot \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ . \\ . \\ R(p) \end{bmatrix} \qquad \textit{2. 18}$$

The $\boldsymbol{p}$ x $\boldsymbol{p}$ matrix of autocorrelation values is a symmetric and positive definitive matrix within which each descending diagonal from left to right is contestant, also known as a Toeplitz. This means it can be solved efficiently through the Levinson-Durbin (L-D) recursion through a process of two steps, initialisation and recursion, as outlined below[68] .

Compute the error energy associated with the order- $\boldsymbol{i}$ solution

$$\boldsymbol{\varepsilon^{(i)} = \left(i - k_i{}^2\right)\epsilon^{(i-1)}} \qquad \textit{2. 19}$$

$\boldsymbol{\varepsilon^{(i)}}$is the total squared error for a predictor of order i and k, is the PARCOR coefficient. The end solution is given as follows:

$$\boldsymbol{a^i = a_i^{(p)} 1 \leq i \leq p} \qquad \textit{2. 20}$$

This process of computing LPC coefficients is called the autocorrelation method because of the presence of this operation within the equations. Although there are other methods to compute LPC parameters, for instance the covariance method, the autocorrelation method is the most commonly used because of its computational efficiency and inherent stability.

The LPC filter magnitude response represents the spectral envelope of speech magnitude spectrum of each frame, wherein the choice of $\boldsymbol{p}$ effects the representation of the spectral envelope of the speech spectrum with accuracy. With increase in p, the LPC filter response gives a more accurate speech spectral envelope, however the increase in $\boldsymbol{p}$ also results in the need for more memory and computation (Figure 2.11). Therefore, a happy medium must be established between computational and memory requirements, and spectral accuracy. The authors of [67] suggest that a total of fs (sampling frequency in kHz) poles is adequate to represent its contribution to the spectrum. This is because speech spectrum can generally be represented as having an average density of 2 poles per kHz because of vocal tract contribution [59, 69]. It has also been suggested that approximately 2-4 further poles are needed to adequately represent source excitation spectrum and lip radiation effects in the production of the speech signal [70]Therefore, the optimum choice value for p is 18-20 for a 16kHz sampling frequency. For the purposes of this study, a value of 20 is used.

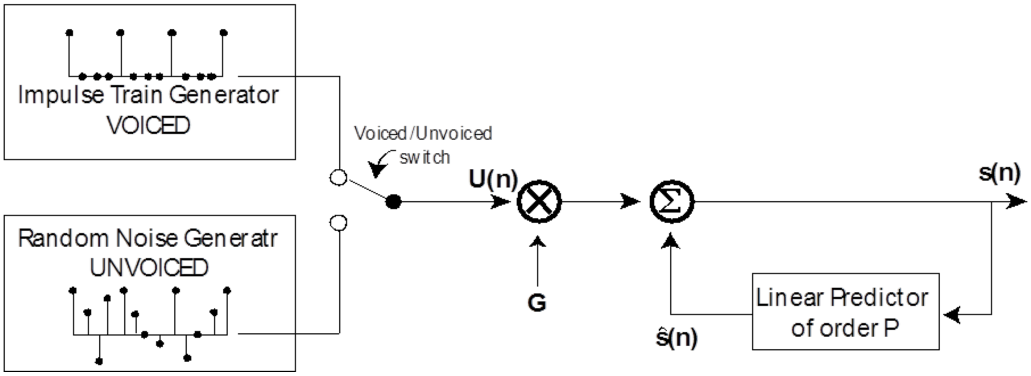

Figure 2.11 LPC model of speech. [72]

2.2.6.3. Linear prediction cepstral coefficients

LPC presents a suitable model of speech production, however in this section, an alternative model is explored, wherein an impulse train and random noise generators drive the vocal tract filters. LPCC is used to separate two components of the speech signal that are convoluted in the time domain. As mentioned before the vocal tract filter is driven by the excitation source, which means the short-term spectrum of speech consists of both slowly varying envelopes corresponding to the vocal tract filter, and rapidly varying envelopes corresponding to the periodic excitations and harmonics. The first component corresponds to unvoiced speech, whereas the latter corresponds to voiced speech. It is safe to state that the observed sequence of speech samples is the result of the convolution of excitation and vocal tract impulse response in the time domain [43].

The main aim of Cepstral analysis is to separate the properties/parameters of the excitation and vocal tract. This is achieved by transforming the two components to a summation using an algorithmic operation in the frequency domain. In frequency domain, the convolution is transformed into a multiplication, which subsequently is transformed into a summation of the log-frequency domain. Transformation back to a time-like domain results in a cepstrum (an anagram of spectrum) which represents the excitation and vocal tract components separately. The vocal tract part of the cepstrum appears at low quefrency (an anagram of frequency), whereas the excitation part appears at high quefrency. The liftering process (anagram of filtering) is then employed to separate the two components by truncating the series of cepstral coefficients obtained.

The cepstral analysis process for a discrete signal is demonstrated in Figure 2.21. Here the logarithm operation is applied to the modulus of $\boldsymbol{S(\omega)}$. This gives a real cepstrum as defined by equation 2.24, which is the most popular form of cepstrum in speech processing applications. However, if the log operation is applied to the complete sequence $\boldsymbol{S(\omega)}$, a complex cepstrum is formed [31].

$$\boldsymbol{c_n = \frac{1}{2\pi}\int_{-K}^{+k} log|S(\omega)|e^{j\omega n}d\omega \quad 0 \le n \le N-1} \qquad \textit{2. 21}$$

$$\boldsymbol{c_n = \frac{1}{N}\sum_{k-0}^{N-1} log\,|X(K)\,|cos\left(\frac{2\pi kn}{N}\right) 0 \le n \le N-1} \qquad \textit{2. 22}$$

The assumption that the logarithm function is real and even means that for discrete cases, the cepstrum may be acquired by using the DCT for the IDFT operation [71]:

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n+\frac{1}{2}\right)\left(k+\frac{1}{2}\right)\right] \quad for\ k$$

$$= 0,1,2,\dots(N-1)$$

2. 23

Alternatively, cepstral coefficients can also be acquired from the LPC coefficients, where the power series expansion $\boldsymbol{z^{-1}}$of the logarithm transfer function of the LPC model is used, as demonstrated in equation 2.27 [59]

$$\boldsymbol{log\, H(z) = C(z) = \sum_{k=1}^{+\infty} c_k\, z^{-1}}$$

2. 24

The relationship between $\boldsymbol{c_k}$ and the LPC coefficients $\boldsymbol{a_k}$ is found by taking the products of both sides of the equation above with respect to $\boldsymbol{z^{-1}}$and equating equal powers of $\boldsymbol{z^{-1}}$. The recursive relationship that results is outlined below where ak represents LPC coefficients and p is the LPC order [59].

$$\boldsymbol{c_1 = -a_1}$$

2. 25

$$\boldsymbol{c_n = -a_1 - \sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right) a_k c_{n-k} \;\; n = 2,3,\dots,p}$$

2. 26

$$\boldsymbol{c_n = -\sum_{i=1}^{n-1}\left(1-\frac{i}{n}\right) a_k c_{n-k} n > p}$$

2. 27

The recursion demonstrated above suggests that the sequence of the cepstral parameters is of infinite length; however, in practice, only the first p terms are used [67]. This cepstrum is referred to as an LPC derived cepstrum (LPCC).

From the review above, it is shown that in automatic speech recognition, MFCC performs better than LPCC in terms of efficiency and accuracy despite the fact that the MFCC algorithm requires more computation than the LPCC algorithm. For instance, in comparing MFCC and LPCC in the recognition of stuttered speech, it was found that the MFCC algorithm slightly outperforms the LPCC algorithm [73]. Furthermore, it is also shown that MFCC performs better than both LPC and LPCC in terms of recognition rate in noisy environments, with the recognition rate of MFCC in such cases rising to as much as 93.33% compared to LPCC's 80% [74]. Moreover, while LPCC is relatively more robust than MFCC under conditions of speaker variability, MFCC shows more robustness than LPCC and LPC under conditions of environmental noise [72][75]. Thus, since the current thesis seeks to enhance voice recognition performance when audio data is obtained under uncontrolled environments, the study focuses on the MFCC algorithm. This method is tested in chapter 3, which focuses on the effects of environmental noise on the accuracy of open-set speaker identification. The following chapter will review open-set speaker identification.

2.3. Open-set speaker identification

The process of open-set text-independent speaker identification (OSTI-SI) for the baseline speaker classifier Gaussian mixture model (discussed further in Section 2.4.2) is summarised in

Figure 2.12. As shown in the Figure, the process involves two stages; identification as illustrated in stage one, and verification as illustrated in stage two. The registered speakers are represented using their corresponding statistical model descriptions $\boldsymbol{\lambda}_{1,}\boldsymbol{\lambda}_{2,} \dots \boldsymbol{\lambda}_{N,}$ with *N* being the number of speakers in the set. Each reference model is built using the short-term spectral features extracted from the training utterances spoken by the corresponding registered speaker.

On the basis of such speaker modelling, the process of speaker identification in the open-set mode (Figure 2.12.) is stated as [11]:

$$\max_{1\le n\le N}\{p(\mathbf{O}|\boldsymbol{\lambda}_N)\} \gtrless \theta \ \rightarrow \mathbf{O} \in \begin{cases} \lambda_i, i = \arg\max_{1\le n\le N}\{p(\mathbf{O}|\lambda_N) \\ \boldsymbol{unknown\ speaker\ model} \end{cases} \qquad \textbf{\textit{2. 28}}$$

Where *θ* is the pre-determined threshold, while, $\mathbf{O}$ denotes the feature vector sequence extracted from the test utterance. is assigned to the speaker model that yields the maximum likelihood over all other speaker models in the registered set. In the case of the maximum likelihood score being greater than the threshold *θ*, the utterance is accepted. Otherwise, it is declared as having originated from an unknown speaker.

It should be noted that in an OSTI-SI scenario, the universal speaker set consists of two subsets of known (registered) speakers and unknown speakers. Within this scenario, there are three possible types of errors as shown in Figure 2.2, and the mentioned errors as follows [73]:

- Open-set Identification Error (OSI-E) or Mislabelling (ML): when $\mathbf{O}_{\boldsymbol{a}}$ belongs to $\boldsymbol{\lambda}_{\boldsymbol{a}}$, but yields the maximum likelihood for another speaker model within the registered set, e.g. $\boldsymbol{\lambda}_{b}$,
- Open-set identification False Rejection (FR): declaring that $\mathbf{O}_{\boldsymbol{a}}$ has originated from an unknown speaker when it actually belongs to $\boldsymbol{\lambda}_{\boldsymbol{a}}$

- Open-set Identification False Acceptance (OSI-FA) occurs when $\mathbf{O}_a$ is assigned to one of the models in the set when it has originated from an unknown speaker.

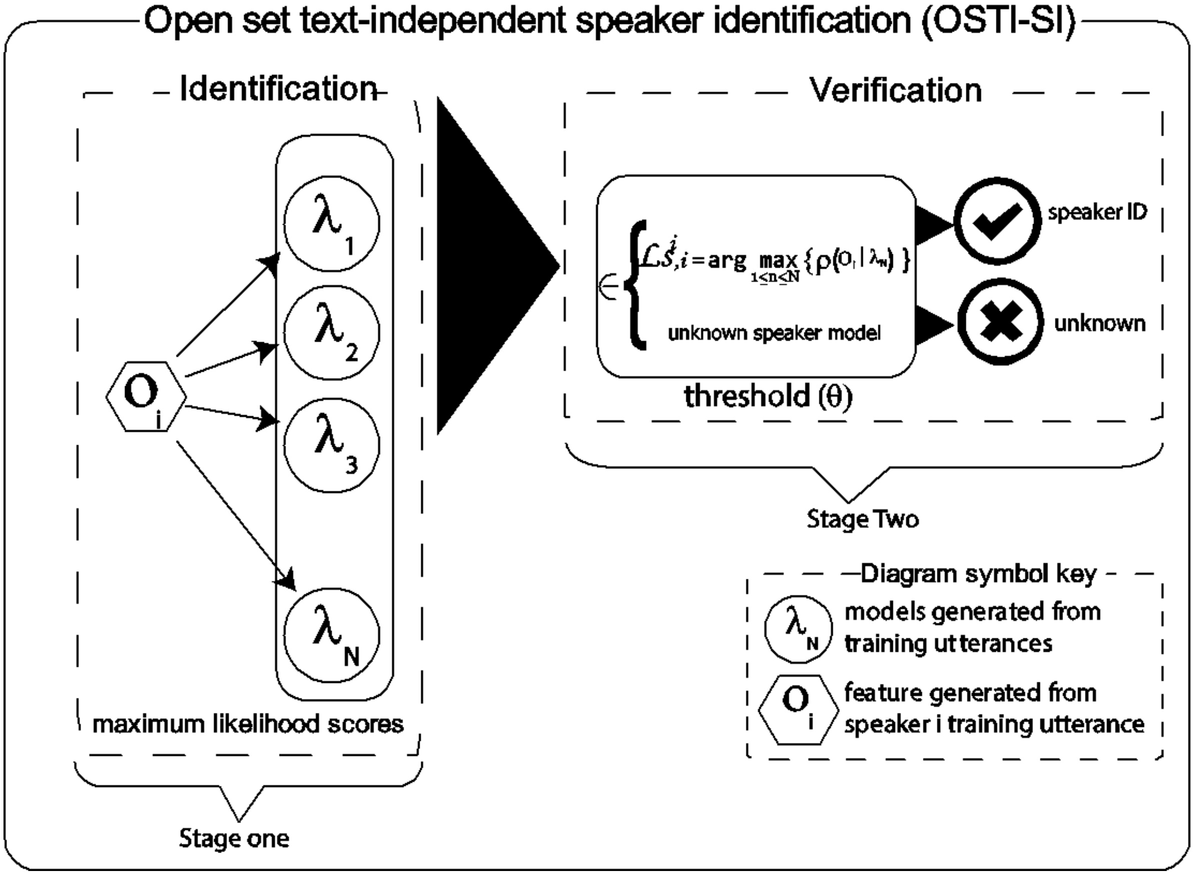
Open set text-independent speaker identification (OSTI-SI)
Identification
λ_1
λ_2
λ_3
λ_N
O_i
maximum likelihood scores
Stage one
Verification
$\in \begin{cases} \mathcal{L}\dot{S}^{i},i = \arg\max_{1 \le n \le N} \{\rho(O_i \mid \lambda_n)\} \\ \text{unknown speaker model} \end{cases}$
threshold (θ)
speaker ID
unknown
Stage Two
Diagram symbol key
models generated from training utterances
feature generated from speaker i training utterance

Figure 2.12 Stages one and two in the process of OSTI-SI

Whilst the identification stage is responsible for ML errors, FA and FR are the consequences of the decision made in the verification stage. An important point to note is that each member of the unknown speakers can be falsely hypothesised as one of the registered speakers, when (s)he achieves a sufficiently high score against one of the registered speaker models. In practice, a key factor affecting the OSTI-SI performance is the size of the registered speakers population [62]. Moreover, the identification decision is expected to become more difficult when considering the challenges proposed in this study. The following scenario considers the hypothetical case where two speakers are enrolled to the system (OSTI-SI) one of which is trained with short, and the other is trained with long reference material. In such a scenario, because of the lack of phonetic content representation of the short speaker model, there is an increased probability that the test material of that particular speaker may more closely match the other speaker model trained with longer reference utterance (within the registered set, which is phonetically much richer). Hence there is increased likelihood of ML error.

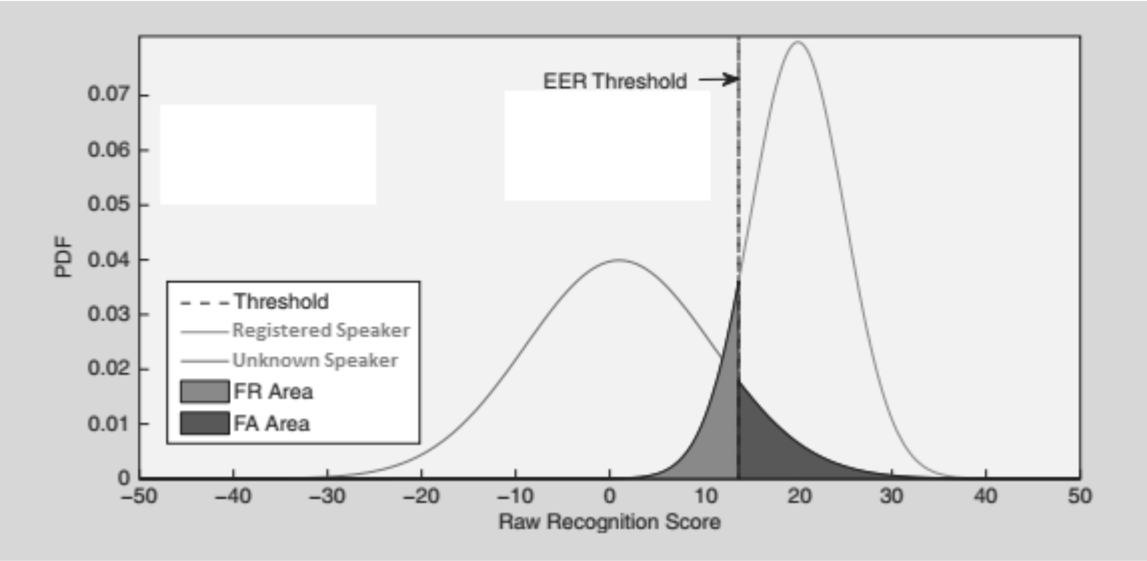

Figure 2.13 Stage two errors in the process of OSTI-SI

It should be noted that an error in the identification stage (ML) would always lead to an overall error regardless of the decision in the verification stage and that this error is unrecoverable [74]. In the case of security applications, such an ML error in the first stage has severe consequences. For example, mislabelling may result in the target of interest being missed. For this reason, it is important to concentrate on approaches for enhancing the reliability of operation in the first stage. As the interest of this study is of open-set text independent speaker recognition, conducting a literature review will assist in gaining further understanding of the current methods and classifiers used to enhance recognition performance in the case of Speaker Identification.

The area of speaker identification has been the subject of investigation and many robust methods and algorithms have been presented in the past, for instance the study[75] includes a scheme for using the fast-scoring method which has been proposed for speaker verification. Furthermore, it provides an evaluation of various score normalisation methods in the proposed OSTI-SI framework, such as Test normalisation (T-norm) and a combination of Test and Zero normalisation (TZ-norm). The dataset used for the experimental investigation was based on NIST SRE2003 1-speaker detection task. They concluded that significant improvements can be achieved if only a single mixture is used in the fast-scoring technique. Furthermore, it has been shown experimentally that, unlike in speaker verification, only the best-scoring mixture needs to be included in the likelihood for achieving the best performance. Additionally, the study has confirmed the significance of score normalisation as a valuable component in OSTI-SI. It has been shown that, whilst Z-norm enhances the performance accuracy considerably, further improvement over the Z-norm performance can be achieved using either of the Unconstrained Cohort Normalisation techniques, T-norm or TZ-norm. In the study [15] the focus was to improve the identification rate of SI. They worked at feature level where the performance of MFCC technique was evaluated in a quiet environment. A speaker database containing 30 male and 30 female speakers was created. Two separate experiments were conducted for the performance evaluation of MFCC technique when applied to K means clustering. In the first case the speech features were directly matched. In the second case a VQ codebook was created by clustering the training features of these 60 speakers. They found that the choice of number of clusters plays a vital role in the recognition rate. The failure rate of speaker recognition in first case was found to be 10% while in the second case was found to be 14%. The percentage rise in mean distortion for all the five test cases for the clustered case was found to be 13.18%. This gives intuitive ideas regarding the choice of the ideal number of clusters for a better recognition. Also in this study [76] the focus was on feature level, the use of three features to improve the performance of OSTI-SI were proposed. The new method called LPCC and F0(the reverse of MFCC), extract more speaker-dependent information than the traditional MFCC. These features were then combined optimally to give the final score. The TIMIT dataset was used and they recorded significant improvement in performance. Some further study was conducted by [77], they believed verification methods are variable and use different types of features, but each system alone does not provide satisfactory results. For this reason a comparison of different features and methods for score fusion for an independent speaker verification application was implemented. Several types of spectral features were used as speaker data. The scores obtained with these types of features were fused with combination methods (as: mean, sum, max, min, weighted sum) and classification methods (as: SVM, linear discriminant). These methods' performances have been compared using a text independent speaker verification method with GMM-UBM, by using a clear speech database for Romanian language. They concluded the best combination method (weighted sum) and achieved an EER two times smaller than the best ones obtained by a baseline system. Further work included the study [78], the use of discriminative training scheme based on maximum mutual information (MMI) criteria for speaker recognition was considered. It was believed discriminative training has been limited to training GMM with a small number of Gaussian components. They present the discriminative training on both target and cohort speaker models specifically for OS-SI problem. Experiment results showed that notable performance improvement was obtained from MMI discriminative (MMI-DISC) approach, as compared to the classic GMM-maximum a posterior (MAP). A new approach was presented by the study [79], employing additional information which is dialect detection with a novel parameterization of the speech to improve the task of speaker identification. The proposed system demonstrated the use of different kernels function of Support Vector Machines (SVM) improves speaker recognition with speakers taken from TIMIT database. Since dialect is among the important and complicated aspects of speaker variability, they demonstrated in this work that it can establish useful indicators to specify a speaker's identity. This method has focused on the formulation of a regional system based on SVM with a novel parameterization. This new technique improves the system performance and succeeds to obtain better performance in EER. New scoring techniques were considered by the study [80], two scoring techniques were compared, SVM and fast scoring. Both techniques were based on a cosine kernel applied in the total factor space, where vectors are extracted using a simple factor analysis. The best results were obtained using fast scoring when LDA and WCCN combinations are applied in order to compensate for the channel effects. The use of the cosine kernel as a decision score makes the decision process faster and less complex. Further investigation into improving the performance of OSTI-SI was in the study [81].They believed that speaker identification systems focus on the speech features used for modelling the speakers without any concern for the speech being input to the system. Knowing how reliable the input speech information is can be very important and useful. The idea of SID-usable speech was to identify and extract those portions of corrupted input speech, which renders the speech data more reliable. For this reason they presented what is called SID-usable speech. Here the speaker identification system itself is used to determine those speech frames that are usable for accurate speaker identification. Two novel approaches to identify SID-usable speech frames were presented, which resulted in 78% and 72% correct detection of SID-usable speech. The experimental results show that SID performance can be quantified by comparing the amount of speech data required for correct identification. The amount of SID-usable speech was approximately 30% less than entire input data without the SID system performance being compromised. Therefore, they concluded that using only SID-usable speech improves the speaker identification performance. Other feature extraction techniques were considered such as the study [82] wherein a new feature extraction Neurogram technique was adopted to enhance the performance of speaker identification. Neurogram is a 2-D time-frequency representation which was constructed by combining the neural responses (i.e., feature) from 25 auditory nerve (AN) fibres. In this study, the neurogram coefficients were extracted for each speaker to be used as a feature for identification. The average size of the neurogram over three databases (considering all speech signals) was 190 × 25, where the number of frames was 190, and the number of AN fibres was 25. The performance of the proposed method was compared to the identification results of three traditional baseline feature-based methods (MFCC, Frequency domain linear prediction (FDLP) and Frequency cepstral coefficients (GFCC). They claimed neural-response-based metric worked well for both text-dependent and text-independent tasks. The proposed neural feature successfully captured the important distinguishing information about speakers to make the system relatively robust against different types of degradation of the input acoustic signals. The neural feature was extracted from the responses of a physiologically-based model of the auditory periphery. the proposed method was relatively better than the results of most of the existing methods, especially at negative SNRs. Also, the proposed neural feature provided a relatively consistent performance across different types of noise irrespective of the speech materials used. Recent work has mainly been involved in improving the current state of the art i-vectors for instance in the study[83] the emphasis was on identifying an efficient way to implement dimension compactness in total variability space and using cosine distance scoring to predict a fast output score for small size utterance. They claimed that the proposed methodology sufficiently reduces the computation time and works for small size of test utterance. The cosine scoring provides fast predictions about the matching. Further work was conducted to reduce calculation time in [84] and introduces some simplifications to the i-vector speaker recognition systems. I-vector extraction as well as training of the i-vector extractor can be an expensive task both in terms of memory and speed. Under certain assumptions, the formulas for i-vector extraction—also used in i-vector extractor training—can be simplified and lead to a faster and more efficient code. They first assumed that the GMM component alignment is constant across utterances and is given by the UBM GMM weights. They further assumed that the i-vector extractor matrix can be linearly transformed so that its per-Gaussian components are orthogonal. In this study they propose to use Principal component analysis (PCA) and Heteroscedastic linear discriminant analysis (HLDA) to estimate this transform. They claim that they managed to reduce the memory requirements and processing time for the i-vector extractor training so that higher dimensions can be now used while retaining the recognition accuracy. Furthermore, in the i-vector extraction, they managed to reduce the complexity of the algorithm with sacrificing little recognition accuracy, which makes this technique usable in small-scale devices.

Furthermore the new method called intersession compensation and scoring was presented in [85]. This new approach claims to contributes to a better understanding of the session variability characteristics in the total factor space. They presented a set of simple linear and non-linear transformations to remove the session effects and a simple scoring technique based on a statistical classifier. Compared to the baseline and to Linear Discriminant Analysis (LDA) + with in class covariance normalisation (WCCN) +cosine scoring, they claimed that the method they proposed gives the best performances. Furthermore some realistic scenarios was considered such as the work in this study [86] , where they propose a novel approach for noise-robust speaker recognition, where the model of distortions caused by additive and convolutive noises is integrated into the i-vector extraction framework. They adopted Vector Taylor Series (VTS) approximation widely successful in noise robust speech recognition. The model allows for extracting "cleaned-up" i-vectors which can be used in a standard i-vector back end. They evaluate the proposed framework on the PRISM corpus, a NIST-SRE like corpus, where noisy conditions were created by artificially adding babble noises to clean speech segments. Results show that using VTS i-vectors present significant improvements in all noisy conditions compared to a state-of-the art baseline speaker recognition. They further claim that the proposed framework is robust to noise, as improvements are maintained when the system is trained on clean data. Additionally the effect of environmental noise on the identification rate of speaker identification was evaluated by the author [87]. In this study experimental investigations were conducted using a protocol developed for the identification task, based on the NIST speaker recognition evaluation corpus of 2008. In order to closely cover conditions in the considered application areas where users are not expected to cooperate and investigate the identification performance in such scenarios, the speech data is contaminated with a range of real-world noise. It was found that white noise doesn't give a clear representation of environmental noise as it affects all components. Furthermore normalisation techniques played a significant role in improving recognition accuracy of baseline when contaminated with noise (please see chapter 3). The effect of short reference material on the current state of the art was considered in this study[88] . They tried several normalisation techniques to enhance the performance. they investigated how the current selection of factor analysis techniques perform when utterance lengths are significantly reduced. Overall, the current factor analysis approaches have not provided any clear differences in performance for short speech, with the alterative between log likelihood based joint factor analysis (JFA) and Gaussian Probability Linear Discriminant Analysis (GPLDA) offering marginally better performance to LDA + WCCN or SDNAP + WCCN based i-vector systems in lieu of the efficiencies available through operating in the lower-dimensional i-vector space [88]. They concluded all the systems still exhibit performance which declines sharply once utterance lengths fall below 10 seconds. More realistic scenarios were evaluated by the author to identify the effects of varied reference material on the recognition performance of speaker identification [89]. The investigation shows clearly that the current state of the art (i-vector) and the baseline (GMM-UBM) performance are similar in some cases when the reference data is insufficient. They further concluded that the I-vector is more effective when the reference data is varied in duration. Furthermore both classifiers performances drop extremely when the reference and test data is varied please (see chapter 4).

The literature review clearly demonstrates that the area of open set text independent speaker recognition has been the subject of investigation and many robust methods have been presented, what is also clear is that the realistic scenarios presented in this study has not been explored. For this reason the main focus of this study is to identify the performance of

OSTI-SI under realistic conditions and propose novel method in improving its performance.

2.4. Speaker modelling

A speaker's voice becomes known to the system through the process of enrolment. In this process, the feature parameters from the speaker training data are used to construct a speaker model. The speaker model constitutes a unique representation of each registered speaker in the recognition system. In this study, the Gaussian Mixture Model is adopted, which will be used as the baseline system and to represent the current state of the art I-vectors.

2.4.1. Modelling

The representation of the speakers is referred to as speaker models. The speaker model is a reference parametric set for each speaker, generated from the feature vectors as described above [90]. There are two types of modelling methods: deterministic and statistical. Deterministic methods are techniques such as Dynamic Time Warping (DTW) and Vector Quantization (VQ). However, statistical methods include techniques such as Hidden Markov Models (HMM) and Gaussian Mixtures Models (GMM). For text-independent speaker recognition systems, where there is no prior knowledge of what text is stated by the speaker, one of the most successful likelihood functions has been the Gaussian Mixture Model [91]. It is also a common approach in speaker verification, and speaker identification.

2.4.2. Gaussian mixture model

An M component GMM is a weighted combination of M Gaussian Probability Density functions (PDFs) and it is represented by $\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, \dots, M$, where μ_i is the mean vector and Σ_i is the covariance mixture of the i^{th} Gaussian component in the GMM,w_i is the associated weight for the Gaussian component. The weight for the component Gaussians sum to unity and the probability of the GMM producing an observation **o** is[48, 59, 91]:

$$\boldsymbol{P(o|\lambda) = \sum_{i=1}^{M} w_i\, N(o|\mu_i, {\textstyle\sum_i})} \qquad \textbf{\textit{2. 29}}$$

where $\boldsymbol{o}$ is the p-dimensional feature vector, the weight of each of the *M* components, which is constrained by $\sum_i w_i$=1. $N(\boldsymbol{o}|\mu_i, \sum_i)$ is the p-variate Gaussian density function and is given by:

$$\boldsymbol{N(o|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_i|^{\frac{1}{2}}} exp\,\{-\frac{1}{2}(o-\mu_i)\sum_i^{-1}(o-\mu_i)\}\,,} \qquad \textbf{\textit{2. 30}}$$

$$\boldsymbol{i = 1, \dots, M}$$

In 2.30 Σ_i indicates the determinant of the covariance matrix Σ_i. As mentioned previously, PDFs are represented by the set of parameters $\lambda = \{w_i, \mu_i, \Sigma_i\}$, and this set is referred to as the speakers GMM model.

The GMM can have three possible forms of covariance matrix. These matrixes include the nodal, grand and global covariance. In the nodal case one covariance matrix is used for each mixture. In the grand case the mixture is shared amongst all mixture densities, while in the global case, one covariance mixture is shared amongst all speakers in the recognition system.

In this work, cepstral feature parameters are used which means the feature parameters are highly uncorrelated and therefore their covariance are negligibly small. For this reason, the best solution is to implement the GMM with diagonal nodal covariance matrices. The use of one covariance matrix per component density offers a much greater modelling capability in comparison to grand and global covariance. It is also reported in[92] that this form of GMM performs better in text independent speaker recognition system which is the focus of this study.

2.4.2.1. Motivation for the gaussian mixture model

The Gaussian Mixture Model (GMM) is a specific type of Mixture Model (MM). A Mixture model is a probability type which can be used to represent the presence of subpopulations within an overall population, such as block frame discretized windows in a voice stream chosen for analysis. It does not require an observation data set to identify the sub-population to which a specific observation belongs, however. A GMM assumes the data follow a normal, or Gaussian, distribution and may be Bayesian or non-Bayesian in form. A GMM speaker representation is motivated by two interpretations. The first is, the mixture density provides a smooth approximation to the underlying long term distribution of the features, where the features are of parameters extracted from a speech of a given speaker [13]. The second is, each individual Gaussian component in a speaker's dependent GMM represents the spectral structure associated with broad phonetic classes.

2.4.3. Modelling the feature distribution

The statistical representation of a set of feature vectors, generated from the training data for the speaker in question, is referred to as a speaker model. One possible approach to evaluate the speaker model is to use a single (multivariate) Gaussian distribution as shown in the Figure 2.14. In this case, the fundamental assumption is that the super-vectors are independent. A modelling work example is demonstrated in the Figure 2.14. In order to obtain a plot (i.e. the histogram and the Gaussian fit), one dimension of a set of just under a 1000 cepstral vectors from a single speaker is used.

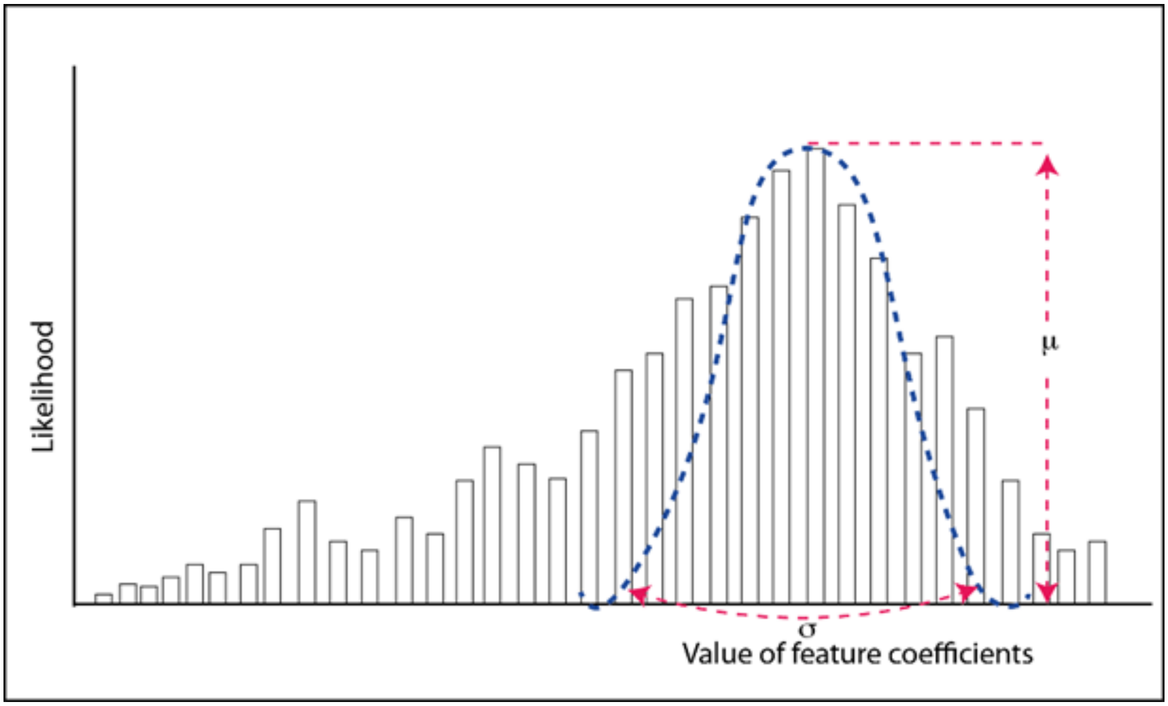

Figure 2.14 Illustrating modelling with a single Gaussian

It is obvious from the figure that an accurate fit cannot be obtained when using a single Gaussian distribution. Also, it is obvious from observing the result in Figure 2.14 that the feature vector distributions are multi-model and a single Gaussian function may not be the best approach to fit it accurately. In order to tackle this problem a finite mixture of Gaussian distributions can be used. Figure 2.15 illustrates such an approach where the same vector sequence as the previous experiment was used, but with multiple Gaussian distributions applied. The figure clearly demonstrates, in this case of four Gaussian densities used, it provides a better representation of the empirical distribution.

The data shown in Figures 2.14 and 2.15 show an even better increase in fit accuracy when the number of Gaussian densities are increased to 32. As has been shown, it appears that a better approximation of the empirical distribution of the feature vectors can be achieved when increasing the number of Gaussian densities. However, having too many Gaussian density mixtures to model a speaker can result in the model becoming highly tuned to the training data. This may cause the analysis to lose the ability to generalize the training data.

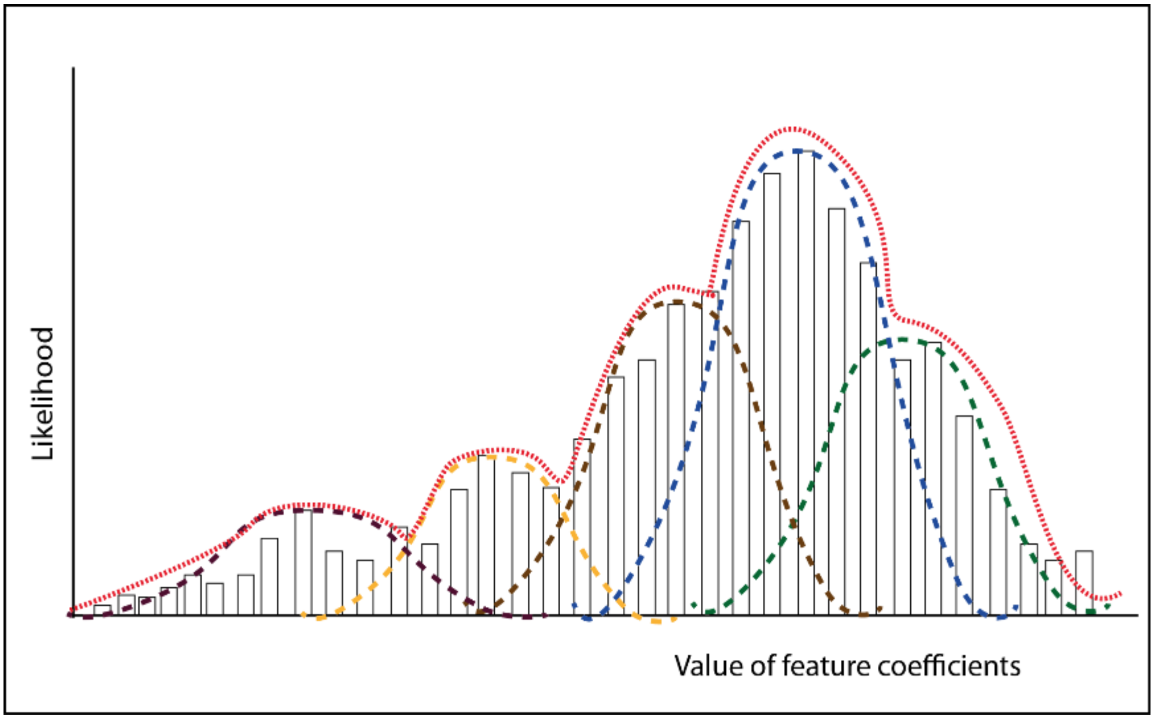

Figure 2.15 Illustrating modelling with mixture of Gaussians

2.4.4. Modelling broad acoustic classes

In the GMM voice modelling technique, each mixture component can be associated with different acoustic classes that represent some broad phonetic events such as vowels and fricatives. The phonetic events, in turn, reflect some general speaker-dependent vocal tract configurations that are useful for characterizing speaker identity [13]. Those mixture parameters represent the average and variation of the vocal tract configurations for the associated acoustic classes. The example shown in Figure 2.15 is an example of a magnitude spectrum associated with voiced and unvoiced acoustic classes.

The parameters of the GMM model are estimated using the Maximum Likelihood (ML) estimation method with a set of training vectors. ML is the process of clustering of the feature vectors into M clusters within the feature space in an unsupervised manner. The information about the origin of each feature vector in terms of acoustic class is not known. The ML estimation is used to generate the speaker model. Computationally, this can be achieved using the Expectation Maximisation (EM) algorithm [93]. The main object of EM is to improve GMM parameter estimates by increasing the probability (on each iteration), such that the model estimated matches the distribution of the training feature vectors [94].

2.4.5. Maximum likelihood estimation

As stated previously, the GMM parameters are estimated by the Maximum Likelihood (ML) estimation. The ML-based GMM model is generated by forming unsupervised clustering of the training feature vectors into a number of clusters within the feature space. The mixture of the GMM is applied to correspond to these feature vector clusters. Since the information related to the origin of each sequence vector acoustic class is unavailable, this makes the procedure an unsupervised one [90]. If acoustic information had been available, the features could have been grouped into associated acoustic classes, so that a mean vector and covariance mixture for each group could be generated. The ML estimation is used to find the models parameters w_i, μ_i and $\sum_i$ that maximize the likelihood function of GMM, given a set of training vectors $\mathbf{O}$ as follows:

$$\boldsymbol{P(O|\lambda) = \prod_{t=1}^{T} p\,(o_i|\lambda)} \qquad \textbf{\textit{2. 31}}$$

Maximizing the above function involves differentiating it with regards to the parameter set $\lambda = \{w_i, \mu_i, \sum_i\}, i = 1, \dots, M$ and then, setting the equation to zero [90] as follows:

$$\boldsymbol{\frac{\partial p(O|\lambda)}{\partial \lambda} = 0} \qquad \textbf{\textit{2. 32}}$$

However, obtaining a solution for the above expression is extremely difficult. Therefore, an iterative process based on the expectation – maximization (EM) algorithm [95] is used instead. The EM algorithm consists of two steps, E-step and the M-step. Its main purpose is to guarantee a monotonic increase in the likelihood function. During the E-step a new estimate of the parameters is computed based on the initial (or current) parameter estimates and the training data. During this step the EM procedure handles the incomplete data problem by estimating feature vector labels using the posteriori probability for acoustic class ι given o_i the observation. Where the probability is $\boldsymbol{p(i|o_t, \lambda)}$

$$\boldsymbol{p(i|o_t, \lambda) = \frac{w_i p_i(o_t)}{p(\ o_t, \lambda)} = \frac{w_i p_i(o_t)}{\sum_{k=1}^{M} w_k p_k(o_t)}} \qquad \textbf{2. 33}$$

A posteriori probability is used to label each training observation. Each mixture components weight, means and covariance is estimation as follows:

$$\boldsymbol{\widehat{w}_i = \frac{1}{T}\sum_{t=1}^{T} p\ (i|o_t, \lambda)} \qquad \textbf{\textit{2. 34}}$$

$$\boldsymbol{\widehat{\mu}_i = \frac{\sum_{t=1}^{T} p\ (i|o_t, \lambda) o_t}{\sum_{t=1}^{T} p\ (i|o_t, \lambda)}} \qquad \textbf{\textit{2. 35}}$$

$$\boldsymbol{\widehat{\sigma_i^2} = \frac{\sum_{t=1}^{T} p\ (i|o_t, \lambda) o_i^2}{\sum_{t=1}^{T} p\ (i|o_t, \lambda)} -\ \widehat{\mu_i^2}} \qquad \textbf{\textit{2. 36}}$$

In the M-step, the current model parameters are replaced by those computed during the E-step. Which serves as the initial model estimation for the iteration. The iteration process is repeated until the likelihood function converges. The likelihood is considered as covered when it is below the pre-set threshold. The idea is to begin with an initial model $\bar{\lambda}$ such that $\boldsymbol{p(o|\bar{\lambda}) \geq\ p(o|\lambda)}$.

2.4.6. Maximum a posteriori estimation

Given a universal background model (UBM) λ_{UBM} and T training observation $O = \{o_3, o_2, .., o_T\}$ extracted from a speech segment, the probabilistic alignment of the training observation into the $\boldsymbol{M}$ mixture component of the UBM is determined first. The probabilistic alignment of the feature vector o_t with mixture I in the world model, is given by the posteriori probability as [91]:

$$\boldsymbol{p(i|o_t) = \frac{w_i N(o_t, \mu_i, \sigma_i^2)}{\sum_{m=1}^{M} w_m N(o_t \mu_m, \sigma_m^2)}} \qquad \textit{2. 37}$$

Where w_x is the weight of the mixture x and $N(o_t, \mu_i, \sigma_i^2)$ the corresponding Gaussian density, evaluated for observation vector o_t

Equation 2.36 is evaluated for all the training observation o_t in a similar manner as the EM algorithm. For each mixture I a new estimation for the weight, means and variance is derived by the respective sufficient statistics, which is computed by [63, 91]:

$$\boldsymbol{c_i = \sum_{t=1}^{T} P\ (i|o_t, \lambda_{UBM})} \qquad \textit{2. 38}$$

$$\boldsymbol{E_i(O) = \frac{1}{n_i} \sum_{t=1}^{T} P\ (i|o_t, \lambda_{UBM}) o_t} \qquad \textit{2. 39}$$

$$\boldsymbol{E_i(O^2) = \frac{1}{n_i} \sum_{t=1}^{T} P\ (i|o_t, \lambda_{UBM}) {o^2}_t} \qquad \textit{2. 40}$$

Where c_i,$E_i(O)$ and $E_i(O^2)$ are the count first and second moment of training feature respectively.

The degree of adaptation depends on the number of training vectors observed in each mixture in the following way.[91]

$$\boldsymbol{k_i^p = \frac{c_i}{c_i + r^p}} \qquad \textit{2. 41}$$

Where the relevance factor for parameter p is r^p. Since the number of training feature vectors observed that are associated with mixture I is c_i, (2.39) will only allow significant parameter updating if the count itself is relatively high. Therefore the relevancy factor controls how much new data should be observed in a mixture before the new parameters begin to replace the world model [91].

2.4.7. Maximum likelihood classification

Having developed a model using a speaker's voice, the nest step is to explore how this model can be used for recognizing the speakers of the input speech segment. Consider a single feature vector $\mathbf{o}$ during the ML classification case. The goal is to select a speaker model λ_s that maximizes the probability of $\boldsymbol{p(\lambda_s|}\mathbf{o})$ where $\mathbf{o}$ is a single feature vector from the sequence of input feature vectors used to identify speaker S. This probability can be written as the following using the Bayes theorem:

2. 42

$$\boldsymbol{P(\lambda_s|o) = \frac{P(\lambda_s|o)p(\lambda_s)}{p(o)}}$$

where $p(\lambda_s|\mathrm{o})$is the a priori probability of speaker$S$, and $p(\mathrm{o})$is the unconditional probability of an observation, o, being produced by a specific speaker. This assumes the system is registered with more than one speaker (N registered speakers). In order to classify o as coming from speaker S the following condition must be met:

2. 43

$$\boldsymbol{S = \underset{1\leq n\leq N}{argmax}\, P(\lambda_n|o) = \underset{1\leq n\leq N}{argmax}\{\frac{P(\lambda_n|o)p(\lambda_n)}{p(o)}\}}$$

If all speakers have the same a priori probability$P(\lambda_n)$, the above equation can be simplified as the following:

2. 44

$$\boldsymbol{S = \underset{1\leq n\leq N}{argmax}\{P(\lambda_n|o)\}}$$

The product in equation 2.43 is known as the likelihood function for λ_n and it is represented by $\ell(\lambda)$. For convenience, this term is often evaluated in the log domain and is termed the log likelihood function$L(\lambda)$. As given bellow:

$$\boldsymbol{LLR_{avg}(O, \lambda_{target}, \lambda_{UBM}) = \frac{1}{T}\sum_{t=1}^{T}\{log\, p(o_t, \lambda_{target}) - log\, p\,(o_t, \lambda_{UBM})\}} \qquad \textbf{2. 45}$$

An alternative to the decoupled GMM is to use an adapted GMM-UBM, which is a high order GMM trained on a large quantity of speech, which has been obtained from a wide sample of speaker population of interest and is designed to capture the general form of the speaker model [96]. The dominant approach to background modelling is to use a single speaker independent GMM to represent $p(o|\lambda_{UBM})$. Using the GMM as a likelihood function, the UBM model is typically a large GMM trained to represent the speaker-independent distribution of features [91]. Practice has shown it is advantageous to train the universal background models with 50 % female and 50 % male speakers.

In practice, the process generally follows these steps:

Step one

A Universal Background Model is produced using the EM algorithm, which is an estimation of a large amount of speaker samples, typically with a size of 1024 or 2048 Gaussian components. The model is produced from a sample set which is constructed

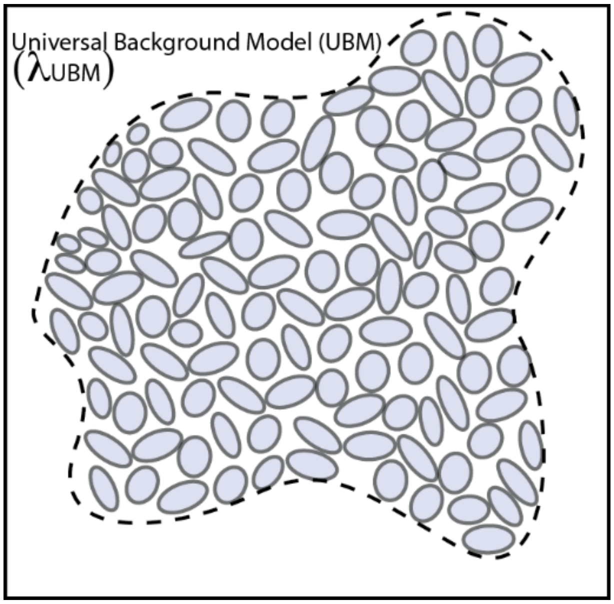

Figure 2.16 Illustrating the UBM world model generated by the EM algorithm [47]

With half of the speaker samples from male speakers and the other half is from female speakers.

Step two

Given the UBM produced in the first step and the training vector assignable to the speaker to be enrolled to the system (feature vectors produced during the feature extraction stage) $=$ $o_1, o_2, \dots, o_T$, the probabilistic alignment of the training vector into the UBM mixture components is determined as shown in Figure 2.16.The probabilistic alignment of the feature vector o_t with mixture I in the world model, is given by the posteriori probability as [91]:

$$\boldsymbol{p(i|o_t) = \frac{w_i N(o_t, \mu_i, \sigma_i^2)}{\sum_{m=1}^{M} w_m N(o_t \mu_m, \sigma_m^2)}} \qquad \textbf{2. 46}$$

Where w_x is the weight of the mixture x and $N(o_t, \mu_i, \sigma_i^2)$ the corresponding *Gaussian density, evaluated for observation vector* o_t

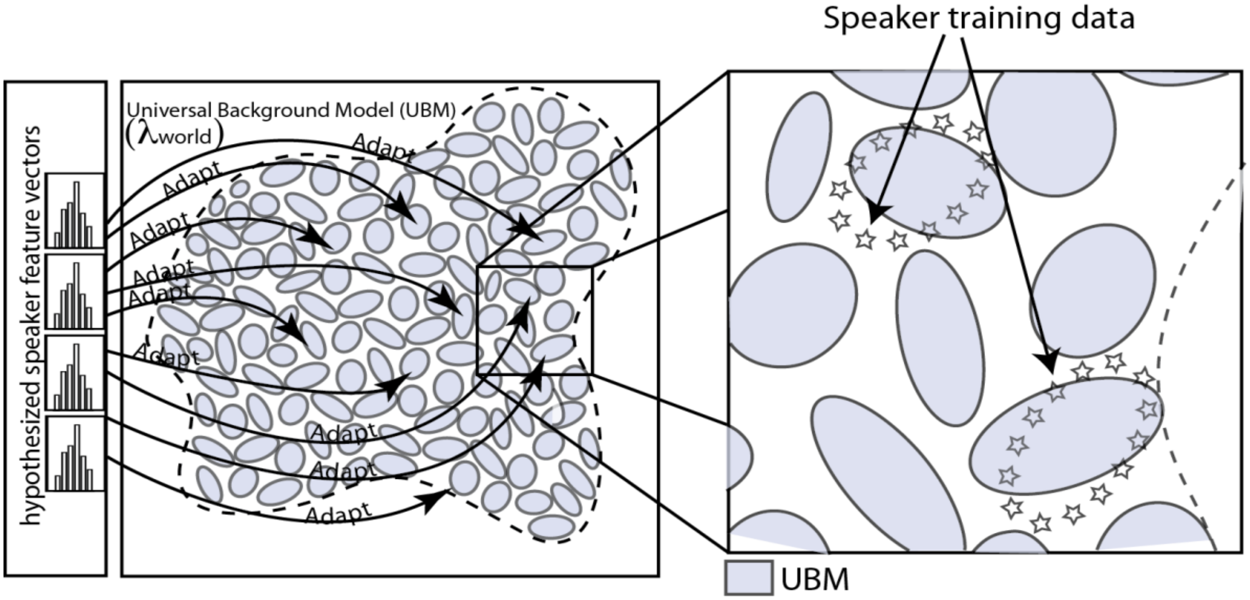

Figure 2.17 Illustrating the training vectors (’s) are probabilistically mapped into the UBM mixtures [47]

Step three

A GMM speaker model is obtained using MAP adaptation technique. The new sufficient

Fortuna statistics from the training data are used to update the old UBM sufficient statistics for

mixture. Evaluation of the training observations o_t, similar to EM algorithm, for each mixture i a new estimation for the weight, means and variance is derived by the respective sufficient statistics [91] as shown in Figure 2.17. The figure shows how the adapted mixture parameters are derived by using the new data statistics with the UBM mixture parameters. The adaptation is data dependent, thus the UBM mixture parameters are adapted by different amounts.

Step four: Testing stage

In the recognition mode, the MAP-adapted model and the UBM are coupled. The match score depends on both the target model (λ_{target}) and the background model (λ_{UBM}) via the average log likelihood ratio as given by the following computation [97, 98]:

$$\boldsymbol{LLR_{avg}\left(O,\lambda_{target},\lambda_{UBM}\right) = \frac{1}{T}\sum_{t=1}^{T}\{log\,p(o_t,\lambda_{target}) - log\,p\,(o_t,\lambda_{UBM})\}} \qquad \textbf{\textit{2. 47}}$$

This essentially measures the difference of the target and background models in generating the observations $\chi = \{x_{1,}x_{2,} \dots, x_t\}$

2.4.8. Weighted bilateral scoring (WBS)

The GMM-UBM technique, which has been one of the dominating approaches in the field of speaker recognition for the past two decades, [99, 100] is considered in the experimental part of this study as the baseline. Weighted bilateral scoring (WBS)[101] provides an extension of this traditional approach. In the context of this study, its potential benefit is related to a scenario within the GMM-UBM paradigm, where the training utterance (utterance x) from a speaker is too short, whereas the testing utterance (utterance y) from the same speaker is considerably longer. Essentially, the weighted bilateral scoring approach solves the problem of lack of reciprocity between two different speakers in open-set speaker identification, in which test utterances tend to be shorter compared to training utterances [101]. To solve this problem the weighted bilateral scoring approach arrives at a final identification score on the basis of weighted combinations between independently normalized reverse and forward scores [101]. In this case, matching an utterance y against the poorly adapted model obtained using an utterance x is likely to yield a low score [102]. However, bilateral scoring involves combining the above forward score with a reverse score obtained by matching utterance x against the richer model obtained by using an utterance y. It should be further emphasised that fusing the reverse score with the traditional GMM-UBM forward score can be specifically beneficial for the real-world applications, which are likely to involve reference and test speech data of varied lengths [101, 103, 104].

In the GMM-UBM paradigm, the framework for weighted bilateral scoring can be summarised as follows.

$$\boldsymbol{L_i^{forward}} = \boldsymbol{log}(\boldsymbol{p}(\boldsymbol{O^u}|\boldsymbol{\lambda_i^k})) - \boldsymbol{log}(\boldsymbol{p}(\boldsymbol{O^u}|\boldsymbol{\lambda_{UBM}})). \tag{2. 48}$$

$$\boldsymbol{L_i^{reverse}} = \boldsymbol{log}(\boldsymbol{p}(\boldsymbol{O_i^k}|\boldsymbol{\lambda^u})) - \boldsymbol{log}\left(\boldsymbol{p}(\boldsymbol{O_i^k}|\boldsymbol{\lambda_{UBM}})\right). \tag{2. 49}$$

$$\boldsymbol{L_i^{bilateral}}(\boldsymbol{f}) = (\boldsymbol{1}-\boldsymbol{f})\boldsymbol{L_i^{forward}} + \boldsymbol{f}\boldsymbol{L_i^{reverse}}. \tag{2. 50}$$

In the above expressions,$\boldsymbol{O}_i^{\mathrm{k}}$and $\boldsymbol{O}^{\mathrm{u}}$ are the feature sequences for the i-th (known) target speaker and the (unknown) speaker of the test utterance respectively. λ_i^{k} and λ^{u} are the corresponding adapted models, and f is the weighting factor with a range of 0 to 1.

2.4.9. Joint factor analysis

Recent advances in reducing dimensionality have been involved in the techniques developed by the National Institute of Standards and Technology (NIST) for instance Joint Factor Analysis (JFA) which is an extension to the GMM-UBM system. JFA may be used to address the complexity of the utterance and lower dimensionality in the model towards the issue with variability in a speaker utterance. It assumes that most of the variance contained in the session-dependent GMM supervector may be accounted for by a small number of hidden variables, which can be classified as either speaker-sourced or channel-sourced factors [105]. JFA is then used to analyse those two channels by combining the three MAP (classical, eigen voice, and eigen channel): finding two separate subspaces representing these channels [105] as shown in Figure 2.19. This technique has shown a high degree of success for solving or simplifying the channel variability problem, successfully separating and processing channel data related to emotionality in the speech, and improving accuracy of classification.

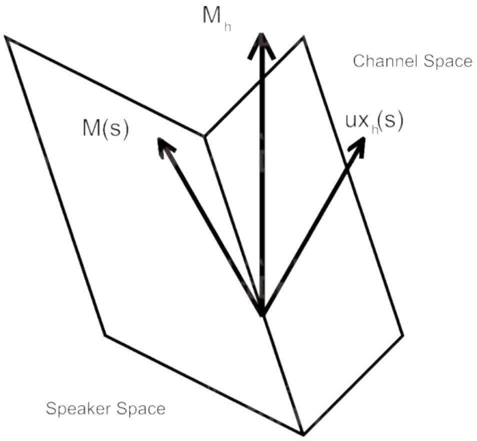

Figure 2.18 Illustration of the JFA super-vector space[106]

JFA was found by Kenny 2004 [107] and it's formulated by combining both eigen voice and eigen channel together, which is accomplished by MAP adaptation for a single model. This model assumes that both speaker and channel variability lie in a lower dimensional sub space of the GMM supervector space. These subspaces are spanned by the matrix V and U. The model assumes for a randomly chosen utterance obtained from a speaker S and session H, that its GMM supervector space can be represented by

$$\boldsymbol{M_h(s) = m + v^* \, y(s) + \, u^* x_h \, (s)} \qquad \textbf{\textit{2. 51}}$$

Where h is the certain utterance of speaker s, m is the speaker- and channel- independent super vector, $m + v^* \, y \, (s)$ describes the part of the super-vector affected by the emotion and the content of speaker v is called speaker space. $y(s)$ is the speaker factor. $u^* \; x_h(s)$ describes the part of the supervector affected by the channel, u is called channel space, $x_h(s)$ is the channel factor. $y(s)$ and $x_h(s)$ are assumed to be independent from each other and normally distributed.

2.4.10. The i-vector total variability space

The extension to JFA is the techniques also developed by the National Institute of Standards and Technology (NIST): which is the total variability space I-Vector and currently it is the state of art classifier for speaker identification. The state-of-the-art i-vector analysis builds on the simplifications of the JFA analysis to reduce dimensionality even further: this increases classification accuracy even more. I-vector analysis framework provides a compact representation of an utterance as a low-dimensional vector, constituting a compression of the utterance which folds into itself the components of the GMM-generated supervector. The i-vector approach trains on one space: the "total variability space", as defined by Dehak, et al.

[90]. For their model, they proposed describing the utterance as a single space that contains and describes the two variabilities of JFA; they named it the ‘total variability space’. Thus, they may be thought of as a kind of JFA modification.

Once the Gaussian mixture model (GMM) has been applied an utterance and the data has been processed to produce the super-vector, as shown in Figure 2.20, the i-vector methodology may be used to simplify it further into what may be described as a compact form with lower dimensionality.

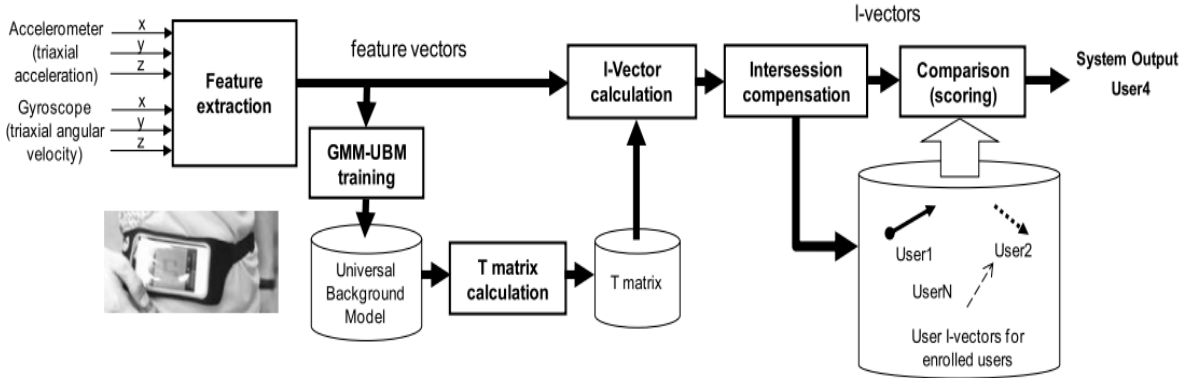

Figure 2.19 I-Vector system architecture

This compaction of the super-vector may be used to solve the following problems with:

- how to directly affect construction of a fixed-sized vector sample, so that a comparison between any pair of sound documents may be made using methods such as cosine similarity or Euclidean distance evaluation, and
- how to eliminate external noise and distortions within a sound document or utterance, as well as compensate for session or channel variance due to background noise, emotional content, or poor

sound volume, so that speaker characteristics may be preserved and minimize issues around voice sample training.

The i-vector approach [12, 90] is related to the GMM-UBM technique, and represents a kind of simplification and compression of the super-vector result. Each i-vector can be regarded as a compact representation of an adapted GMM. To this end, a matrix $\boldsymbol{T}$ called the Total Variability Matrix, or TVM as explained in Figure 2.21, is computed from a large background corpus. The name, 'Total Variability Matrix' refers to the fact that in i-vector space, speaker-specific information is contained within it, together with intra-speaker variability. This matrix $\boldsymbol{T}$ defines a transformation of GMM Gaussian mean super-vectors to the lower-dimensional i-vector space and is described by the following equation:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{T}\boldsymbol{w} \qquad \textbf{\textit{2. 52}}$$

Here, $\boldsymbol{M}$ is the means supervector corresponding to the speech utterance, $\boldsymbol{m}$ is the UBM supervector and $\boldsymbol{w}$ is a standard-normally distributed latent variable of the dimension chosen for the i-vector space. The i-vector $\boldsymbol{w}$ that represents the speech utterance is computed as the MAP estimate of $\boldsymbol{x}$ [108]. As noted above, the total variability matrix $\boldsymbol{T}$ is computed as a Maximum Likelihood estimate from a background corpus. This corpus should be sufficiently large and representative of the speech conditions encountered in relevant applications.

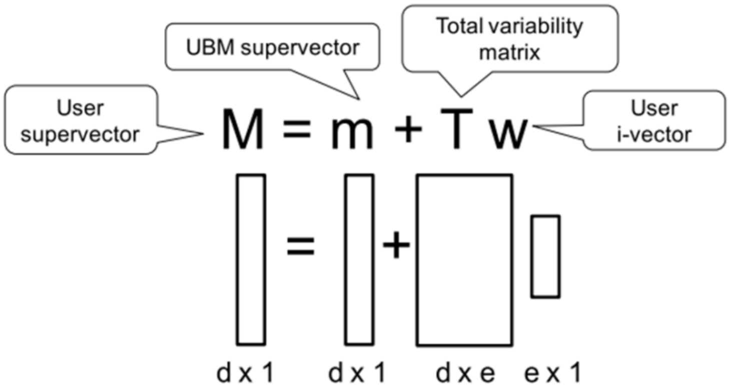

Figure 2.20 Total variability space representation

Having obtained a T matrix, the next step is to extract an i-vector from a sequence of frames. The i-vector w is a hidden variable, which can be defined by its posterior distribution conditioned to the Baum-Welch statistic [12, 109], for a given utterance. This posterior distribution is a Gaussian distribution and the mean of this distribution corresponds exactly to the target i-vector. The Baum-Welch statistics are extracted using the UBM. Suppose there is a sequence of L frames $\{y_1, y_2 \ldots . y_l$ and a UBM, G, composed of C mixture components and defined in some feature space of dimension F. The Baum-Welch statistic needs to estimate the i-vector for a given user activity u are obtained by:

$$N_c = \sum_{t=1}^{L} P\,(c|y_t\,, G) \qquad \textbf{\textit{2. 53}}$$

$$F_c = \sum_{t=1}^{L} P\,(c|y_{t,G})y_t \qquad \textbf{\textit{2. 54}}$$

Where $c = 1 \ldots., C$ is the Gaussian index and $P(c|y_{t\ ,} G)$ corresponds to the posterior probability of mixture component generating an i-vector y_t . in order to estimate the i-vector, it is necessary to compute the centralised first-order Baum-Welch statistics based on the UBM mean mixture components:

$$w = \left(1 + T^t \sum_{t=1}^{-1} N\,(u)T\right)^{-1} .T^t \sum_{t=1}^{-1} FF\,(u) \qquad \textbf{\textit{2. 55}}$$

$N(u)$ is a diagonal matrix of dimension $d \times d$, (d is the multiplication of number of Gaussians, C, by dimension of every Gaussian, F) whose diagonal blocks are $N_c I(\,c = 1 \ldots., C).N_c$ is one scalar per Gaussian that it is replicated F time to compose the matrix $N(u)$. $FF(u)$ is a super vector of dimension $d \times 1$ obtained by concatenating all first-order Baum-Welch statistics FF_c for a given utterance u . Σ is a diagonal covariance matrix of dimension $d \times d$ estimated during factor analysis training [110]. Σ models the residual variability not captured by the total variability matrix T.

In order to identify a user, the scoring module compares an i-vector computed from an input sequence within all i-vectors from the enrolled users, previously calculated and stored in a database. The identified user is the one whose i-vector has the smallest distance to an i-vector extracted from the current frame sequence. The considered distance is the cosine distance.

Unlike traditional GMM-UBM score computation, the i-vector approach is symmetrical in the sense that i-vectors are computed for both the training and test utterances. The comparison of the test i-vector, $\boldsymbol{w}_{test}$, and target i-vector, $\boldsymbol{w}_{target}$, is conducted by using the cosine similarity score (CSS) defined as follows [12]:

$$\boldsymbol{score}(\boldsymbol{w_{target}}, \boldsymbol{w_{test}}) = \frac{< \boldsymbol{w_{target}}, \boldsymbol{w_{test}} >}{\|\boldsymbol{w_{target}}\| \|\boldsymbol{w_{test}}\|} \qquad \textbf{2. 56}$$

Due to the fact that i-vectors represent not only the characteristics of the speaker that is important for the recognition task, but also undesired intra-speaker variability such as channel effects, the suppression n of the latter improves the accuracy of the approach [111].

In order to improve the result, i-vector based recognition systems incorporated different techniques to carryout session compensation in the total factor space. The advantage of applying session compensation in the total factor space is the low dimensions of these vectors, as compared to GMM supervectors. This reduction results in a less expensive computation. One of those techniques is within covariance normalisation, which is discussed in the next section

2.4.10.1. *within-class covariance normalisation* (WCCN)

WCCN has been shown to work well in practice.

To apply this technique, a covariance matrix is computed for each one of the speakers in a background set. Then, the average of all these covariance matrices is calculated to obtain the overall within-class covariance matrix $\boldsymbol{W}$ [12, 111]:

$$\boldsymbol{W} = \frac{\mathbf{1}}{\boldsymbol{S}} \sum_{s=1}^{S} \frac{\mathbf{1}}{\boldsymbol{n_s}} \sum_{i=1}^{n_s} (\boldsymbol{w_i^s} - \overline{\boldsymbol{w_s}})(\boldsymbol{w_i^s} - \overline{\boldsymbol{w_s}})^{\boldsymbol{t}} \qquad \textbf{2. 57}$$

Here, $\overline{\boldsymbol{w_s}}$ is the mean of all i-vectors in the set originating from speaker s $(s = 1, \dots, S)$ and $\boldsymbol{w_i^s}$ is the i^{th} i-vector of speaker s in the background set $(i = 1, \dots, n_s)$. Then, a matrix $\boldsymbol{B}$ is obtained through Cholesky decomposition of the inverse of the within-class covariance matrix; $\boldsymbol{W^{-1} = BB^t}$. Finally, matrix $\boldsymbol{B}$ is multiplied with any i-vector $\boldsymbol{w}$ to calculate its normalized version:

$$\boldsymbol{w_{norm}} = \boldsymbol{B^t w} \qquad \textit{2. 58}$$

In the recognition system, an i-vector system uses a set of low-dimensional total variability (TV) factors to represent each conversation side. Where a GMM-UBM super vector mean matrix, M, is assumed decomposable into speaker independent, speaker dependent, channel dependent, and residual components for standard analysis. Where i-vector analysis is applied, it decomposes s into two factors, M = m + T·x, where m is the UBM mean supervector, T describes the Total Variability Matrix (TVM), and x represents the i-vector. The TVM matrix represents the subspace which encloses the bulk of the speaker-specific information in an utterance, and includes channel variability, which the analyst would like to remove, optimally. To begin to formulate the i-vector, the TVM is trained on the utterance (by using ML estimation - a modified JFA, for example), treating each s conversation component as a separate speaker utterance. By doing so, each component may be treated independently to separate the afore-mentioned speaker variability and channel variabilities. The i-vector, m, treated as a latent variable: in effect, it is the maximum a-posteriori, or MAP point estimate of the standard normal prior [108]. As such, this i-vector extraction methodology successfully and reliably normalizes GMM-UBM super vector covariance's.

2.5. The accumulated error rate

As the concern of this study is the open-set speaker identification. In essence, such a process involves first identifying the speaker model in the database that best matches the given test utterance, and then determining if the test utterance has actually been produced by the speaker associated with the best-matched model (the stages illustrated in Figure 2.12). Whilst, conventionally, the performance of each of these two sub-processes is evaluated independently, it is argued that the use of a measure of performance for the complete process can provide a more useful basis for comparing the effectiveness of different systems. Based on this argument, the accumulated error rate (AER) [112] is adopted in this study. AER was motivated by the approach commonly used in computing DER (Diarisation Error Rate) [113]. It involves a holistic approach to the analysis of the performance in OSTI-SI rather than the independent consideration of the effectiveness in each of the two stages of the process (i.e. identification and verification). For this purpose, the use of three measures of the overall performance in OSTI-SI, i.e. mislabelling (ML), false acceptance (FA) and false rejection (FR) are considered. The integration of these measures has been achieved through the introduction of a metric termed Minimum-Accumulative Error Rate (M-AER). It has been shown the study [112] ML, FA and FR are all influenced by the threshold level adopted in open-set identification, and that it may not be possible to achieve equal rates of these errors using a single threshold level. However, in the study [112] it has been demonstrated that the threshold can be set such as to minimise the Accumulative Error Rate. The Minimum-Accumulative Error Rate provides a valuable basis for comparing the overall effectiveness of different open-set speaker identification systems. The adopted approach efficiently compares performance of open-set speaker identification and presents an analysis of its characteristics.

Open-Set Speaker Identification

To evaluate the full process of OSTI-SI, the recognition accuracy will need to be computed separately for the two stages of identification and verification. Based on this evaluation strategy, just the identification rates are not the optimal approach if it is required to compare the effectiveness of different OSTI-SI techniques. In this case, it is more convenient to adopt a single measure of accuracy that characterises the whole process of open-set identification. Motivated by the approach proposed for the computation of error rate in the diarisation process [92], such a measure for OSTI-SI has been introduced in [112]. This measure is referred to as accumulated error rate (AER) and it allows the evaluation of both stages of OSTI-SI using a single error figure. This is defined as the ratio of the sum of inaccuracies encountered over the total number of identification trials:

$$\boldsymbol{AER(\theta)} = \frac{\boldsymbol{ML(\theta) + FR(\theta) + FA(\theta)}}{\boldsymbol{T}} \tag{2. 59}$$

Here, θ is the threshold adopted in the second stage of the process, $ML(\theta)$is the number of mislabeled (incorrectly identified) clients, $FR(\theta)$ is the number of clients that are falsely rejected, $FA(\theta)$ is the number of impostors that are falsely accepted as clients, and T is the total number of identification trials. This measure of OSTI-SI accuracy is used in the remainder of this paper for a more thorough analysis of the experimental results.

2.6. Summary

Speaker recognition has been a subject of investigation for over two decades, and in that time significant improvements have been made with regards to speaker feature extraction techniques, and modelling techniques as the literature review above shows. The mentioned advancements have led to an increase in recognition accuracy under different conditions where significant speaker data is available. As a result, speaker recognition has been utilised by many applications. Although the literature explores speaker recognition in uncontrolled conditions, (varied speaker reference data, presence of background noise), this is not expansive. With regards to the application of speaker recognition technology in situations of security and surveillance, this thesis further explores the literature available regarding speaker recognition in unfavourable and uncontrolled conditions. The subsequent chapter explores the effect of real world noise on recognition performance, and possible methods of improving recognition performance in such conditions. This is followed by an exploration of varied reference data and its effect on recognition performance.

CHAPTER THREE

3. The Effects of Environmental Noise on the Accuracy of Open-Set Speaker Recognition

3.1. Introduction

3.2. Experimental Investigations

3.2.1. Adopted approach for speaker classification

3.2.2. Speech corpora and protocol for evaluation of OSTI-SI

3.2.3. The conditions of noise contamination

3.2.4. Overview of results for the first stage of OSTI-SI

3.2.5. AER results with telephone-quality data

3.2.6 AER results for speech data contaminated with white noise

3.2.7 AER results for speech data contaminated with car and factory noise

3.3. Summary

3. The Effect of environmental noise on the accuracy of open-set speaker recognition

3.1. Introduction

The presence of environmental noise in speech data has been shown to have a detrimental effect on recognition performance [114]. The aim of this study is to investigate the extent of the effects of environmental noise on the accuracy of open-set speaker recognition. The speaker classification approaches considered for the experiments are (i) the state-of-the-art i-vector method and (ii) the traditional GMM-UBM method supported by score normalisation. To closely cover relevant conditions in the considered application areas and investigate the effects on the identification performance in such scenarios, the speech data used in this study have been contaminated with different types of real-world noise. The work in this chapter has been published in [115].

Recall that open-set speaker identification is the process of determining the correct speaker of a given utterance from a registered population, with the additional requirement to establish if the utterance is not produced by any of the registered speakers. When the speakers are not required to provide utterances of specific texts during identification trials, the process is referred to as open-set, text-independent speaker identification (OSTI-SI). This is the most challenging class of voice biometrics and has a wide range of applications in such areas as audio surveillance, document indexation, and screening [11].

In the last several years, the voice recognition field has given rise to a considerable body of research into enhancing the effectiveness of open-set speaker identification in practical applications. An aspect of this has been related to the introduction of methods for minimising the adverse effects of speech variation due to additive noise [3, 12]. The significance of this arises because, in practice, additive noise causes a mismatch between the test and reference utterances, which in turn can significantly reduce the reliability of OSTI-SI. One of the solutions put forward was the work in [116], the authors worked on creating a free noise corpus from real data. They tested it on the state-of-the-art i-vector classifier and recorded performance of the system under different signal to noise ratios (SNR). They tested two conditions with miss match SNR and matching SNR for enrolment, test and UBM data. The evaluation corpus was of NIST SRC 2010, and the NIST 2008 was used to create a gender dependent 1024 size UBM. They recorded under mismatch SNR EER increased by 13 times when they applied 8db of noise.

Another study in [117] proposed multi-condition training strategy for Gaussian Probabilistic Linear Discriminant Analysis (PLDA) modelling of vector representations of speech utterances. Different real noise conditions were applied as well as white noise to the male speakers of condition interview of the NIST 2010 corpus. They tested the performance of the current state-of-the-art i-vector classifier where they generated a gender dependant UBM of size 2048 on the same data with i-vector dimension of 400. The authors reported that the method proposed showed significant reduction of EER especially in the case of white noise contaminated data. Table 3.1 is a record of EER for the i-vector classifier reported by the authors.

Table 3.1 Experimental results of the study [117], demonstrating the effect of noise on recognition performance

Note log scale of EER is given				
Noise condition	Babble EER	Car EER	Helicopter EER	White noise EER
0	3%	3%	3%	3%
20 db	4%	5%	4%	3.5%
10 db	7%	8%	7%	15%
6 db	12%	15%	10%	23%

The study in [118] presented a new method where the model of distortion caused by attentive and convolution noise is integrated into the i-vector extraction framework. The model is based on a Vector Taylor Series (VTS) approximation widely successful in noise robust speech recognition. The model allows for extracting “cleaned-up” i-vectors, which can be used in a standard i-vector back end. They evaluated the proposed framework on the PRISM corpus, a NIST-SRE like corpus, where noisy conditions were created by artificially adding babble noises to clean speech segments. A 512 diagonal component UBM was trained in a gender dependent fashion on NIST telephone data from the speaker recognition evaluation (SRE) 2004 and 2005. An i-vector extractor of dimension 400 is then trained on a larger set (NIST SRE ’04, ’05, ’06, Switchboard, and Fisher). The dimensionality of i-vectors is further reduced to 200 by LDA, followed by length normalization and PLDA. The results are recorded in Table 3.2 to show the performance of the current state-of-the-art under differed bubble SNR. The proposed method with its intense computational cost did reduce EER by almost 50% when the UBM was trained with noise and clean data and 8db of bubble noise was applied to the evaluation set.

Table 3.2 Experimental results of the study [118], demonstrating the effect of bubble noise on the recognition performance of the current state of the art classifier

Evaluation Condition SNR	UBM trained with clean data, EER	UBM trained on clean and noise data EER
8db	97%	81%
15db	66%	43%
20db	35%	26%
Clean	8.2%	8.6%

In [119], a study concerned with the propagation of uncertainty in the state-of-the-art speaker recognition system. For experiments on the noised NIST SRE 2010 corpus, they used 1024-component diagonal covariance universal background models (UBM), 400 dimension i-Vector trained from Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004, 2005, 2006 SRE enrolment data. The dimensionality was reduced to 200 by LDA, followed by length normalization and PLDA. They contaminated NIST SRE2010 database by artificially adding babble noise at different SNRs assuming that the original NIST SRE10 data is clean. They added babble noise taken from the NOISEX database. In their experiments, they define the oracle uncertainty as the magnitude-squared error between noisy observation features and its clean correspondence. The oracle uncertainty of features was passed into the i-Vector extraction system along with unprocessed noisy features. The results are recorded in Table 3.3 with clear performance improvement demonstrated.

Table 3.3 Experimental results of the study [119], comparing the current state of the art with there proposed methods recognition performance with noise data.

	Clean EER	10db EER	5db EER	0db EER	-5db EER
i-vector	2.1%	5.7%	12.1%	22.1%	35.5%
i-vector-U	2.1%	4.7%	10.2%	20.1%	32.5%

Unlike the previous research mentioned in this chapter, we believe a realistic noise representation is not thoroughly investigated as the majority of the noise used to contaminate were repetitive noise such as helicopter, car and white noise as discussed in the examples presented above. One of the contributions of this study is using more realistic noise audio files with variation in amplitude, to contaminate the clean data which gave more realistic performance of the classifiers especially under uncontrolled environment which was one of the key areas of focus in this study. Further contribution included the conclusion of white noise not being a good representative of real world noise as it contaminates all parts of the audio document because of its repetitive and constant level of amplitude.

The remainder of this chapter is organised as follows. The next section, provides an overview of the approaches to speaker identification adopted in this study. Section 3.2 presents the experimental investigations together with an analysis of the results. Finally, summary of this chapter is discussed in Section 3.3.

3.2. Experimental investigations

3.2.1. Adopted approaches for speaker classification

A total of three speaker recognition techniques are included in the experimental investigations, which are thoroughly discussed in literature review:

(i) GMM-UBM as baseline system, without additional score normalization techniques

(ii) GMM-UBM as in (i), but with TZ-norm [120, 121]

(iii) The i-vector approach, with a dimension of 300 for the total variability space.

3.2.2. Speech corpora and protocol for evaluation of OSTI-SI

The experiments in this chapter are mostly based on the NIST speaker recognition evaluation (SRE) database 2008. An evaluation protocol for open-set identification has been defined on a subset of this telephone-quality database containing 400 registered speakers and 200 out-of-set (unknown) speakers. All the selected material originates from the "short2/short3" core condition [122]. The number of identification trials depends not only on the number of registered speakers and out-of-set impostors, but also on the number of test utterances which varies for different speakers. For this reason, there are a total of 1312 identification trials of enrolled speakers and 627 identification trials of out-of-set impostors. The background corpus used for UBM training is a subset of the NIST speaker recognition evaluation (SRE) database 2005 [123]. This dataset consists of 622 male and 932 female utterances, and the developed gender independent UBM comprises 2048 Gaussian mixture components.

3.2.3. The conditions of noise contamination

Real-world applications of OSTI-SI should be able to cope with a variety of noise types and various degrees of severity of speech signal degradation. Thus, to thoroughly investigate the effect of ambient noise on the OSTI-SI accuracy, a total of seven conditions have been considered in the experiments. The first operating condition is based on the use of the original telephone data from the NIST corpus 2008, without any additional noise contamination. Then, the same speech data have been contaminated with white noise, car and factory noise from the

NOISEX-92 corpus of noise recordings [124]. For each one of these three noise types, two versions of the speech corpus have been generated with signal-to-noise ratios (SNRs) of 5 dB and 15 dB, respectively. It should be noted that for each set of experiments, noise of the same type and level has been added to training and test material, and the same type and level has also been added to the background corpus for TZ normalization. The background set of speech utterances for UBM and i-vector total variability training, on the other hand, is not contaminated with noise. This is because it is considered unfeasible in practice to adapt this large part of the background corpus to changing conditions of the speech data and repeat the computationally demanding processes for UBM and total variability matrix generation each time a new condition is encountered.

3.2.4. Overview of results for the first stage of OSTI-SI

It is worth noting that the open-set, text-independent speaker identification (OSTI-SI) process consists of the two stages of identification and verification. The accuracy of the first stage can be expressed as the identification rate in the closed-set mode. This accuracy rate is essentially computed based on the use of speech data from the registered speaker population (400 samples). In other words, the unknown speakers (out of set) cannot influence the results for this stage.

Table 3.4 gives an overview of the identification (closed-set) rate, for the three classification approaches adopted and the seven noise conditions as defined above. As observed in this table, background noise has a severe effect on the recognition accuracy of the first stage of OSTI-SI. Comparing results of different noise levels at the same approach and the same noise type, unsurprisingly, the drop-in identification rate at the lower SNR of 5 dB is significantly larger than that at 15 dB. Table 3.1 also shows that there are considerable differences between the identification rates for different types of noise (same SNR). As indicated in this table, white noise has the largest effect, followed by factory noise

Table 3.4 Identification (Closed-set) rate at various conditions

		Identification rate		
		GMM-UBM	**GMM-UBM** **TZ-norm**	**I-Vector**
Clean data		39.7%	42.5%	49.5%
White noise contamination	5dB	14.7%	19.8%	27.1%
	15dB	24.6%	29.7%	39.3%
Car noise **Contamination**	5dB	32.1%	37.7%	41%
	15dB	34.8%	40.3%	44%
Factory noise **Contamination**	5dB	22.3%	26.3%	33.8%
	15dB	30%	33.4%	43%

3.2.5. AER results with telephone-quality data

In this part of the experimental investigations, the original training and testing data from the NIST evaluation 2008 is used without any additional noise contamination as baseline result. For this condition, Figure 3.1 shows the accumulated error rate (AER), as defined above, versus the threshold θ. It should be noted that the plots given in this figure are based on applying score range normalisation to the AERs for the three considered methods. This is to facilitate a meaningful comparison of the methods. However, it is noted that in each case, a different threshold still needs to be set in order to achieve the minimum AER. The reason for this is that AER(θ) depends on the method-specific client and impostor score distributions. For the purpose of facilitating the comparison further, an extended procedure for score range normalisation is applied to the plots in Figure 3.2 (and in all subsequent figures), in order to shift the point of minimum AER to the same threshold θ=0.5. Hence each curve is shifted independently, being in the middle of the score range, this value has been chosen to facilitate the graphical representation.

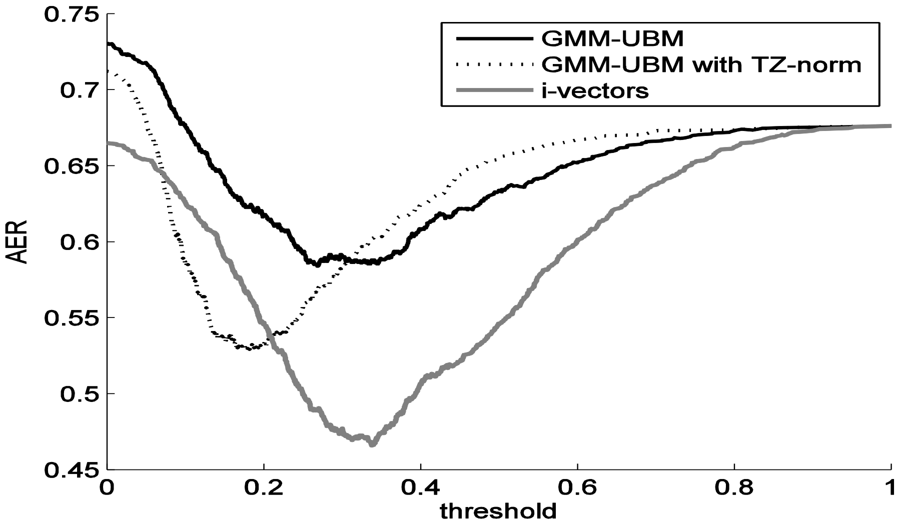

Figure 3.1 Comparison of different methods based on the NIST telephone-quality speech data [81]

The experimental results for the NIST telephone-quality data show that the accuracy in OSTI-SI based on GMM-UBM can be considerably improved by using TZ-normalisation. It is also noted that the highest accuracy in this case is offered by the i-vector approach.

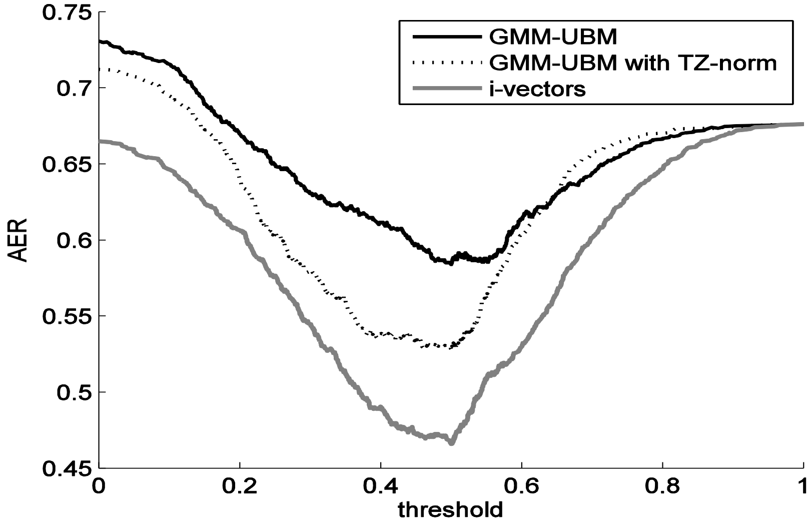

Figure 3.2 Adjusted AER plots for the experiments in the first part of the investigations.

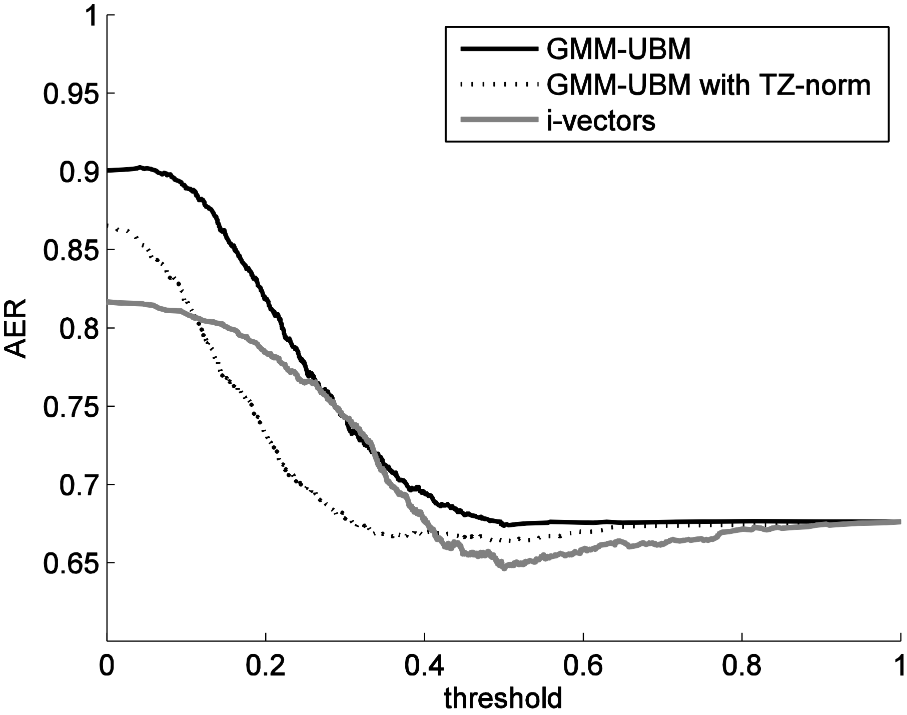

Figure 3.3 Experimental results for different methods based on the use of speech data contaminated with a high level of white noise (SNR = 5 dB).

3.2.6. AER results for speech data contaminated with white noise

The aim of the experiments in this and the following section is to comparatively evaluate the recognition performance of the adopted algorithms for different types and levels of noise in speech signals. The speech data contamination in this part is based on the procedure described in section 3.34. The first part of the investigations in this section is based on using white noise to contaminate speech to achieve signal-to-noise ratios (SNRs) of 5 dB (Figure 3.3) and 15 dB (Figure 3.4).

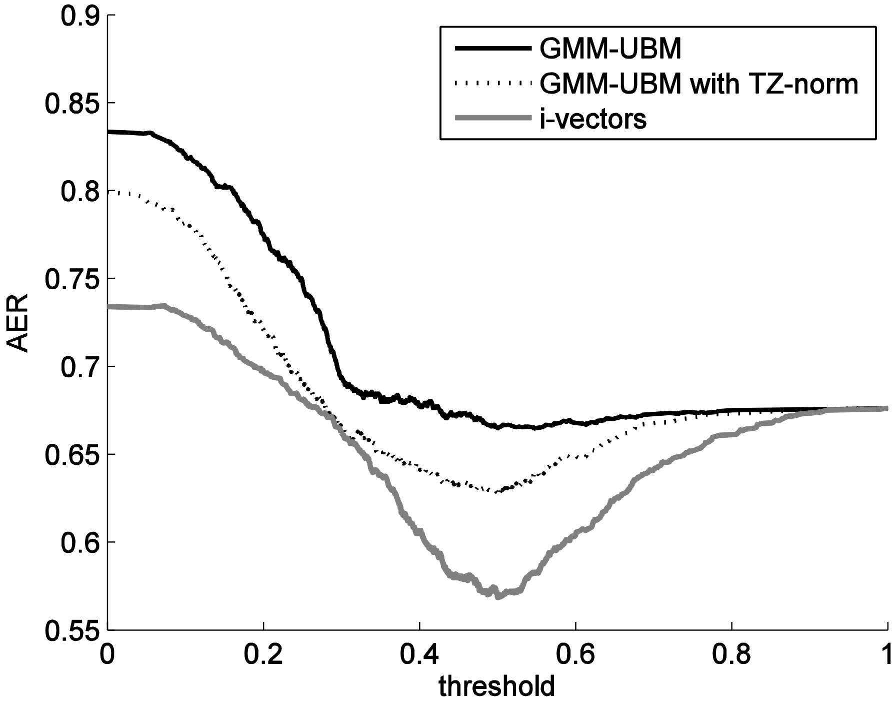

Figure 3.4 AER plots for different methods based on the use of speech data contaminated with a moderate level of white noise (SNR = 15 dB).

The plots in figures 3.3 and 3.4 show that in the case of the lower SNR (5 dB), there is little difference between the three considered approaches as far as the minimal AER is concerned.

This is in spite of the improvement of the identification rate in the first stage that is achieved by TZ-normalisation and i-vectors respectively in comparison with the GMM-UBM baseline. However, when the SNR is increased to 15 dB, the i-vector and GMM-UBM with TZ-norm offer higher accuracy rates than the baseline system (Figure 3.4).

3.2.7. AER results for speech data contaminated with car and factory noise

In order to more realistically reflect the conditions encountered in real applications, the experimental investigations are extended to include car and factory noise. As in the previous section, SNRs of 5 and 15 dB are considered for both types of noise. The experimental results for the resultant four conditions are presented in Figures 3.5 to 3.8.

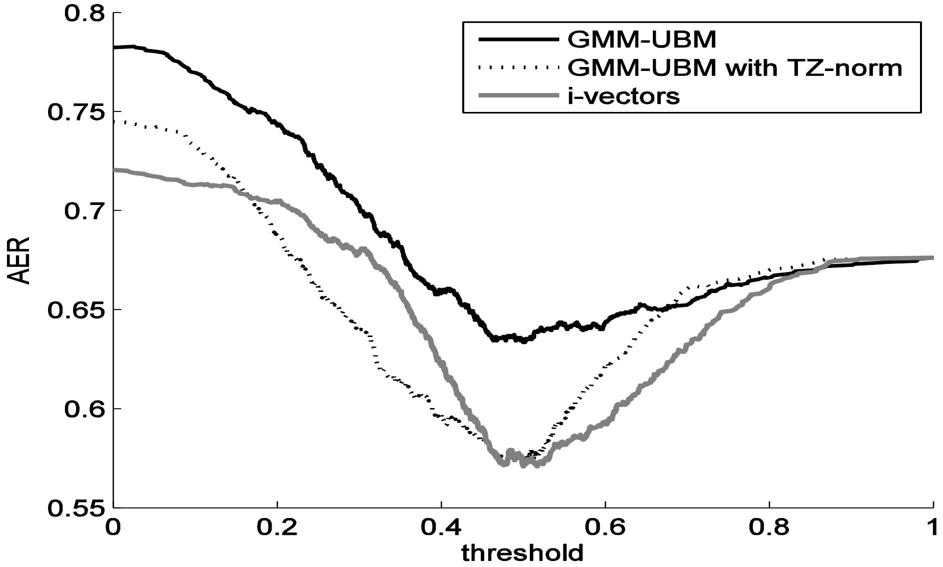

Figure 3.5 AER plots for speech data contaminated with car noise (SNR= 5db)

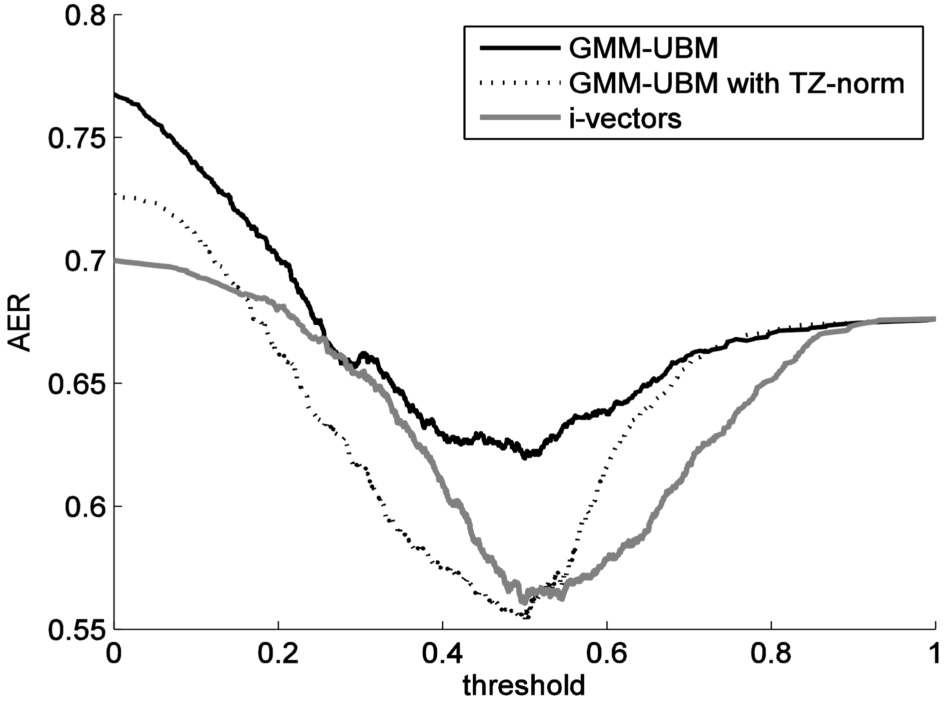

Figure 3.6 Experimental results for speech data that is moderately contaminated with car noise (SNR=15db)

There are a number of interesting observations to be made from these results. For instance, it can be seen that the synthetic white noise has a more severe adverse effect on OSTI-SI accuracy in comparison with the real-world noise types. Moreover, when comparing the minimal AERs for the different classification techniques considered, it can be noted that in the case of car noise, there is little difference in performance between "GMM-UBM with TZ-normalisation" and i-vector for the two SNR levels adopted. However, i-vector performs significantly better when factory noise has been added to the audio files. This is especially the case for the lower noise level (i.e. SNR of 15 dB). It should also be noted in this context that the background speech data used for the TZ-normalisation technique is contaminated with the same level and type of noise as the training and testing data. This is somewhat similar to the CT-norm method presented in [3]. Additionally, the experiments in this study have been based on the use of identical levels and types of noise in the training and testing data. As part of further work in this area, it is important to evaluate the effects of noise mismatch on the performance of OSTI-SI.

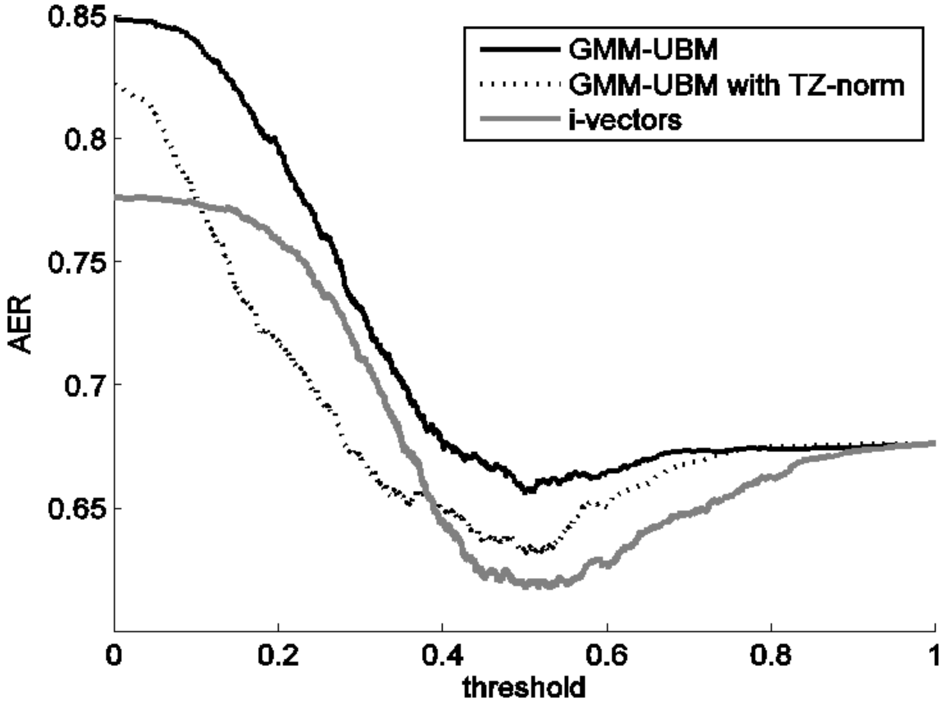

Figure 3.7 AER plot for speech data contaminated with factory noise (SNR =5 dB)

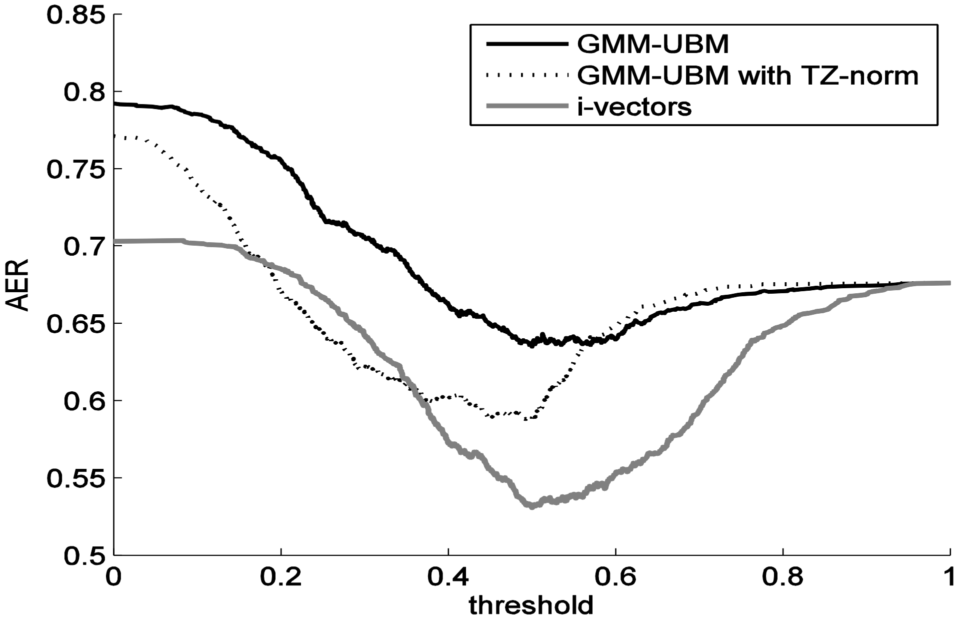

Figure 3.8 Experimental results for speech data moderately contaminated by factory noise (SNR =15dB).

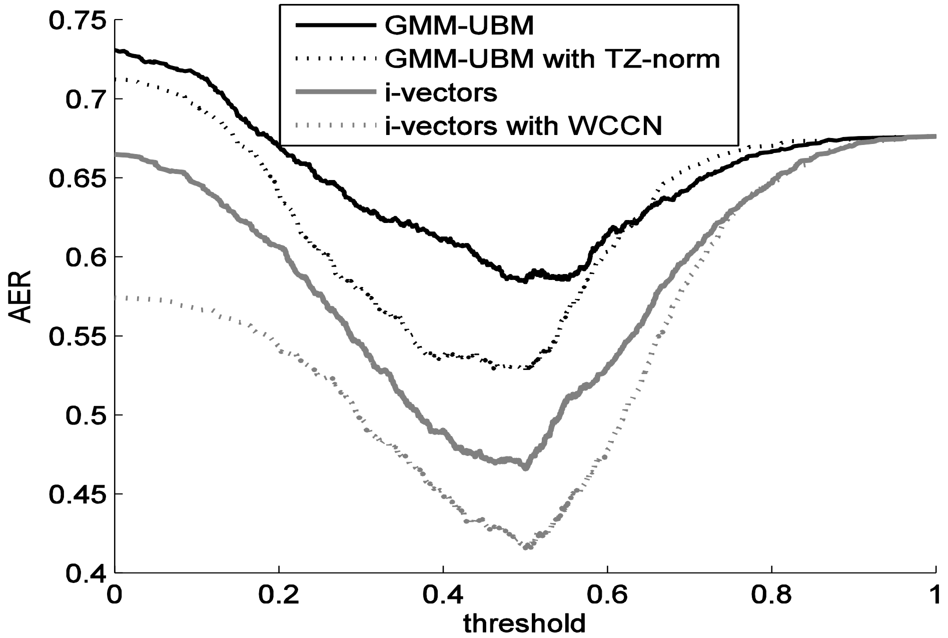

Figure 3.9 Results illustrating the superior performance of i-vector with WCCN in experiments based on NIST telephone quality data

In order to consider the intra-speaker variability compensation offered by i-vector, a set of experiments is conducted using the within class covariance normalization (WCCN) as described in section 2.4.10.1 The result of this experimental investigation is presented in Figure 3.9 for the NIST telephone-quality speech data, together with the results for the same data condition presented earlier in Figure 3.2. As observed, the incorporation of WCCN appears to further improve the recognition performance of the i-vector technique.

3.3. Summary

Overall, the experimental findings for OSTI-SI show that in comparison with the more traditional GMM-UBM approaches, the i-vector technique tends to be more robust against noise contamination of the speech data. However, the level of superiority of this approach appears to vary somewhat with the type and level of additive noise in speech.

For high levels of noise contamination, the outcomes indicate the necessity to consider alternative or additional methods for enhancing the OSTI-SI accuracy. For example, the approach based on multi-SNR UBMs has shown promising results in [4]. A strategy that might further the accuracy of the i-vector approach, even in the presence of high levels of noise, could be that based on using multiple total variability matrices as well as multi-SNR UBMs for various signal-to-noise ratios.

Chapter 4 investigates the effect of diverse duration speech data on the performance of open set text independent speaker identification, and methods of overcoming such effects using weighted bilateral scoring.

CHAPTER FOUR

4. Open-set speaker identification with diverse-duration speech data

4.1. Introduction

The aim of this chapter is to provide a thorough investigation into the effect of short and varied duration reference data in open-set speaker recognition. The contribution is when comparing varied duration and its effect on system performance. Furthermore, a comparison has been done when varied data is enrolled and its effects on the current state of the are i-vector and baseline GMM-UBM classifiers as well as an extension to the baseline system referred to as weighted bilateral scoring (WBS).

In order to represent real world scenarios, four conditions have been exploited. These conditions are referred to as long, medium, short and mixed duration reference material. Generally, research use short duration reference material because this is where the challenge is as discussed earlier in the introduction chapter and also, as discussed in [125] where an investigation was conducted on the current state-of-the-art i-vector, to see the effect of training data length on system performance using the NIST2008 data set. They concluded a rapid reduction in performance as the duration of the training data is reduced, for instance, at unified training and test duration of condition interview, where each recorded session is 10 seconds minimum, they recorded an EER of 25.51%. This was further improved when they applied WCCN. They further investigated, the conditions interview and telephone recordings were adopted, with uniform duration of 10 seconds the EER was further increased to 32.7% because the quality of telephone is not as good as interview. It’s worth mentioning WCCN offered significant improvement under all conditions especially when the channel conditions differ.

Furthermore in the following study [126] they generated a GMM-UBM of 2048 using 4032 unique male and female training utterances from the NIST 2004-2008 data set. They only used the female portion of the NIST 2010 data set for training and testing during evaluation. They recorded an EER of 25.66% when unified duration of test and training utterance of 10 seconds were used. The EER further increased when they reduced the training and test duration to 5 seconds where they recorded 31.1 EER.

In this study [127], a gender dependent UBM was generated using the switchboard I, II phase corpus of size 512. The evaluation data used was of the short2-short3 NIST 2008 condition. They also used data from 150 speakers each with 10 sessions as normalization data. The i-vector performance after applying S-normalisation whole the duration of training and test utterance was at 50 seconds they recorded 6.9% EER, this further increased to 9.5% when the duration of test and training was reduced to 30 seconds. A dramatic increase in EER of 18% was recorded when duration of training and test was further reduced to 10 seconds.

In the study [128], they compared the baseline classifier with the current state of the art i-vector under different test conditions. The NIST SER 2010 telephone speech was used as evaluation data and under all conditions the duration of training data was 20 seconds. Two 1024 dimension diagonal component UBM was generated and the following system performance was recorded. It’s worth mentioning that the multi feature method that was proposed in this study reduced EER significantly under all training duration length

Table 4.1 Experimental results of the study [128] which comparison of the baseline and the current state of the art classifiers performance under different enrolment data duration.

Duration of test data	10 seconds EER	6 seconds EER	2 seconds EER
UBM-EM	12%	18%	32%
I-vector using MFCC features	7%	17%	28%

As seen in the above research short duration is investigated more while no emphasis has been on the challenging case of mixed duration reference material, proposed in this study. Therefore, this will be the focus and main contribution of this chapter.

The remainder of this chapter presents the experimental investigation together with the analysis of the results and finally, the overall conclusion and future work.

4.2. Experimental investigations

4.2.1. The adopted speaker classifiers

For the purposes of the experimental investigations, a total of five speaker recognition methods are considered as follows, which are thoroughly discussed in the literature review chapter.

a) GMM-UBM (providing standard forward scores)

b) Weighted bilateral GMM-UBM (with a weighting factor of f=0.6, determined as appropriate through a set of preliminary experiments as given in the table below)

Table 4.2 Preliminary experiments to identify best value for f

Value of f	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
ID rate %	62	63	71	76	79	82	74	69	60

c) Weighted bilateral GMM-UBM as in (ii) with TZ-norm [120, 121]

d) i-vector

e) i-vector, incorporating within-class covariance normalisation (WCCN)

In the case of i-vector methods, the dimension of the total variability space is 300.

4.2.2. Speech data and OS-SI evaluation protocol

The training and testing data used in this study are from the NIST speaker recognition evaluation (SRE) corpus 2008. As the concern in this research is OS-SI, an appropriate evaluation protocol has been defined on a subset of this corpus, by selecting telephone speech segments from the “short2/short3” core condition [122]. The experiments involve 400 registered speakers and 200 out-of-set speakers. Given the varied number of test utterances for different speakers, the total number of trials comprises 1312 identification trials of enrolled speakers and 627 identification trials of out-of-set speakers.

OS-SI involves the two stages of identification and verification. Traditionally, the recognition accuracy is evaluated separately for these two stages. However, for comparing the performance of several methods against each other, it is more convenient to use a single measure of recognition accuracy that represents the complete process of open-set identification. Motivated by the proposed error computation rate approach in the diarisation process [92], such a measure has been introduced in [112] and is referred to as an accumulated error rate (AER) (described in Chapter 2.5)

The speech dataset for UBM training is taken from the NIST speaker recognition evaluation (SRE) corpus 2005 [123]. This dataset consists of 622 male and 932 female utterances. The developed UBM comprises 2048 Gaussian mixture components.

4.2.3. Experiments with training and test material of uniform length

In the first part of the experimental investigations, it is assumed that the reference and test material is of uniform duration. One set of experiments is conducted using training and testing data of 60 second duration. Figure 4.1 provides the results for these experiments in terms of the accumulated error rate (AER) versus the threshold θ.It should be noted that the plots in Figure 4.1 are obtained by applying a score range normalisation procedure to the AERs (θ) obtained for various methods. Whilst the results clearly illustrate the relative performance of different approaches, it is noted that in each case, the minimum AER is associated with a different threshold. This is because AER (θ) depends on the distributions of client and impostor scores, which are different for individual methods considered. To further facilitate the comparison, a procedure for range normalisation is considered here to shift the minimal point on each plot to $\theta = 0.5$ on the horizontal axis (see Figure 4.2). In general, the results show that if the speech data are of sufficient duration (60 seconds in the given experiments), weighted bilateral scoring does not achieve an improvement over the simpler GMM-UBM baseline. Also, it is noted that the use of TZ-norm can significantly improve the accuracy. As illustrated in Figure. 4.2, the use of i-vector leads to a lower AER, and that this can be further reduced by incorporating WCCN into the i-vector approach.

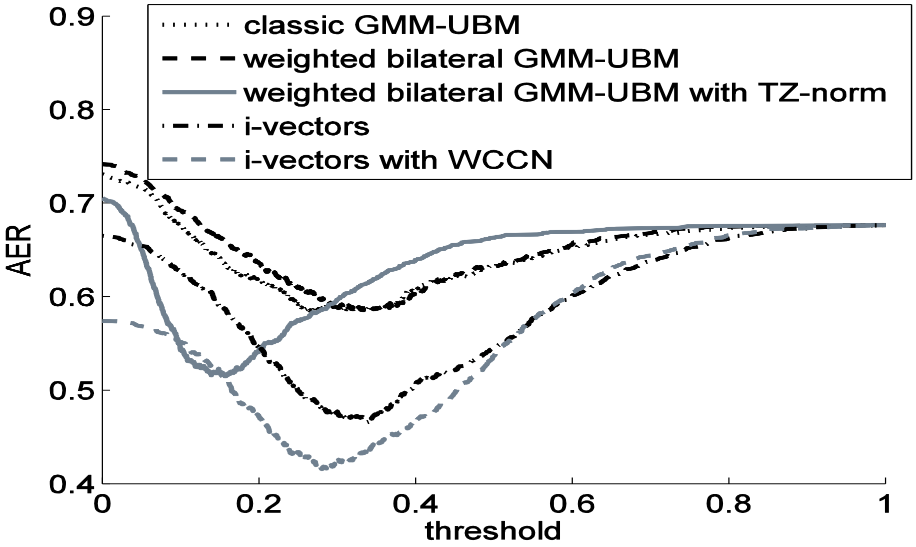

Figure 4.1 Comparison of different methods in experiments with training and testing data of uniformly long duration.

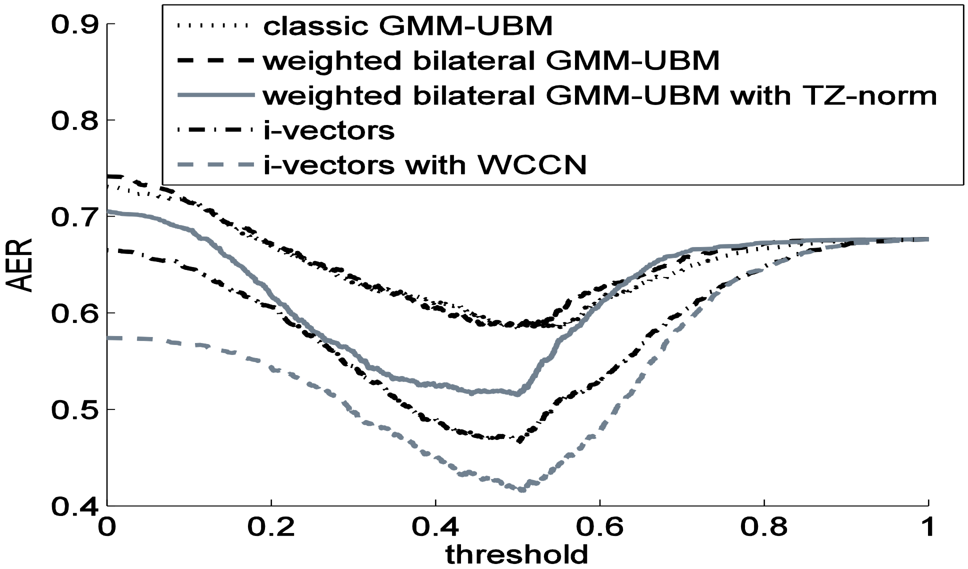

Figure 4.2 Adjusted AER plots for the experiments in the first part of the investigations

To determine how the reduction of the training material duration affects the performance of different speaker recognition methods, the same set of experiments as above is conducted using the reference speech data of two second duration and the testing speech data of 60 second duration. The resulting AER plots are depicted in Figure. 4.3. As observed, the reduction in the training data duration has significantly increased the AER for all the methods considered. Another interesting outcome of this set of experiments is the similar performance offered by different methods.

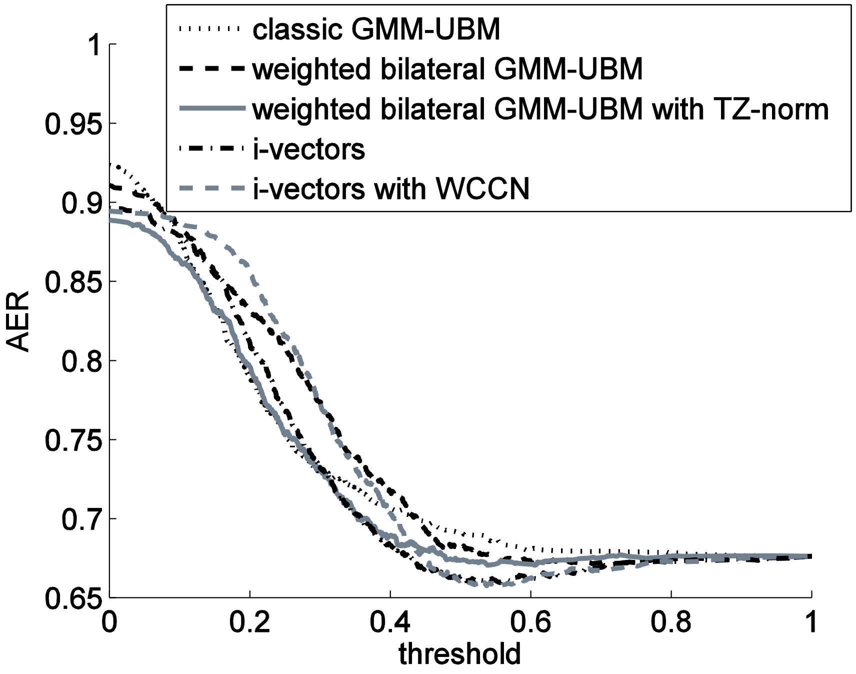

Figure 4.3 AER for different methods in experiments with reduced training data.

4.2.4. Experiments with varied duration reference data

A more realistic scenario in the considered application area is that involving varied duration reference speech data. Therefore, the reference data adopted for the experiments in this part ranges from 1 to 30 seconds for different speakers. The length of test utterances is kept at 60 seconds as in the experiments in the previous section.

Figure 4.4 shows the AER plots for this part of the experimental investigations. It is noted that in this case, the outcome is like the full-length reference data as far as the order of best-performing methods is concerned. The weighted bilateral approach is seen to exhibit a marginal advantage over traditional GMM-UBM. More importantly, it is observed that the i-vector approaches offer very slight improvements over weighted bilateral GMM-UBM with TZ-norm

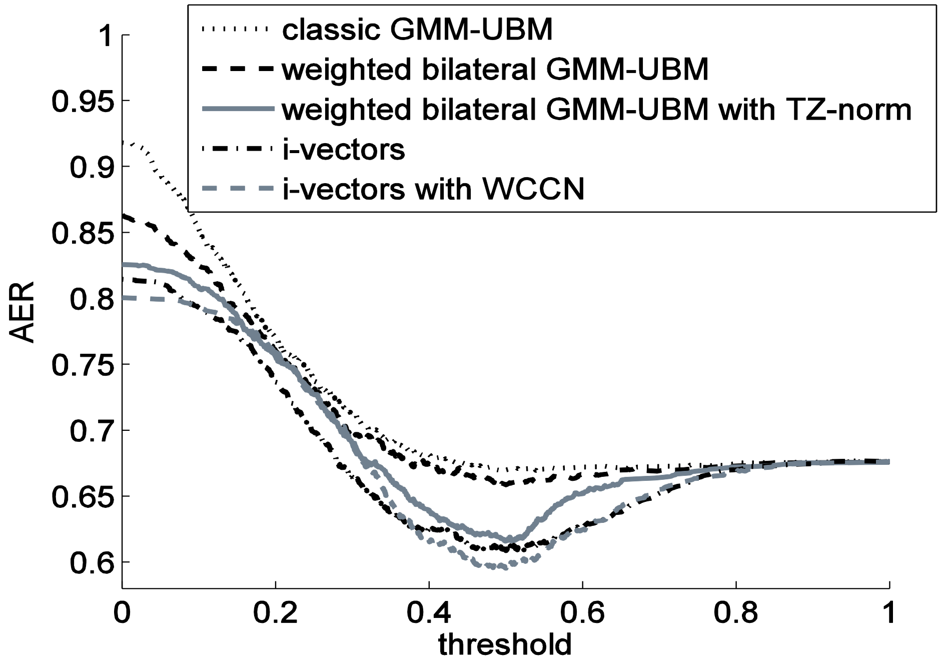

Figure 4.4 Experimental results for varied duration reference and uniform test data.

4.2.5. Experiments with varied duration reference and test material

The last set of experiments is based on the use of training and test utterances of varied duration, ranging from 1 to 30 seconds. This condition is believed to be highly relevant to the considered application area, where there is usually very little control over the duration of speech material captured for training and testing. Figure 4.5 shows the AER plots for this challenging condition. It is noted that, apart from the traditional GMM-UBM approach, all other methods exhibit a similar level of performance. Another interesting observation is that WCCN does not seem to improve the effectiveness of the i-vector approach. This result is contrary to the results obtained with sufficient material of uniform duration, with varied duration references (Figure 4.2), and with uniform duration test data (Figure .4.4), but similar to the uniform, but short reference material (Figure. 4.3).

4.3. Summary

The investigations in this study have been related to the challenges posed by varied duration speech data in open-set speaker identification. This represents the scenario in a number of important applications in such areas as surveillance and criminal investigations.

The experimental results show that with sufficient enrolment and test data, state-of-the art speaker identification approaches such as i-vector with WCCN (for intra-speaker variability compensation) attain a high degree of accuracy in open-set speaker identification and achieve significant improvement over more traditional techniques (Figure. 4.2). However, when the reference data is of short and varied duration, the i-vector technique offers marginal improvement over bilateral GMM-UBM with TZ-norm, and WCCN appears to be less effective.

Furthermore, if reference data is too short or if both reference and test data are varied, a significant drop in OS-SI accuracy is experienced, and there appears to be little difference between the performances achievable by the methods considered in the study.

This chapter concludes that in realistic conditions there is no significant difference between the classifiers explored. For this reason the novel approach is presented in the following chapter, of vowel boosting, where emphasis is applied to the chunks of the speech data which contain the most speaker information.

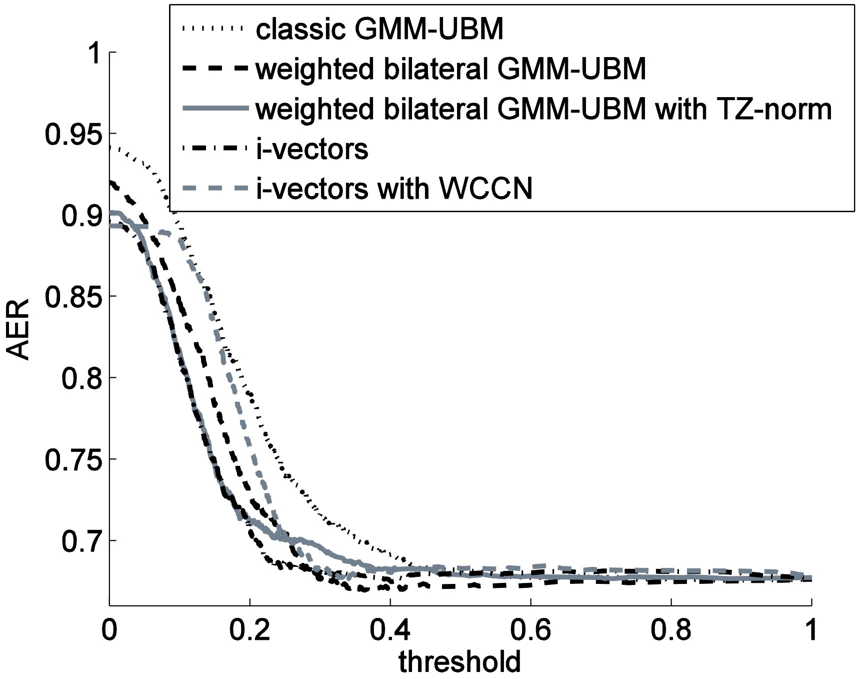

Figure 4.5 Performance of different methods considered in terms of AER in experiments with varied duration reference and test material.

CHAPTER FIVE

5. Vowel Boosting: A Novel Approach to Enhance the Reliability in Speaker Identification

5. Vowel Boosting: A Novel Approach to Enhance the Reliability in Speaker Identification

5.1. Introduction

From the first experiment on ‘The Effects of Environmental Noise on the Accuracy of Open-Set Speaker Recognition’, a need was identified to improve voice recognition for speech data contaminated by various types of noises including car and factory noises. In the second experiment on ‘Open-set Speaker Identification with Diverse-Duration Speech Data’, it was determined that the weighted bilateral approach has an advantage over traditional GMM-UBM approaches in identifying diverse-duration data. In addition, the I-vector was also found to offer an improvement on weighted bilateral GMM-UBM with TZ-norm. In this chapter, a novel approach named ‘Vowel Boosting’ is presented to meet the growing need for effective extraction of intelligence and evidence from audio recordings in the fight against crime based on the results from the two experiments.

A main difficulty in the considered application of speaker identification is that obtaining audio material without speakers’ knowledge or cooperation under uncontrolled environmental conditions results in undesired variations in the utterances acquired for the process in terms of quality and duration. The problem associated with the lack of control over the utterance duration is twofold. First, utterances of short duration are limited in terms of phonetic content. This in turn detracts from the quality of the reference models built for the purpose of speaker identification. Second, in the case of multiple targets enrolled onto the recognition system, (e.g. through a number of recordings obtained from different events), the duration of training utterances can vary from speaker to speaker. For example, in some cases the reference utterances for a subset of speakers can be significantly longer in duration in comparison to those for other enrolled speakers.

Another factor influencing the complexity of OSTI-SI in practice is the size of the population of registered speakers. As the population grows the difficulty in discriminating amongst the registered speaker voices increases. In addition, the growth in the population also increases the difficulty in confidently declaring a test utterance belonging to or not belonging to the initially nominated registered speaker [115].

The research into speaker identification over the past several years has resulted in considerable advances in the field and the establishment of well-defined approaches. These approaches are based on firm pattern matching principles, and incorporate capabilities for dealing with the effects of noise and other causes of variation in speech characteristics [3, 12]. However, to date, there has been limited attention to the problems posed by operating under uncontrolled conditions resulting in the lack of control over the duration of reference speech data. Establishing the extent of the challenge in this case requires experimental evaluations of the effects of varied duration reference material on the speaker identification accuracy. An undesired issue expected in such an operating condition is that the phonetic content of a given test utterance may not be well represented in the reference model for the true speaker, when this relies on short training speech. One possible solution put forward by the study in [101] is based on a weighted bilateral scoring using the GMM-UBM classifier. A significant increase in recognition accuracy was recorded in that study. The work in [12] presents a comparison of the performance the joint factor analysis and i-vector classifiers, involving the use of different normalisation techniques to compensate for channel variability. Whilst the outcomes show considerably drop in accuracy for short training data, it is also indicated that in such a condition, regardless of the classifier type, the phonetic richness of the evaluation data plays a significant role.

An additional difficulty in the application area considered here is that, in the case of varied duration training data for the registered population, a test utterance spoken by a particular enrolled speaker may achieve a better match score against the model for another speaker that provides a fuller representation of the phonetic elements in the test utterance. Previous work by the authors [115] investigated weighted bilateral scoring with test and zero normalisation for recognition performance improvement using varied duration training data. It is worth mentioning that limited improvement was recorded, motivating further investigation into the subject area.

This study proposes a novel approach termed “Vowel Boosting (VB)” to enhance the reliability in speaker identification when operating with varied duration speech data under uncontrolled conditions. The proposed method involves the classification of the given speech data into broad phonetic units, and the emphasis on units offering a relatively higher discrimination capability. To analyse the characteristics of the challenges involved, a thorough investigation into the effects of Long, Medium, Short and Mixed training durations on different classifiers, the baseline GMM-UBM with score normalisation, GMM-UBM with weighted bilateral scoring and the current state of the art i-vector with WCCN normalisation is presented.

Recall, the process of open-set text-independent speaker identification (OSTI-SI) involves two stages; identification and verification. The universal speaker set consists of two subsets of known (registered) speakers and unknown speakers. Within this scenario, there are three possible types of errors Miss Labelling, False acceptance and False rejection. Where if an ML error occurs its unrecoverable for this reason the main focus of this research is to reduce ML errors as discussed earlier in the literature review.

This chapter is organised as follows. In Section 5.2 investigates the current methods of phonetic-based speaker recognition systems. In section 5.3 the proposed vowel boosting method is introduced. Section 5.4 details the experimental investigations and provides an analysis of the results. Finally, in section 5.5 the summary of this chapter is discussed.

5.2. Phonetic-based speaker recognition

In automatic speaker recognition, the phonetic content of the training and testing material plays a vital role in matching the test utterance from a registered speaker to his/her model in the reference set, and this has been the focus of many studies. The research in [97] presents investigations into the distance between different vowels from different speakers. The study concludes that a vowel segment by a given speaker should have a high probability of matching well with one of the vowels extracted from his/her own training utterance. The study in [98], which involves the use of neural network speaker recognition, emphasises that vowels contain the most speaker-specific information. Their approach is to spot all the vowels within the utterance using the feed forward multi-layer perceptions (MLP) and to represent the features using perceptual linear predictive (PLP) speech analysis technique. The study concluded that the higher the number of vowels spotted the higher overall recognition performance. In other studies, [93, 129] attempts are made to tackle the challenges in speaker recognition with short utterances. In [93], the vowel sounds are categorised into eight sub-categories based on the IPA vowel chart [130] and for each sub-category a universal phoneme model is trained (UBVCM) from the background data. During the enrolment, the training utterance, which is assumed to be of significant duration (to meet the requirement of the approach), is first passed through a phoneme recogniser. The vowels within the training utterance are then categorised and eight models are produced, which are referred to as Vowel-Class (VC). During the test phase (short durations), the same procedure of enrolment is repeated, eight scores are obtained by evaluating the VCs against UBVCMs and the scores are fused to obtain the final score for each trial. This was later enhanced by reducing the categories of vowels from eight to five [129]. In another study [56], a comparison of the relative speaker discrimination properties of broad phonetic classes is presented. The classifiers used in this work are the baseline GMM-UBM, and vector quantisation (VQ). The study concludes that certain phonetic groups contain more speaker-specific information than others. It is further concluded that the performance achieved by using vowels exclusively is similar to that obtained by using the entire training utterance. Moreover, it is indicted that the phonetic content of the training speech material is more important for the task than simply its duration.

By considering the first stage of OSTI-SI and the challenges associated with the application area in this study, a main factor affecting the recognition performance is that of the duration of the reference material. To further clarify this point it should be noted that, in general, training a speaker model with a short utterance cannot be expected to provide a strong representation of the broad phonemes. In the text-independent mode of operation, this adversely affects the quality of speaker model. As a result, the reliability of OSTI-SI, which normally entails the enrolment of a considerable number of speakers, can be significantly reduced. To be more specific, the test utterance of a registered speaker can potentially be identified as originating from another speaker model (generated with long training utterance), which is also registered in the set and has a better phonetic representation. The fact that the application area considered in this study involves the deployment of speaker recognition in uncontrolled operating conditions means that there is no control over the duration or the phonetical content of the speech material obtained. As indicated above, previous studies show that vowels' contribution to speaker segregation is more than other phonemes because they contain more speaker discriminative information [56, 93, 97, 98, 129]. However, it should be noted that aiming to focus specifically on vowels (or any other particular phoneme) requires a significant amount of audio material which in this case is an uncontrollable variable, making the methods discussed unsuitable for dealing with the scenarios covered in this study. This motivates investigations to identify effective methods for enhancing the speaker recognition performance under the considered challenging operating conditions. The main facet of these challenges is the OSTI-SI operation based on short training material, which results in a poorly adapted model.

In relation to the above problem, the conclusions given in [56, 93, 97, 98, 129] prompt the inference that if the phonetic areas within the speech material that contain relatively more speaker discriminative information are emphasised during the recognition process, then the probability of correctly matching a given test utterance to the poorly adapted model of the true speaker can increase. Establishing an effective approach for this purpose necessitates experimental investigations in order to determine the relative influence of different phonetic groups on the speaker recognition performance. The outcomes of such investigations together with the conclusions in the previous studies can then be used as a basis to determine procedures for enhancing the accuracy in speaker recognition in the considered application area.

5.3. Proposed method of vowel boosting

As indicated earlier, the specific challenges that arise in the considered security application of SI are due to operating with short and varied reference speech data. In such a scenario, it is reasonable to assume that there is no user cooperation, as the audio recording may be obtained without the users' awareness. Therefore, there is effectively no control over the duration and phonetic richness of the material obtained. This in turn can result in poor representation of a speaker's voice characteristics in his/her reference model. In other words, the generated speaker model can be undesirably limited in terms of phonetic content. Matching a test utterance to a poorly adapted speaker model when they both originated from the same speaker is a very challenging task. The task becomes further complicated when multiple speakers are enrolled, as is the case in the identification process. In this case, such a modelling problem can increase the risk of mislabelling, i.e. a test utterance from a particular registered speaker achieving a higher match score against another speaker model that offers a richer representation of phonemes (one that is trained with a long duration training utterance).

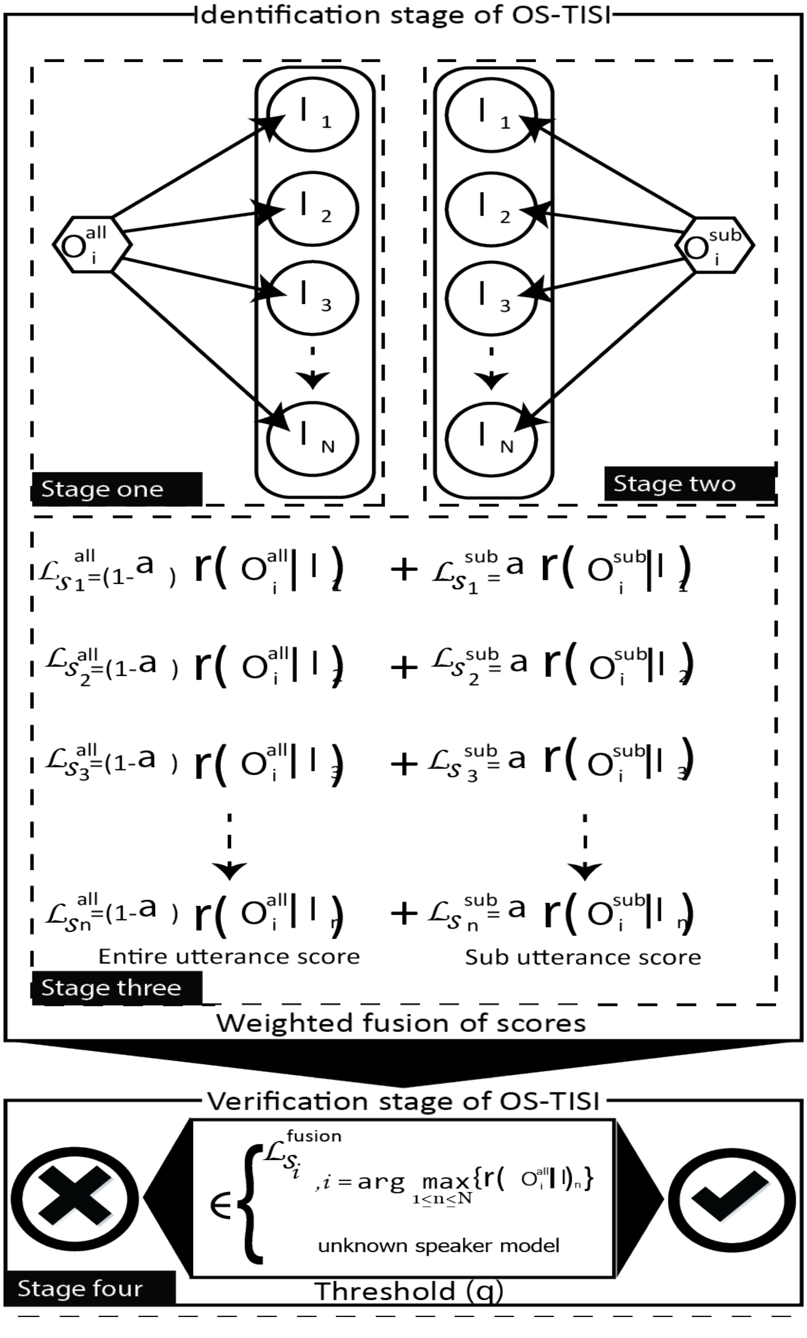
Identification stage of OS-TISI
I 1
I 2
I 3
I N
O all i
O sub i
Stage one
Stage two
$\mathcal{L}_{S1}^{all}$=(1-a) r(O all i | I 1)
+ $\mathcal{L}_{S1}^{sub}$= a r(O sub i | I 1)
$\mathcal{L}_{S2}^{all}$=(1-a) r(O all i | I 2)
+ $\mathcal{L}_{S2}^{sub}$= a r(O sub i | I 2)
$\mathcal{L}_{S3}^{all}$=(1-a) r(O all i | I 3)
+ $\mathcal{L}_{S3}^{sub}$= a r(O sub i | I 3)
$\mathcal{L}_{Sn}^{all}$=(1-a) r(O all i | I n)
+ $\mathcal{L}_{Sn}^{sub}$= a r(O sub i | I n)
Entire utterance score
Sub utterance score
Stage three
Weighted fusion of scores
Verification stage of OS-TISI
$\in$ $\mathcal{L}_{S_i}^{fusion}$,i = arg max 1≤n≤N {r(O all i | I)n}
unknown speaker model
Stage four
Threshold (q)

Figure 5.1 Illustration of proposed method.

O_i^{all} Entire test utterance of instant i, O_i^{sub} sub (vowels only) test utterance of instant i, I_N reference model of instant $\boldsymbol{\mathcal{L}}_{sN}^{fusion}$, a weight $a = 0.2$, $\boldsymbol{\mathcal{L}}_{s}^{all}$ score obtained when the test utterance O_i^{all} is evaluated against I_N, $\boldsymbol{\mathcal{L}}_{sN}^{sub}$ score obtained when test utterance $O_i^{sub}$is evaluated against I_N

Research has shown that within the phonetic groups of a language, some offer more speaker specific information than others [56]. Furthermore, the studies in [56, 93, 97, 98, 129, 130] all conclude that the vowel phonemes are the greatest contributors to the speaker recognition performance. It is also worth noting that, according to the study in [56], the phonetic content of the training speech material is more important than its quantity. In addition, the studies in [56, 93, 97, 98, 129, 130] conclude that vowel phonemes always have a better chance of matching to another vowel when originated from the same speaker.

Table 5.1 Experimental results of previous research in using vowels

Ref	Duration of reference material	Performance	
[56]	2 sec	GMM	20% ID rate
[93]	9 sec	GMM-UBM	42% EER
[128]	2 sec	GGM-UBM	33% EER
		i-vector	30% EER
[97]	100 sec	GMM	70% ID rate

The table above is a summary of previous work presenting duration of training material as the greatest contributors to the speaker recognition performance. The authors concluded that the more vowel phenomes result in better performance. The majority presents their experimental results using EER's. In this work, to evaluate the proposed "vowel boost" approach AER is used as discussed in the literature review Section 2.5, which provides a better classifier performance evaluation as it takes both stages of OS-SID (identification and Verification) into consideration.

The above conclusions further support the motivation for reducing the identification error rate through the introduction of a method that incorporates the relative speaker discriminative characteristics of different phonetic classes. On that basis, a new approach is proposed here, which is briefly outlined in Figure. 5.1. As noted in this figure, whilst the method involves placing a relatively higher emphasis on the vowel content, the entire test utterance data is used when forming the match scores in the identification process. The proposed method as indicated in Figure. 5.1, involves the following four stages.

Stage one: The probability of the entire test utterance against the pre-registered speaker models is obtained.

$$\mathcal{L}s_i = \rho \left(\mathbf{O}_{all} \middle| \lambda_i \right) \qquad \textbf{\textit{5. 1}}$$

Where $\mathcal{L}s_i$ is the score obtained, ρ is the probability (loglikelihood) of the *i*-th trained speaker with reference model $\boldsymbol{\lambda}_i$, evaluated against the feature vectors of the entire test utterance $\mathbf{O}_{all}$.

Stage two: The vowel phonemes of the given test utterance are determined and extracted using a phonetic recognition engine. The score for the sub test utterance is then obtained as the probability of the vowels against the pre-registered speaker models.

$$\mathcal{L}s_i^{sub} = \rho \left(\mathbf{O}_{sub} \middle| \lambda_i \right) \qquad \textbf{\textit{5. 2}}$$

Where $\boldsymbol{\mathcal{L}s}_i^{sub}$ is the sub score obtained, ρ is the probability of the reference model λ_i of the i-th speaker, evaluated against feature vectors belonging to the vowels elements of the test utterance $\mathbf{O}_{sub}$.

Stage three: This stage involves fusing the scores obtained in stage one ($\boldsymbol{\mathcal{L}s}_i$) and stage two ($\boldsymbol{\mathcal{L}s}_i^{sub}$) using an appropriate weighting procedure. Through a set of preliminary experiments, a weight factor of α=0.2 was determined as appropriate.

$$\boldsymbol{\mathcal{L}s}_i^{fusion} = (1-\alpha)\boldsymbol{\mathcal{L}s}_i + \boldsymbol{\mathcal{L}s}_i^{sub} \qquad \textbf{\textit{5. 3}}$$

$\boldsymbol{\mathcal{L}s}_i^{fusion}$is the final score obtained when fusing $\boldsymbol{\mathcal{L}s}_i$and $\boldsymbol{\mathcal{L}s}_i^{sub}$with the weight values of (1-α) and α respectively.

Stage four: In this verification stage of OSTI-SI, the test utterance $O_{all}$is verified against $\boldsymbol{\mathcal{L}s}_i^{fusion}$based on a pre-set threshold *θ*. In the case of the fused score being greater than the threshold *θ*, the utterance is accepted. Otherwise, it is declared as originated from an unknown speaker.

$$max_{1\le i\le n}\{\boldsymbol{\mathcal{L}s}_i^{fusion}\} \gtreqless \theta \longrightarrow \mathbf{O}_{all} \in \left\{\begin{matrix} \boldsymbol{\lambda}_l, l = \overset{\text{argmax}}{}_{1\le i\le n}\{\boldsymbol{\mathcal{L}s}_i^{fusion}\} \\ Unknown\ Speaker \end{matrix}\right\} \qquad \textbf{\textit{5. 4}}$$

Fusing $\boldsymbol{\mathcal{L}s}_i$ and $\boldsymbol{\mathcal{L}s}_i^{sub}$with the appropriate weights will emphasise the influence of the vowel elements in the scoring procedure. Since vowel elements contain more speaker specific information, the fused score will be expected to increase the likelihood of identifying the correct speaker of the test utterance from the registered set.

5.4. Vowel boosting evaluation

For the purpose of investigations, two sets of experiments are considered. The purpose of the first set of, presented in section 5.5, is to establish the contribution of each phonetic group to the recognition performance of OSTI-SI. This is referred to as relative effectiveness of phonetic classes.

The second set of experiments is to evaluate the relative effectiveness of the proposed VB approach, under realistic conditions in terms of the training data duration, as expected in the particular applications of OSTI-SI considered in this study. To be more specific, the experiments are designed to determine

- The effect of training data duration on the recognition performance (Section 5.7).
- The recognition performance under the mixed data duration condition (Section 5.8.2).

Since OSTI-SI involves the two stages of identification and verification, traditionally, the recognition accuracy is evaluated separately for these two stages. However, for the purpose of comparing the performance of several methods against each other, it is more convenient to use a single measure of recognition accuracy that represents the complete process of open-set identification. Motivated by the approach proposed for error rate computation in the diarisation process [92], such a measure has been introduced in [112] and is referred to as accumulated error rate (AER). This is defined by:

$$max_{1\le i\le n}\{\boldsymbol{\mathcal{L}s}_i^{fusion}\} \gtrless \theta \longrightarrow \mathbf{O}_{all} \in \begin{Bmatrix} \boldsymbol{\lambda}_l, l = \underset{1\le i\le n}{\operatorname{argmax}}\{\boldsymbol{\mathcal{L}s}_i^{fusion}\} \\ Unknown\ Speaker \end{Bmatrix} \qquad \mathbf{5.\ 5}$$

where θ is the threshold adopted in the second stage of the process, ML(θ) is the number of mislabelled (incorrectly identified) clients, FR(θ) is the number of clients that are falsely rejected, FA(θ) is the number of out-of-set speakers that are falsely accepted as clients, and T is the total number of identification trials.

The database adopted for the experimental investigations is that of TIMIT as it provides phonetic labelling for each recorded session, hence eradicating the possibility of phoneme recognition errors. The TIMIT database contains recordings from 630 different speakers with a 438 to 192 male to female speaker ratio. There are 10 sessions per speaker, with each session consisting of a sentence read out by the speaker. The duration of each spoken utterance varies from 3 to 6 seconds. In total, there are 63000 recordings in the database.

5.5. Relative effectiveness of phonetic classes

In the TIMIT database, each speaker's recorded utterance is labelled phonetically. The label file indicates the starting sample and end sample of each phoneme spoken by the speaker. In this experiment the phonetic label is used to divide the individual parts of each utterance into three groups. These groups are as follows.

- Vowels covering the following phonemes.
  iv , ih , eh , ey , ae , aa , aw , ay , ah , ao , oy ,ow, uh, uw, ux, er, ax, ix, axr, ax-h
- Fricatives containing the following phonemes
  s, sh, z, f, th, v, dh, m, n, ng, em, en, eng, nx
- Others phonemes

Three experimental conditions are considered to determine the relative contributions of the above phonetic classes to the recognition performance. The adopted speech data is divided into the required three phoneme groups based on the information provided in the TIMIT database documentation. As indicated in Table 5.1, the three experimental conditions differ from each other in terms of the phonetic class of data used for training, testing and UBM construction.

Table 5.2 Experimental conditions, where Sub-UBMv, Sub-UBMo and Sub-UBMf are the UBMs generated exclusively with certain phonemes within the background data where v represents vowels, o represents others, and f represents fricatives

Experiment	Training material condition	Test material condition	UBM material condition
Vowels	Vowels only	Vowels only	Sub-UBM^v
Others	Others only	Others only	Sub-UBM^o
Fricative	Fricative only	Fricative only	Sub-UBM^f

5.5.1. Performance of phonetic classes: experimental setup

The initial experiments detailed here are aimed to determine the relative effectiveness of the considered phonetic classes for open-set speaker identification. As the well-known GMM-UBM technique has been one of the dominating approaches in the field of speaker recognition for the past two decades [92, 99, 100] , this approach is selected for the purpose of performance evaluation here and also as the baseline in other experimental studies in this research. Test normalisation (T-norm) and Zero normalisation (Z-norm) are the two score normalisation approaches used in this study because of their capability to improve the GMM-UBM recognition performance. T-norm compensates for inter–session score variation, attempting to reduce any acoustic or session mismatch between testing and training data from the same speaker [10]. Z-norm, on the other hand, tries to compensate for the inter-speaker score variation which is a primary concern with the mismatch in the training condition (e.g. different microphones). The aim is to align the speaker models, which are generated under different training conditions, prior to the test phase [87]. As indicated earlier, the purpose of this investigation is not to compare the recognition performance of different classifiers, but to assess the contributions of each phonetic class to the recognition performance. For this purpose, the baseline classifier (i.e. GMM-UBM) is considered sufficient.

In this study, the TIMIT dataset is used in the following manner

- 200 speakers for building the required UBMs,
- 120 speakers as the registered speakers,
- 150 speakers as the background speakers for score normalisation, and
- 66 speakers as the out of set (unknown) speakers.

Data preparation for UBM: For the 200 UBM speakers, the content of each utterance is divided into three groups (Vowels, Fricatives and others) based on the phonetic labelling provided by TIMIT. The speech material in the individual groups is then used to build the corresponding sub UBMs (i.e. one sub UBM for each of the phonetic groups defined above).

In order to generate gender balanced sub UBMs, recordings from 100 male speakers and 100 female speakers are used. There is 1 utterance in each of the 10 recording sessions for each speaker, providing a total of 2000 utterances for sub UBMs. The size of the sub UBMs in terms of the number of Gaussian components varies according to the distribution of the phonemes within the entire dataset. The sub UBMs' sizes are as follows:

- *Vowel's sub UBM: 512*
- *Fricative's sub UBM: 256 mixtures*
- *Other's sub UBM: 256 mixtures*

Evaluation set (registered speakers): In total, 120 speakers, consisting of 80 male and 40 female speakers are registered. The utterances available from all the ten sessions are segmented into the three considered phonetic groups with each segment labelled accordingly. As indicated in Table 5.2, for each speaker, the data from the first four sessions is reserved for training, and the speech in the remaining six sessions is used to provide three test tokens. It is worth noting that, prior to phonetic segmentation, the training utterance for each speaker does not exceed 20 seconds in duration.

Table 5.3 Foreground material partitions

Evaluation data	Training material				Test material one		Test material two		Test material three	
Speech recording Sessions	1	2	3	4	5	6	7	8	9	10
	Number of registered speakers: 120				Total number of test trials for registered speakers: 360					

Out-of-set speakers (unregistered speakers): The unknown speaker set consists of 66 speakers (i.e. 44 male and 22 female speakers). Although there are 10 utterances available for each unknown (unregistered) speaker, only six utterances per speaker are used here. For each such speaker the available utterances from different sessions are individually divided into the three considered phonetic groups and labelled accordingly. For the purpose of experiments, three out-of-set groups are then formed, each based on material from two of the recording sessions as illustrated in Table 5.3.

Table 5.4 Out of set speaker data

Out-of-set speakers	Out-of-set one	Out-of-set two	Out-of-set-three
Speech recording Sessions	5 6	7 8	9 10
Number of unknown speakers	66	66	66

Background normalisation data: The speech data of 150 speakers reserved for this purpose (120 from male speakers and 30 from female speakers) are divided using the phonetic label into three phonetic groups, and then labelled accordingly. The sub-utterances in each phonetic group are then adopted as background normalisation data, giving 150 T-norm models and 450 Z-norm trials. The background normalisation data is divided as shown in

Table 5.5 Background normalisation segregation.

Normalisation	T-norm	Z-norm one	Z-norm two	Z-norm three
Speech recording Sessions	1 2 3 4	5 6	7 8	9 10
	Total number of T-norm models per phonetic group: 150	Total number of Z-norm trials per phonetic group: 450		

5.6. Performance of phonetic Classes: experimental results

As stated earlier, the purpose of the experiments presented here is to determine the relative recognition performance offered by the three phonetic classes considered in the study. For the purpose of completeness, the identification performance in the first stage OSTI-SI is presented in Table 5.5, whilst the overall AERs are provided in Figure. 5.2 Similar to the conclusion of previous investigations [4], here the vowels are found to offer the highest recognition effectiveness (the lowest AER), outperforming the other two phonetic groups.

It is worth noting that the TIMIT database contains some background noise but it is still considered a clean database. In general, obtaining speech data under uncontrolled operating conditions means that the data can be subject to degradation due to such factors as background noise and channel effects. This can potentially affect the speaker recognition performance unfavourably [13, 87, 131]. The main focus of this study, however, is the effect, on the recognition performance, of variations in the audio material duration. The specific effects of noise and channel variation, which are additional undesired issues in the considered application area, will be covered as part of the future work.

Table 5.6 Relative effectiveness of phonetic classes' experimental results where training material duration does not exceed 20 seconds

System	Vowels	Others	Fricatives
Correctly identified speakers	48%	28%	33%

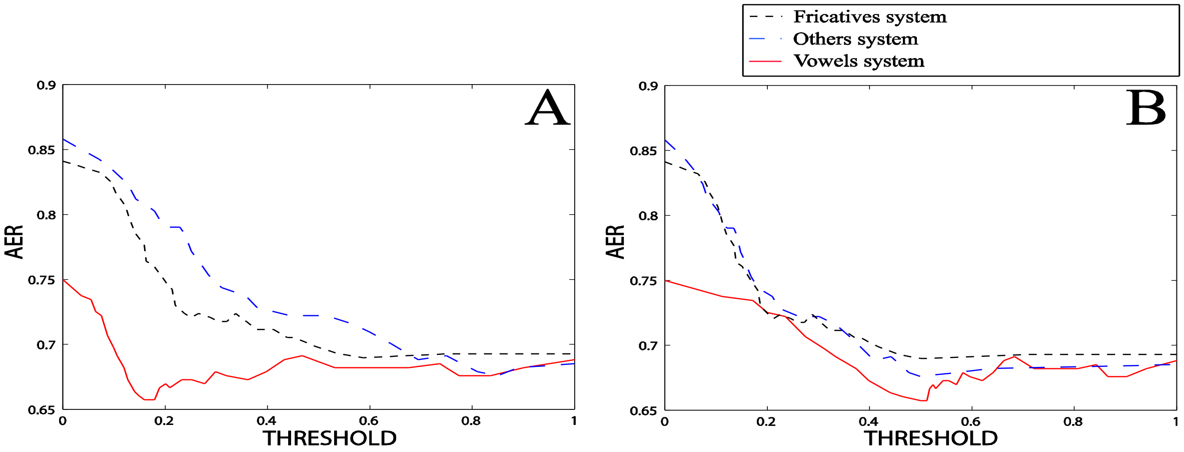

Figure 5.2 Relative effectiveness of different phonetic classes in OS-SI experiments

Subfigure A provides the results for in terms of the accumulated Error Rate (AER) versus the threshold (θ). It should be noted that the plots in subfigure A are obtained by applying a score range normalisation procedure to the AERs obtained for various methods. Whilst the results clearly illustrate the relative performance of different approaches, it is noted that in each case, the minimum AER is associated with a different threshold. This is due to the fact that AER depends on the distributions of client and impostor scores, which are different for individual methods considered. In order to further facilitate the comparison, a procedure for range normalisation is considered here to shift the minimal point on each plot to θ= 0.5 on the horizontal axis (see subfigure B)

5.7. Vowel boosting experiments

5.7.1. Adopted classifiers for the vowel boosting experiments

For the purpose of the study, three speaker recognition classification methods are used to investigate the performance of vowel boosting including and they are as follows:

.

a) GMM-UBM (providing standard forward scores) with score normalisation TZ-norm (discussed in Section 2.4.2),
b) Weighted bilateral GMM-UBM with score normalisation (with a weighting factor of f=0.6, determined as appropriate through a set of preliminary experiments as given in the table 4.2)
c) Weighted bilateral GMM-UBM as in (ii) with TZ-norm [120, 121] (discussed in Section 2.4.8)
d) i-vector (discussed in Section 2.4.10)
e) i-vector, incorporating within-class covariance normalisation (WCCN)

In the case of i-vector methods, the dimension of the total variability space is 300.

The purpose of the experiments in this section is to investigate the effectiveness of the proposed method for reducing the ML error that occurs in the first stage of OSTI-SI (Figure 5.2). For this purpose, the experiments are first conducted to analyse the effects, on the identification accuracy, of short, medium and long training data duration. The experiments are then extended to establish the challenge posed by diverse duration training data in the identification process, and the effectiveness of the proposed approach for addressing it. Table 5.7 presents the actual lengths of training data used in the experiments.

Table 5.7 Training data conditions in terms of duration

	Condition name	Time duration of training material
To establish the relative effect of training duration on recognition performance experiment	Long	18 – 20 seconds
	Medium	8 – 10 seconds
	Short	Approximately 2 seconds
To investigate the effect, on the recognition accuracy, of varied duration training material	Mixed	Equal combination of Long, Mix and Short conditions

The TIMIT database is partitioned as follows:

Universal background model (UBM): As in the earlier experiments detailed above, the UBM built for this part of the study is 1024 in size. It is based on 2000 utterances provided by 200 speakers in 10 recording sessions. Again as before, by using 100 male speakers and 100 female speakers in the process, it is ensured that the UBM is gender balanced.

Foreground speech material: The speech materials from 186 speakers are used as the foreground material. 120 speakers in this group are used as the registered speakers. These consist of 80 male and 40 female speakers. The speech data for these speakers, which is captured in 10 recording sessions, is deployed as shown in Table 5.7.

Table 5.8 The structure of foreground speech material

Experimental condition	Training material				Test material					
	Sessions				Sessions					
Long	1	2	3	4	5	6	7	8	9	10
Medium	1	2			5	6	7	8	9	10
Short	1				5	6	7	8	9	10
	Total number of registered speaker models(in each case)120				Total number of test trials (in each case) 360					

The remaining 66 speakers are used as the unknown speakers (unregistered speakers), with a 44 to 22 male to female ratio. The utterances from these speakers are used to form three sets of unknown speakers.

As before, the utterances from 150 speakers are used as the background data for the normalisation purposes. 120 of these speakers are male speakers and the remaining 30 are female speakers.

In the case of *GMM-UBM and weighted bilateral GMM-UBM* the background data is used as shown in Table 5.4. A total of 150 T-norm models are used, which originate from the first four recording sessions of the relevant speakers, while there are 450 Z-norm utterances produced from the 150 background speaker data set.

For i-vector with WCCN the utterances from all 10 sessions of the 150 speakers are used as WCCN normalisation data.

5.7.2. Experimental results for vowel boosting

The experimental conditions considered are shown in Table 5.7. The proposed VB method is implemented with a weighting factor of α=0.2, determined through a set of preliminary experiments demonstrated in the Figure 5.3. This results in boosting the influence of the vowel phonemes on the recognition decision.

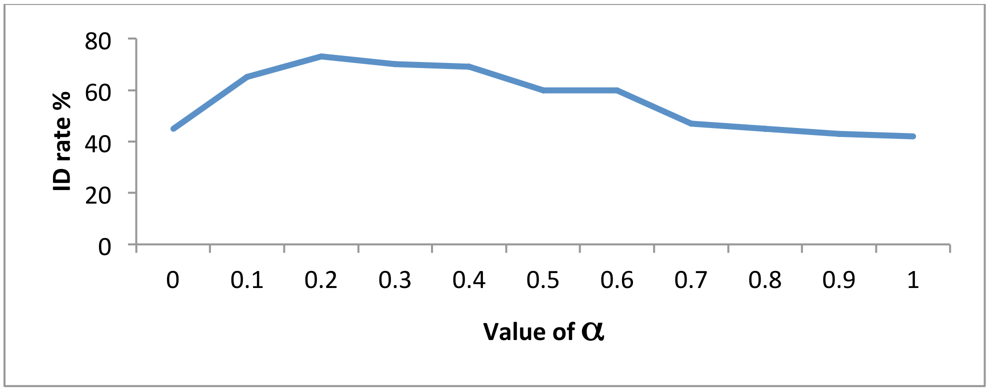

Figure 5.3 Preliminary experiment to determine best value for a, to be noted the duration of reference material was of 40 seconds using the TIMIT 2005 data set

The AER results for *GMM-UBM with TZ-norm* are illustrated in Figure 30, and the identification performance for this system is given in Table 5.8. The results clearly demonstrate that the recognition performance drops and AER increases as the duration of the training material is reduced. It is worth noting that the training utterance duration in the "Long data" condition did not exceed 20 seconds. For this reason the recognition performance is low in comparison to previous studies. The results also demonstrate that the introduction of the proposed method improves identification rate considerably and leads to lower AERs for all conditions. It is interesting to note that in the case of short data conditions, where the training utterances used does not exceed 2 seconds in duration per speaker, a higher increase in the number of correctly identified speaker is recorded. The results recorded demonstrated that emphasising the vowel's contribution in recognition decision reduces the mislabelling error in the first stage of OSTI-SI.

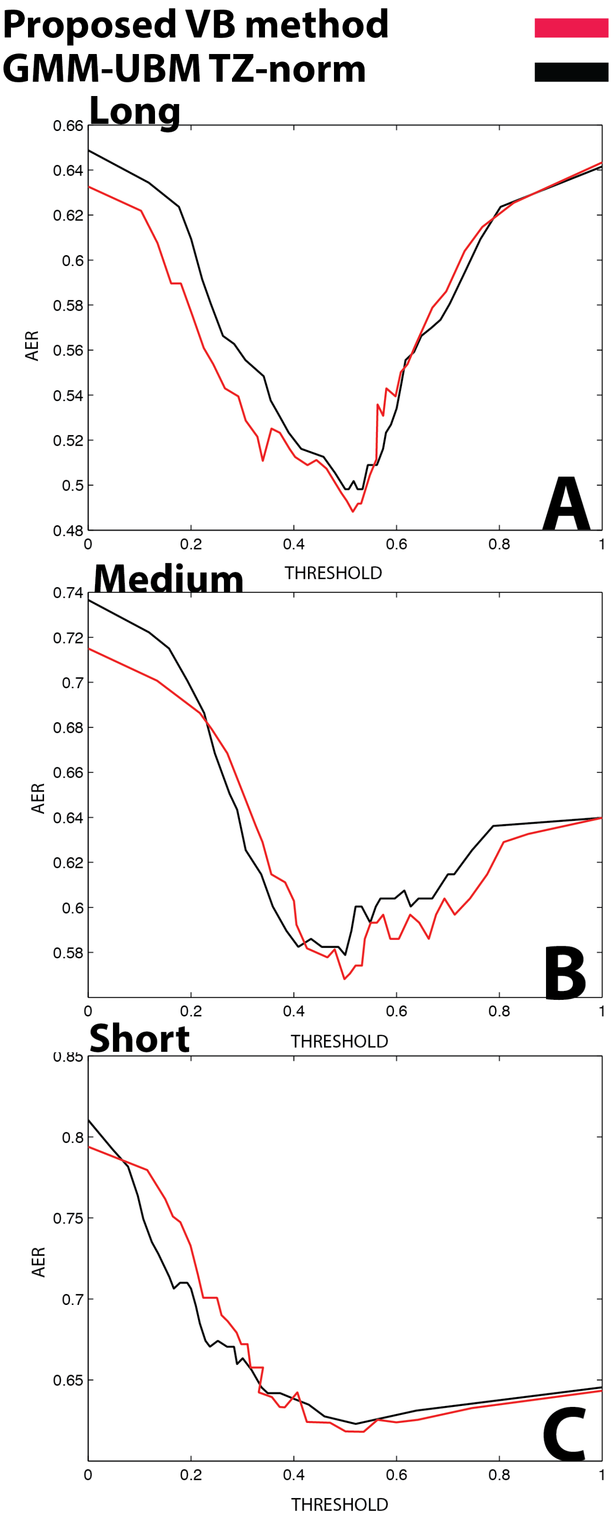

Figure 5.4 Performance of proposed method based on GMM_UBM with TZ-norm, the subfigure A, B and C are the AER results under considered training conditions (i.e. long, medium, and short training data conditions).

Table 5.9 Baseline identification rate

	Long training condition	Medium training condition	Short training condition
GMM-UBM (with TZ-norm)	78%	58%	41%
GMM-UBM the VB method (with TZ-norm)	83%	64%	47%

Figure.5.5 illustrates the AER results obtained through *weighted bilateral scoring (WBS) with and without the VB method, whilst* Table 5.10 provides the identification rate for the two approaches. It should be noted that WBS was a solution proposed by the authors earlier [3, 101] to tackle the effects, on the recognition performance of varied duration training utterances. The experimental results presented below clearly demonstrate that similar to the baseline classifier, the recognition performance drops and AER increases as the duration of the training material is reduced. It is also noted that again the use of VB method results in increasing the identification rate and lowering the minimum achievable AER.

Table 5.10 Identification rates for the WBS classifier with and without the VB method

	Long training condition	Medium training condition	Short training condition
Weighted bilateral score (with TZ-norm)	77%	61.2%	46%
Weighted bilateral scoring with the VB method (with TZ-norm)	82.8%	65%	48%

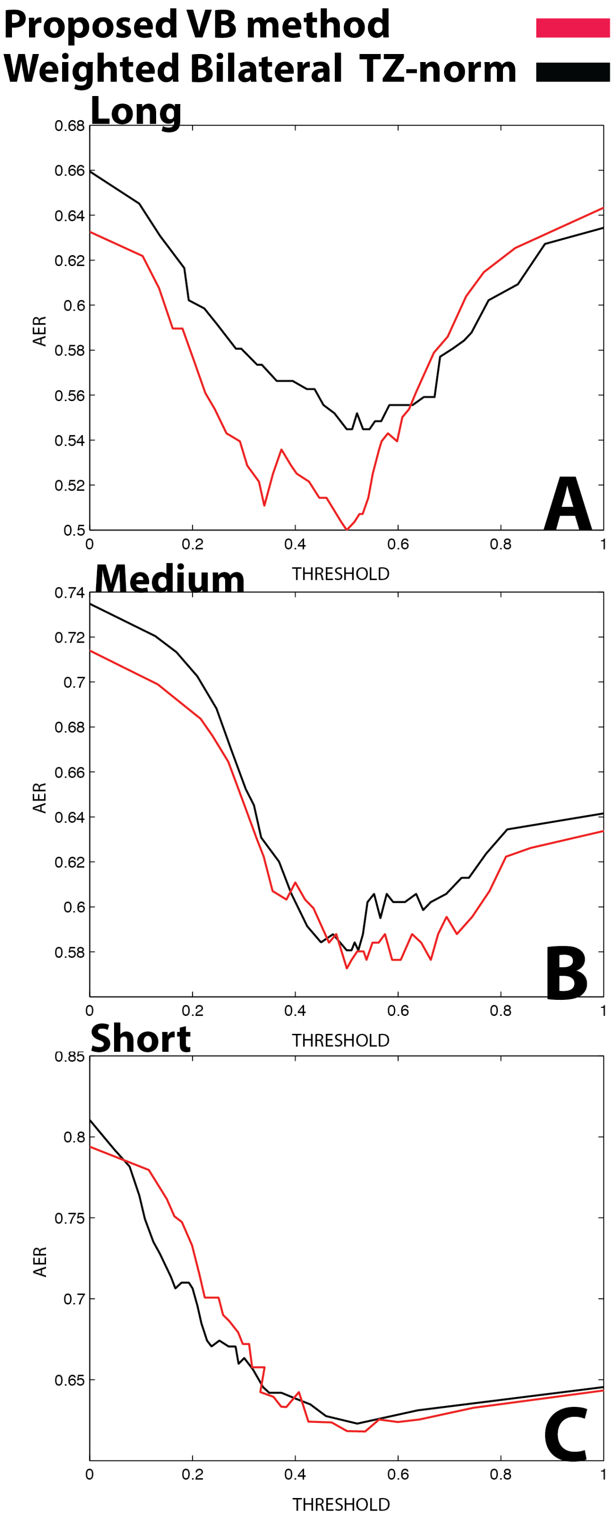

Figure 5.5 Performance of the proposed method based on WBS with TZ-norm. The subfigures A, B and C are the AER results under considered training conditions (i.e. long, medium, and short training data conditions).

The results for experiments with i-vector *with WCCN* are presented in Figure. 5.6 And Table 5.11. These results demonstrate that, as in the previous cases, the performance of i-vector depends on the duration of training material, i.e. as the reduction in the training material duration adversely affects both AER and the identification rate. It is also noted that, as expected, the i-vector approach out performs both the baseline and WBS classifiers. More importantly, the results clearly demonstrate that introduction of the proposed VB method also enhances the recognition performance of this state-of-the-art approach. In fact, it is noted that, in this case, the performance enhancement achieved in terms of both AER and the identification rate is higher than those for other classifiers considered.

Table 5.11 Identification rates for the i-vector classifier with and without The VB method

	Long training condition	Medium training condition	Short training condition
i-vector (with WCCN)	80%	65%	48%
i-vector with the VB method (with WCCN)	85%	69%	54%

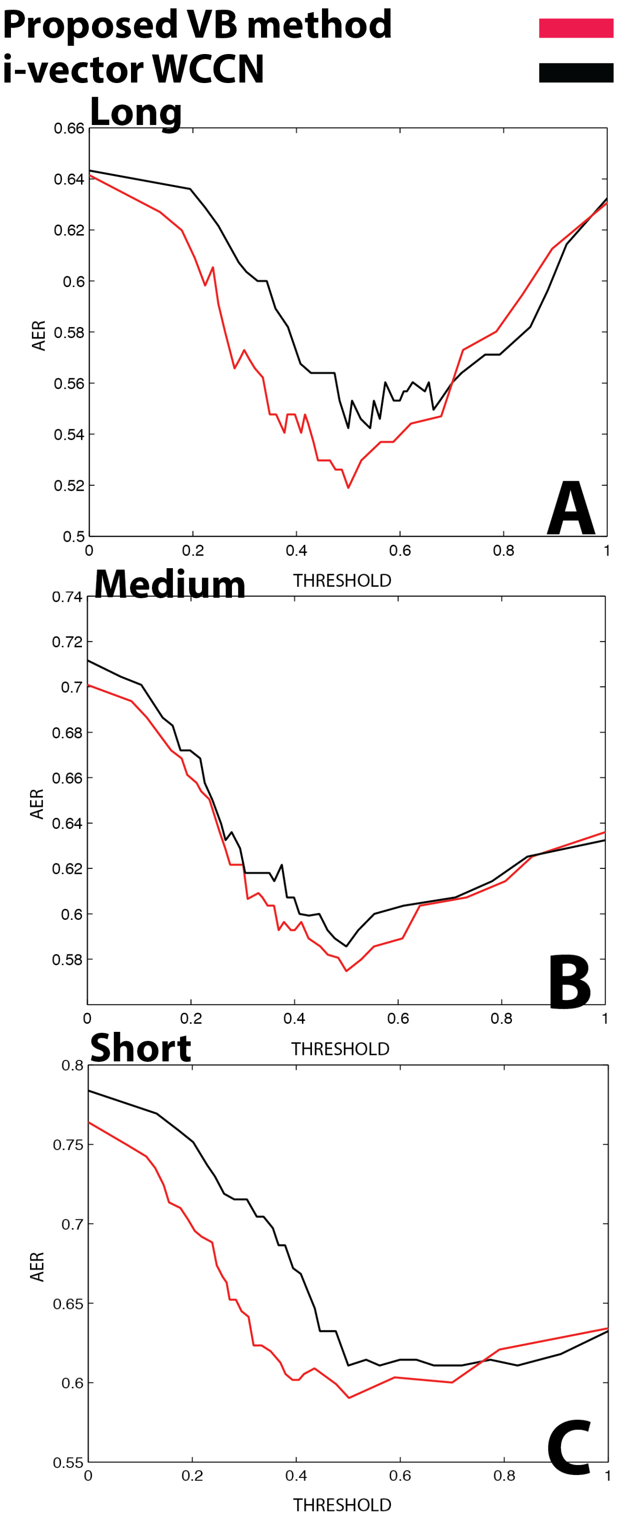

Figure 5.6 Performance of the proposed method based on i-vector with WCCN. The subfigures A, B and C are the AER results under considered training conditions (i.e. long, medium, and short training data conditions).

5.7.3. Effectiveness of VB in OSTI-SI with varied duration training material

The purpose of the set of experimental investigations presented here is to determine the effectiveness of the proposed vowel boosting approach in open-set speaker identification with diverse duration training material ranging from two to twenty seconds. The experimental results of the number of correctly identified speakers, for each classifier adopted in this study is presented in Table 5.12

These results clearly demonstrate that, similar to the findings in the previous set of experiments, the proposed method improves the identification rate in the first stage of OSTI-SI significantly (Table 5.12) and also results in a reduction in AER (Figure. 5.7). It should be noted that the test utterances belonging to each enrolled speaker originated from division of two recording sessions. The duration of each test utterance did not exceed 5 seconds. It is believed that in the case of longer duration test utterances, a further increase in identification rate can be achieved using the proposed method. The reason for this is that as the increase in the duration of the test utterance results in higher number of vowel phonemes. This in turn enhances the vowel representation and thereby provides a more effective means for boosting vowels, and lowers mislabelling error.

Table 5.12 Identification rates for the three considered classifiers with and without the VB method in experiments based on varied duration training data duration.

Classifier adopted	Standard performance	Proposed method of VB
GMM-UBM with TZ-norm	59%	63%
Weighted bilateral score with TZ-norm	60%	64%
i-vector with WCCN	62%	65%

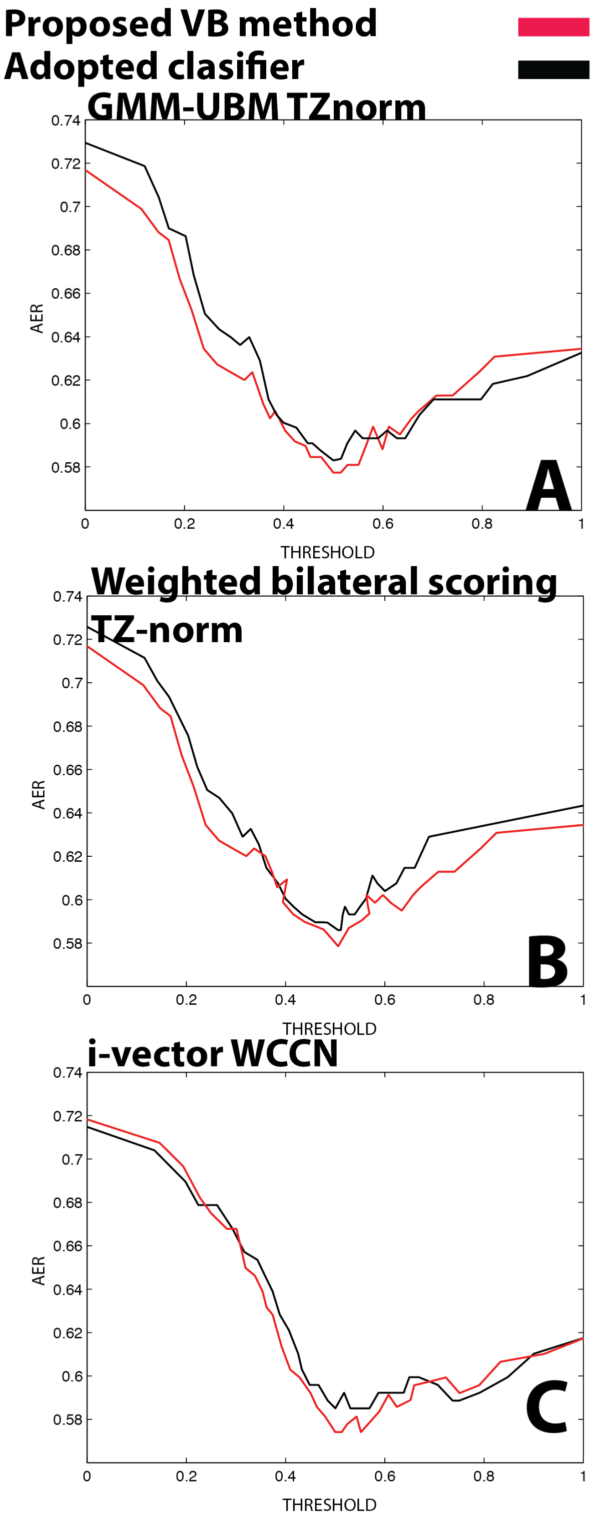

Figure 5.7 Performance improvement in AER offered by the proposed method under the mixed training data conditions. Where subfigure A is of GMM-UBM with TZ-norm, B is of Weighted bilateral scoring with TZ-norm and finally C is of the state of the are i-vector classifier

5.8. Summary

In summary, the novel “vowel boosting” method that enhances the vowels of speakers speech within an audio document and improve speaker identification process is proposed and thoroughly evaluated in this chapter. The experiments clearly demonstrate major improvement in speaker identification using short, medium, long and mixed duration reference material using the baseline GMM-UBM, Weighted bilateral scoring and the current state-of-the-art i-vector as seen in tables 5.9-5.12.

In comparison to previous work, long and short duration were reported as in table 5.1 with low performance observed using GMM-UBM and i-vector whereas in this work, the proposed vowel boosting method produces high identification performance using short, medium, long and mixed duration reference material with the baseline GMM-UBM, Weighted bilateral scoring and the current state-of-the-art i-vector . Future work to investigate the proposed method further is proposed in chapter 6.

CHAPTER SIX

6. Conclusions and future work

6. Conclusions and future work

The investigations in this study evolved around the need for more efficient speaker recognition when considering its application for recognition of individuals who are uncooperative. Obtaining data in uncontrolled environments without user cooperation is likely to introduce a great many artefacts and variations into a speech sample, such as variable channel conditions, background noise and varied duration (length) of training data. These all play a negative role on speaker recognition accuracy. Technically, the work carried out focussed on speaker identification as part of speaker recognition process in particular, reducing miss-labelling error that occurs during the identification stage of OS-SID process and it is unrecoverable,

The investigations began with experiments identifying the effect of environmental noise on the accuracy of open set speaker recognition as explained in Chapter 3. Three variations of noise (white noise, car noise and factory noise) were used to test their effects on the performance of the baseline and proposed state of the art methods of speaker recognition. The results presented a number of observations:

- White noise was found to affect all components of the audio document negatively. However white noise was also identified to be non-applicable in real word settings, and therefore would have little effect on the real-world application of speaker recognition of uncooperative subjects.
- Other forms of background noise (factory and car noise) were shown to have only partial effects on the audio document
- State of the art systems are conventionally deemed superior to the baseline systems however experiments comparing both systems in severe conditions of background noise found that the performance of the baseline system and current state of the art systems are very similar. The state of the art system gives only minimal improvements in recognition.

- Normalisation techniques for both classifiers play a significant role in improving system accuracy
- i-vector systems conventionally deemed as a more efficient alternative to baseline models, however experiments in this section demonstrated only a minimal improvement of system accuracy when the audio document has car noise added.
- However, i-vector outperformed the baseline model significantly when testing them on audio data including factory noise, because of the fact that factory noise is less monotonous the systems are better able to distinguish between audio data and background noise in comparison to the monotonous white and car noise.

As mentioned previously, another major challenge to speaker recognition performance is the duration of reference material, which in the case of the uncooperative subjects is highly variable and uncontrolled. It is argued that such an operating condition can significantly reduce the effectiveness of speaker identification by increasing the mislabelling (ML) error in the first stage of the process. For the application area considered in this study, the mislabelling error can indeed have severe consequences. This is because an ML error in the security application of OSTI-SI effectively means that the target of interest is completely missed in the first stage of the process. To address the problem, the adverse effects of varied duration training data were experimentally analysed, to understand their relative contributions to the recognition performance. The focus of experiments in Chapter 4 explored this challenge, by considering the effect of varied duration reference material on three systems, GMM-UBM (baseline), Weighted Bilateral Scoring with score normalisation (an extension of GMM-UBM) and i-vector with score normalisation (current state of the art system). The key findings of these experiments were as follows:

- When testing the performance of the three systems using sufficient reference material (60 seconds), as expected the current state of the art i-vector with score normalisation outperformed the baseline and weighted bilateral scoring systems.
- When reference material duration was decreased (1 second), the overall performance of all systems dropped dramatically, and there were some interesting observations. Firstly, weighted bilateral scoring with normalisation demonstrated a marginal improvement in performance as compared to the baseline system, and i-vector demonstrated a slight improvement over weighted bilateral scoring. Also, the performance of i-vector under such conditions is very similar to when applying score normalisation, hence there was no performance improvement gained with the additional computation costs.
- Similar results were obtained for when the systems were tested with varied duration reference material (1-30 seconds), where the performance of weighted bilateral scoring with score normalisation and i-vector without score normalisation was very similar. However, when applying score normalisation, i-vector marginally outperformed the other systems. Therefore, with the lack of normalisation data, there would be less computational cost to consider if using weighted bilateral scoring in comparison to i-vector, for the same performance.
- The last set of experiments explored variation in duration of both reference and test material of system performance. In such conditions, all systems performance dropped dramatically. Surprisingly, GMM-UBM baseline outperformed all other systems. What's more, the higher computation costs of i-vector with normalisation yielded no improvement on performance when compared to I-vector without normalisation.

The findings of this chapter demonstrated that there was no system that was compatible for speaker recognition when applied to the conditions expected from uncooperative subjects, and highly uncontrolled speech data. Therefore this prompted the exploration of an alternative system.

Speaker Recognition Biometrics

Based on research conducted, it was demonstrated that certain phonetic content provide more speaker information than others (see Chapter 5.3). An experiment was conducted (see Chapter 5.7) to test this, and results demonstrated that vowels contributed the most speaker information that assist in recognition performance. Therefore, a novel approach was explored named vowel boosting. The proposed vowel boosting method was applied to a range of classifiers adopted in this study, i.e. i-vector with WCCN, WBS and GMM-UBM with TZ-norm. Each range of classifiers adopted was tested under three adopted conditions: long, medium and short duration reference material. The proposed vowel boosting method was applied to each adopted condition on the same classifiers. The results were as follows:

- Regarding GMM-UBM with score normalisation baseline classifier, under conditions of long reference material a 5% improvement of identification rate was observed. In conditions of both medium and short duration reference material a 6% improvement was observed.
- For weighted bilateral scoring with score normalisation, under conditions of long, medium and short duration reference material, a 5.8%, 3.8% and 2% improvement in identification rate was observed respectively.
- For i-vector with score normalisation, in conditions of long, medium and short duration reference material, a 5%, 4% and 6% improvement in identification rate was observed respectively.

The findings of these experiments demonstrated the significant improvement in performance that vowel boosting gives for all adopted conditions and classifiers. Specifically, previous experiences of this study clearly demonstrated that the performance of the current state of the art recognition accuracy is subject to the duration of the reference material. Therefore, the increase in 6% identification rate in the case of short reference material when vowel boosting is applied to the current state of the art classifier is a significant achievement.

Speaker Recognition Biometrics

The second phase of experiments considered the effect of vowel boosting in a more realistic scenario, with varied duration reference material on the same adopted classifiers mentioned. The results were as follows:

- When vowel boosting was applied to the baseline GMM-UBM with score normalisation, and weighted bilateral scoring, a 4% improvement was observed in identification rate in both cases.
- For the state of the art i-vector with score normalisation an improvement of 5% was observed for the identification rate.

Therefore, the results demonstrate that the vowel boosting method has been shown to obtain significant improvements in identification rate for each training condition considered on the adopted classifiers. Most significantly, it has been shown to improve efficacy of current state of the art i-vector classifier.

For future work, investigation of the performance of the proposed VB method under varied channel (data obtained from example, telephone as reference material and data obtained from microphone as test material) characteristics and environmental noise will be explored. The environmental noises that can be investigated are more realistic noise where the amplitude of the noise varies, which will give a more realistic representation of the real world speaker scenarios. In addition, the investigation can also focus on varied reference and test material, which in the case of uncontrolled environment and uncooperative subjects is very likely.

7. References

[1] N. Lebovic, "Biometrics, or The Power of the Radical Center," *Critical Inquiry,* vol. 41, no. 4, pp. 841-868, 2015.

[2] M. G. Bulmer, *Francis Galton: pioneer of heredity and biometry*. JHU Press, 2003.

[3] S. Pillay, A. Ariyaeeinia, P. Sivakumaran, and M. Pawlewski, "Effective speaker verification via dynamic mismatch compensation," *Biometrics, IET,* vol. 1, no. 2, pp. 130-135, 2012.

[4] B. W. Husted, "Global Environmental and Social Strategy," *Global Strategy Journal,* vol. 3, no. 2, pp. 195-197, 2013.

[5] G. R. J. P. o. t. I. Doddington, "Speaker recognition—Identifying people by their voices," vol. 73, no. 11, pp. 1651-1664, 1985.

[6] K. H. Davis, R. Biddulph, and S. Balashek, "Automatic recognition of spoken digits," *The Journal of the Acoustical Society of America,* vol. 24, no. 6, pp. 637-642, 1952.

[7] P. Saini and P. Kaur, "Automatic speech recognition: A review," *International Journal of Engineering Trends and Technology,* vol. 4, no. 2, pp. 1-5, 2013.

[8] O. Rebollo, D. Mellado, and E. Fernández-Medina, "A Systematic Review of Information Security Governance Frameworks in the Cloud Computing Environment," *J. UCS,* vol. 18, no. 6, pp. 798-815, 2012.

[9] E. B. Bazan, "The Foreign Intelligence Surveillance Act: An Overview of the Statutory Framework and Recent Judicial Decisions," DEFENSE ACQUISITION UNIV FORT BELVOIR VA DAVID D ACKER LIBRARY AND KNOWLEDGE REPOSITORY2005.

[10] G. o. Sweden, "Act on Criminal Responsibility for Terrorist Offences," ed. http://www.government.se/contentassets/f84107eae6154ce19e65d64151a1b25f/act-on-criminal-responsibility-for-terrorist-offences.pdf, 2003.

[11] A. M. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, "Verification effectiveness in open-set speaker identification," *Vision, Image and Signal Processing, IEE Proceedings -,* vol. 153, no. 5, pp. 618-624, 2006.

[12] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, no. 4, pp. 788-798, 2011.

[13] K. Kichul and Y. Moo, "Robust speaker recognition against background noise in an enhanced multi-condition domain," *Consumer Electronics, IEEE Transactions on,* vol. 56, no. 3, pp. 1684-1688, 2010.

[14] N. McLaughlin, J. Ming, and D. Crookes, "Speaker recognition in noisy conditions with limited training data," in *Signal Processing Conference, 2011 19th European*, 2011, pp. 1294-1298.

[15] A. K. Singh, R. Singh, and A. Dwivedi, "Mel frequency cepstral coefficients based text independent Automatic Speaker Recognition using matlab," in *2014 International Conference on Reliability Optimization and Information Technology (ICROIT)*, 2014, pp. 524-527.

[16] N. M. AboElenein, K. M. Amin, M. Ibrahim, and M. M. Hadhoud, "Improved text-independent speaker identification system for real time applications," in *2016 Fourth International Japan-Egypt Conference on Electronics, Communications and Computers (JEC-ECC)*, 2016, pp. 58-62.

[17] S. Jongseo, K. Nam Soo, and S. Wonyong, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters,* vol. 6, no. 1, pp. 1-3, 1999.

[18] I. Arroabarren and A. Carlosena, "Voice Production Mechanisms of Vocal Vibrato in Male Singers," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 1, pp. 320-332, 2007.

[19] !!! INVALID CITATION !!!

[20] T. M. MP, "HM Government Transparency Report 2015: Disruptive and Investigatory Powers," British government, https://www.gov.uk/government/publications4 November 2015, Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/473603/51973_Cm_9151_Transparency_Accessible.pdf.

[21] (2001). *USA patriot act, Anti-Terrorism Logislation*.

[22] J.-C. Junqua and J.-P. Haton, *Robustness in automatic speech recognition: fundamentals and applications*. Springer Science & Business Media, 2012.

[23] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on speech and Audio Processing,* vol. 7, no. 6, pp. 697-708, 1999.

[24] C. Gussenhoven, *The phonology of tone and intonation*. Cambridge University Press, 2004.

[25] E. Sapir, "The psychological reality of phonemes," *Selected writings of Edward Sapir in language, culture, and personality,* pp. 46-60, 1949.

[26] H. Veisi and H. Sameti, "Hidden-Markov-model-based voice activity detector with high speech detection rate for speech enhancement," *Signal Processing, IET,* vol. 6, no. 1, pp. 54-63, 2012.

[27] M. Asgari, A. Sayadian, F. Tehranipour, and A. Mostafavi, "Novel Voice Activity Detection Based on Vector Quantization," in *Computer Modelling and Simulation, 2009. UKSIM '09. 11th International Conference on*, 2009, pp. 255-257.

[28] A. M. Aibinu, M. J. E. Salami, A. R. Najeeb, J. F. Azeez, and S. M. A. K. Rajin, "Evaluating the effect of voice activity detection in isolated Yoruba word recognition system," presented at the Mechatronics (ICOM), 2011 4th International Conference On, 17-19 May 2011, 2011.

[29] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *the Journal of the Acoustical Society of America,* vol. 55, no. 6, pp. 1304-1312, 1974.

[30] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *The journal of the acoustical society of America,* vol. 50, no. 2B, pp. 637-655, 1971.

[31] L. Rabiner, *Fundamentals of speech recognition*. 1993.

[32] Y. D. Cho, K. Al-Naimi, and A. Kondoz, "Improved voice activity detection based on a smoothed statistical likelihood ratio," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, vol. 2, pp. 737-740 vol.2.

[33] B. V. Harsha, "A noise robust speech activity detection algorithm," presented at the Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on, 20-22 Oct. 2004, 2004.

[34] H. Matsumoto and M. Moroto, "Evaluation of mel-LPC cepstrum in a large vocabulary continuous speech recognition," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on*, 2001, vol. 1, pp. 117-120 vol.1.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 28, no. 4, pp. 357-366, 1980.

[36] S. Ghai and R. Sinha, "Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition," in *2010 International Conference on Signal Processing and Communications (SPCOM)*, 2010, pp. 1-5.

[37] H. Trang, L. Tran Hoang, and N. Huynh Bui Hoang, "Proposed combination of PCA and MFCC feature extraction in speech recognition system," in *2014 International Conference on Advanced Technologies for Communications (ATC 2014)*, 2014, pp. 697-702.

[38] p.sivakumaran, "robust text-dependant speaker verification," 1998.

[39] C. Hanilci and F. Ertas, "Impact of voice excitation features on speaker verification," in *Electrical and Electronics Engineering (ELECO), 2011 7th International Conference on*, 2011, pp. II-157-II-160.

[40] J. Li, Y. Tian, and L. Zhang, "Research and implementation of speaker recognition algorithm based on FPGA," in *2012 24th Chinese Control and Decision Conference (CCDC)*, 2012, pp. 1155-1158.

[41] S. Maraboina, D. Kolossa, P. K. Bora, and R. Orglmeister, "Multi-speaker voice activity detection using ICA and beampattern analysis," in *Signal Processing Conference, 2006 14th European*, 2006, pp. 1-5.

[42] K. Ishizuka, S. Araki, and T. Kawahara, "Speech Activity Detection for Multi-Party Conversation Analyses Based on Likelihood Ratio Test on Spatial Magnitude," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 18, no. 6, pp. 1354-1365, 2010.

[43] S. S. Bharali and S. K. Kalita, "Zero crossing rate and short term energy as a cue for sex detection with reference to Assamese vowels," in *Convergence of Technology (I2CT), 2014 International Conference for*, 2014, pp. 1-4.

[44] S. A. Soleimani and S. M. Ahadi, "Voice Activity Detection based on Combination of Multiple Features using Linear/Kernel Discriminant Analyses," in *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, 2008, pp. 1-5.

[45] H. Othman and T. Aboulnasr, "A Gaussian/Laplacian hybrid statistical voice activity detector for line spectral frequency-based speech coders," in *Circuits and Systems, 2003 IEEE 46th Midwest Symposium on*, 2003, vol. 2, pp. 693-696 Vol. 2.

[46] E. Dong, G. Liu, Y. Zhou, and X. Zhang, "Applying support vector machines to voice activity detection," in *Signal Processing, 2002 6th International Conference on*, 2002, vol. 2, pp. 1124-1127 vol.2.

[47] Q. H. Jo, J. H. Chang, J. W. Shin, and N. S. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing,* vol. 3, no. 3, pp. 205-210, 2009.

[48] Amit S. Malegaonkar, "Efficient Speaker Change Detection Using Adapted Gaussian Mixture Models," *IEEE,* vol. 15, AUGUST 2007.

[49] K. M. Chugg and A. Polydoros, "Front-end processing for joint maximum likelihood channel and sequence estimation," in *Global Telecommunications Conference, 1994. Communications Theory Mini-Conference Record, 1994 IEEE GLOBECOM., IEEE*, 1994, pp. 51-55.

[50] R. Vergin and D. O'Shaughnessy, "Pre-emphasis and speech recognition," vol. 2, pp. 1062-1065: IEEE.

[51] B. Upc, "Analysis of voice signals for the Harmonics-to-noise Crossover Frequency," 2008.

[52] D. O. Shaughnessy, *Speech Communication: Human and Machine*. Massachusetts: Addison-Wesley Publishing Company.

[53] R. W. Hamming, *Digital filters*. Courier Corporation, 1989.

[54] B. S. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE,* vol. 64, no. 4, pp. 460-475, 1976.

[55] J. Schroeter and M. M. Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions on Speech and Audio Processing,* vol. 2, no. 1, pp. 133-150, 1994.

[56] M. Antal and G. Toderean, "Speaker Recognition and Broad Phonetic Groups," in *SPPRA*, 2006, pp. 155-159.

[57] W. Zufeng, L. Lin, and G. Donghui, "Speaker recognition using weighted dynamic MFCC based on GMM," in *Anti-Counterfeiting Security and Identification in Communication (ASID), 2010 International Conference on*, 2010, pp. 285-288.

[58] A. Mezghani and D. O'Shaughnessy, "Speaker verification using a new representation based on a combination of MFCC and formants," in *Electrical and Computer Engineering, 2005. Canadian Conference on*, 2005, pp. 1461-1464.

[59] J. Fortuna, "SPEAKER INDEXING BASED ON VOICE BIOMETRICS," Electronics, communication and Electrical Engineering, hertfordshire university, March 2006

[60] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," presented at the Signal Processing and Communication Systems (ICSPCS), 2010 4th International Conference on, 13-15 Dec. 2010, 2010.

[61] K. R. Rao and P. Yip, *Discrete cosine transform: algorithms, advantages, applications*. Academic press, 2014.

[62] J. Fortuna, "Speaker Indexing Based on Voice Biometrics," PhD, Electronic, Communication and Electrical Engineering, University of Hertfordshire, March 2006.

[63] f. Bimbot, "A tutorial on text-independent speaker verification," *EURASIP,* 2004.

[64] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE,* vol. 63, no. 4, pp. 561-580, 1975.

[65] A. Harma, "Linear predictive coding with modified filter structures," *Speech and Audio Processing, IEEE Transactions on,* vol. 9, no. 8, pp. 769-777, 2001.

[66] j. G.Proakis, *Digital Signal Processing*. 2007.

[67] Y. Yuan, P. Zhao, and Q. Zhou, "Research of speaker recognition based on combination of LPCC and MFCC," in *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, 2010, vol. 3, pp. 765-767.

[68] Y. Hai-Yan and X. X. Jing, "Performance test of parameters for speaker recognition system based on SVM-VQ," in *2012 International Conference on Machine Learning and Cybernetics*, 2012, vol. 1, pp. 321-325.

[69] T. D. Pham, "LPC cepstral distortion measure for protein sequence comparison," *NanoBioscience, IEEE Transactions on,* vol. 5, no. 2, pp. 83-88, 2006.

[70] J. G. M. Britto and S. S. Kumar, "Speaker change detection - an comparative study using support vector machines," in *Sustainable Energy and Intelligent Systems (SEISCON 2012), IET Chennai 3rd International on*, 2012, pp. 1-5.

[71] S. V. Chapaneri, "Spoken digits recognition using weighted MFCC and improved features for dynamic time warping," *International Journal of Computer Applications,* vol. 40, no. 3, pp. 6-12, 2012.

[72] U. Bhattacharjee, "A comparative study of LPCC and MFCC features for the recognition of Assamese phonemes," *International Journal of Engineering Research and Technology,* vol. 2, no. 1, pp. 1-6, 2013.

[73] P. S. J. Fortuna, A. M. Ariyaeeinia and A. Malegaonkar, "RELATIVE EFFECTIVENESS OF SCORE NORMALISATION METHODS IN OPEN-SET SPEAKER IDENTIFICATION," ed, 2004.

[74] S. G. Pillay, A. Ariyaeeinia, and M. Pawlewski, "Effectiveness of speaker-dependent feature score pruning in speaker verification," in *Communications, Control and Signal Processing, 2008. ISCCSP 2008. 3rd International Symposium on*, 2008, pp. 372-376.

[75] J. S. Fortuna, P. / Ariyaeeinia, A. / Malegaonkar, "Open-Set Speaker Identification Using Adapted Gaussian Mixture Models," presented at the INTERSPEECH-2005, Lisbon, Portugal, 2005.

[76] H. Do, I. Tashev, and A. Acero, "A new speaker identification algorithm for gaming scenarios," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5436-5439.

[77] F. Răstoceanu and M. Lazăr, "Score fusion methods for text-independent speaker verification applications," in *2011 6th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, 2011, pp. 1-6.

[78] C. Gao, G. Saikumar, A. Srivastava, and P. Natarajan, "Open-set speaker identification in broadcast news," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5280-5283.

[79] R. Chakroun, L. B. Zouari, M. Frikha, and A. B. Hamida, "A novel approach based on Support Vector Machines for automatic speaker identification," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, 2015, pp. 1-5.

[80] N. D. Dehak, Réda / Kenny, Patrick / Brümmer, Niko / Ouellet, Pierre / Dumouchel,, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," presented at the INTERSPEECH-2009, Brighton, United Kingdom, 2009.

[81] S. S. Khanwalkar, B. Y. Smolenski, R. E. Yantorno, and S. J. Wenndt, "Enhancement of Speaker Identification using SID-usable speech," in *2005 13th European Signal Processing Conference*, 2005, pp. 1-4.

[82] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery," *PLOS ONE,* vol. 11, no. 7, p. e0158520, 2016.

[83] S. Kanrar, "i Vector used in Speaker Identification by Dimension Compactness," presented at the ARXIV, 2017.

[84] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4516-4519.

[85] P.-M. M. Bousquet, Driss / Bonastre, Jean-François, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," presented at the INTERSPEECH-2011, Florence, Italy, 2011.

[86] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 6788-6791.

[87] R. Karadaghi, H. Hertlein, and A. Ariyaeeinia, "Effectiveness in open-set speaker identification," in *Security Technology (ICCST), 2014 International Carnahan Conference on*, 2014, pp. 1-6.

[88] A. Kanagasundaram, R. Vogt, D. B. Dean, and S. Sridharan, "i-vector based speaker recognition on short utterances," presented at the Proceedings of the 12th Annual Conference of the International Speech Communication A ssociation, Firenze Fiera, Florence, 2011.

[89] R. Karadaghi, H. Hertlein, and A. Ariyaeeinia, "Open-set speaker identification with diverse-duration speech data," in *SPIE Defense + Security*, 2015, vol. 9457, p. 7: SPIE.

[90] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification," presented at the 10th Annual Conference of the International Speech Communication Association (Interspeech), Brighton, United Kingdom, 2009.

[91] D. A. Reynolds, "Speaker Verification Using Adapted GaussianMixture Models," a. R. B. D. Thomas F. Quatieri, Ed., ed, 2000.

[92] A. Miró, "Robust speaker diarization for meetings," Department of Signal Theory and Communications, Universitat Politècnica de Catalunya, http://www.xavieranguera.com/phdthesis/node108.html, 2006.

[93] N. Fatima, X. Wu, T. F. Zheng, C. Zhang, and G. Wang, "A universal phoneme-set based language independent short utterance speaker recognition," in *11th National Conference on Man-Machine Speech Communication (NCMMSC'11), Xi'an, China*, 2011, pp. 16-18: Citeseer.

[94] A. Matza and Y. Bistritz, "Speaker recognition with rival penalized EM training," in *2011 IEEE International Workshop on Machine Learning for Signal Processing*, 2011, pp. 1-6.

[95] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," presented at the Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on, 17-21 May 2004, 2004.

[96] R. Vogt, "Factor Analysis Modelling for Speaker Verification with Short Utterances," *ieee,* 2008.

[97] K. Li and E. Wrench, Jr., "An approach to text-independent speaker recognition with short utterances," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.*, 1983, vol. 8, pp. 555-558.

[98] N. Fakotakis and J. Sirigos, "A high performance text independent speaker recognition system based on vowel spotting and neural nets," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, vol. 2, pp. 661-664 vol. 2.

[99] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on,* vol. 3, no. 1, pp. 72-83, 1995.

[100] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, no. 1–3, pp. 19-41, 2000.

[101] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna, "On the enhancement of speaker identification accuracy using weighted bilateral scoring," presented at the Security Technology, 2008. ICCST 2008. 42nd Annual IEEE International Carnahan Conference on, 13-16 Oct. 2008, 2008.

[102] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna, "Unsupervised speaker change detection using probabilistic pattern matching," *Signal Processing Letters, IEEE,* vol. 13, no. 8, pp. 509-512, 2006.

[103] E. S. Parris and M. J. Carey, "Multilateral techniques for speaker recognition," presented at the 5th International Conference on Spoken Language Processing (ICSLP), Sydney, Australia, 1998.

[104] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalisation methods in open-set speaker identification," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[105] L. Chen and Y. Yang, "Applying emotional factor analysis and I-vector to emotional speaker recognition," *Biometric Recognition,* pp. 174-179, 2011.

[106] V. Alonso Moreno, "Joint factor analysis for forensic automatic speaker recognition," 2011.

[107] P. Kenny, "Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms " CRIMP-06/082006.

[108] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," presented at the 12th Annual Conference of the International Speech Communication Association (Interspeech), Florence, Italy, 2011.

[109] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," pp. 293-298.

[110] P. Kenny, N. Dehak, V. Gupta, and P. Dumouchel, "A new training regimen for factor analysis of speaker variability," *Proc. ICASSP 2008,* 2008.

[111] A. Kanagasundaram, D. Dean, R. Vogt, M. McLaren, S. Sridharan, and M. Mason, "Weighted LDA techniques for i-vector based speaker verification," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25-30 March 2012, 2012.

[112] A. Malegaonkar and A. Ariyaeeinia, "Performance Evaluation in Open-Set Speaker Identification," presented at the The Third European Workshop on Biometrics and Identity Management (BioID 2011), Brandenburg, Germany, 2011.

[113] O. Galibert, "Methodologies for the evaluation of speaker diarization and automatic speech recognition in the presence of overlapping speech," pp. 1131-1134.

[114] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW)*, 2000.

[115] R. Karadaghi, H. Hertlein, and A. Ariyaeeinia, "Open-set speaker identification with diverse-duration speech data," 2015, vol. 9457, pp. 94570G-94570G-7.

[116] Y. Lei, L. Burget, L. Ferrer, M. Graciarena, and N. Scheffer, "Towards noise-robust speaker recognition using probabilistic linear discriminant analysis," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4253-4256: IEEE.

[117] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4257-4260: IEEE.

[118] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 6788-6791: IEEE.

[119] C. Yu, G. Liu, S. Hahm, and J. H. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 4017-4021: IEEE.

[120] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems," *Digital Signal Processing,* vol. 10, no. 1–3, pp. 42-54, 2000.

[121] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalisation methods in open-set speaker identification," presented at the ODYSSEY 2004 - The Speaker and Language Recognition Workshop, Toledo Spain, 2004.

[122] "The NIST Year 2008 Speaker Recognition Evaluation Plan,"

[123] "The NIST Year 2005 Speaker Recognition Evaluation Plan,"

[124] A. Varga and H. J. M. Steenken, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *Speech Communication,* pp. 247-252, 1993.

[125] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication,* vol. 59, pp. 69-82, 2014.

[126] G. Bhattacharya, J. Alam, and P. Kenny, "Deep speaker embeddings for short-duration speaker verification," in *Proc. Interspeech*, 2017, pp. 1517-1521.

[127] A. Kanagasundaram, D. Dean, S. Sridharan, and C. Fookes, "Domain adaptation based Speaker Recognition on Short Utterances," *arXiv preprint arXiv:1610.02831,* 2016.

[128] C. Zhang, X. Li, W. Li, P. Lu, and W. Zhang, "A novel i-vector framework using multiple features and PCA for speaker recognition in short speech condition," in *Audio, Language and Image Processing (ICALIP), 2016 International Conference on*, 2016, pp. 499-503: IEEE.

[129] N. Fatima and T. F. Zheng, "Vowel-category based Short Utterance Speaker Recognition," in *Systems and Informatics (ICSAI), 2012 International Conference on*, 2012, pp. 1774-1778.

[130] P. Delattre, A. M. Liberman, F. S. Cooper, and L. J. Gerstman, "An experimental study of the acoustic determinants of vowel color; observations on one-and two-formant vowels synthesized from spectrographic patterns," *Word,* vol. 8, no. 3, pp. 195-210, 1952.

[131] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4249-4252.

[132] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 15, no. 4, pp. 1435-1447, 2007.

[133] J. Ellenbogen, *Reasoned and Unreasoned Images: The Photography of Bertillon, Galton, and Marey*. Penn State Press, 2012.

[134] O. S. Adeoye, "A survey of emerging biometric technologies," *International Journal of Computer Applications,* vol. 10, 2010.

[135] J. M. Naik, "Speaker verification: a tutorial," *IEEE Communications Magazine,* vol. 28, no. 1, pp. 42-48, 1990.

[136] P. Domain, "Public Domain.“Biography: Alphonse Bertillon (1853-1914)”," ed: https://www.nlm.nih.gov/visibleproofs/galleries/biographies/bertillon.html, 2014.

[137] A. Bertillon, *Alphonse Bertillon's Instructions for Taking Descriptions for the Identification of Criminals, and Others*. Ams PressInc, 1977.

[138] C. L. Lofdahl, "On the environmental externalities of global trade," *International Political Science Review,* vol. 19, no. 4, pp. 339-355, 1998.

[139] F. Asche, C. A. Roheim, and M. D. Smith, "Trade intervention: Not a silver bullet to address environmental externalities in global aquaculture," *Marine Policy,* vol. 69, pp. 194-201, 2016.

[140] S. K. Kopparapu, A. Imran, and G. Sita, "A two pass algorithm for speaker change detection," in *TENCON 2010 - 2010 IEEE Region 10 Conference*, 2010, pp. 755-758.

[141] A. G. Adam, S. S. Kajarekar, and H. Hermansky, "A new speaker change detection method for two-speaker segmentation," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, 2002, vol. 4, pp. IV-3908-IV-3911.

Appendix

Appendix A

The term “Biometrics” is composed of two Greek words − Bio (Life in the human context) and Metrics (that by which anything is Measured). In modern usage, “*Biometrics*” is a term that literally describes the metrics of human characteristics, specifically referring to technologies used to detect and recognize human characteristics [133]. A particular type of biometrics, “*Voice Biometrics*”, is a term that is used to describe several technologies which can look for, identify, or authenticate unique speech patterns belonging to an individual. *Voice Biometrics* may be employed in several ways: for authentication for access (in the way physical passwords, pass cards, fingerprint, or retinal scans are used), for speech recognition in listening applications, for cellular and secure voice communications, or for speaker recognition (who’s talking and/or who said what?) or authentication (who is this person and are they who they *claim* they are?). These speech patterns are made up of identifiable characteristics (biometric identifiers), which are statistically unique to each person, and thus may be used for authentication, access, and/or verification. Note that these *biometric identifiers* are assumed distinctive and measurable, and can be further classified as either physiological (body shape, size, symmetry), or behavioural (unique gestures, vocal tics and habits, or emotional state) [134].

Each human being has a set of unique characteristics, which distinguish each individual (him/her) from one another [135]. These unique characteristics can be categorised into physical attributes and behavioural characteristics. The physical attributes include a wide range of features, including vocal cavity geometry, finger print patterns, palm print, hair colour, iris colour, retinal configuration, and hand geometry. In contrast, the behavioural characteristics include attributes that distinguish a person from the rest of humanity, such as their tone, pitch, mode, and accent of speech, their way of typing on a key board, the way they walk, or their signature.

The interest in using biometrics for identification has existed and been pursued for over 100 years, and grew out of efforts in the 19th century to explore characteristics of humans which might differ between different groups in society. Biometrics as a discipline developed from the work of Alphonse Bertillon (1853–1914), a French police officer and pioneering biometric researcher in the 19th century [1]. Bertillon was the inventor of the mug shot, but is best known for applying anthropological techniques to law enforcement, in one of the earliest known attempts to build a criminal profiling database. Bertillon attempted to establish a "science of identity" by making photographic records of criminal bodies and analysing physical characteristics and comparing those against characteristics of other criminals as well as individuals which had not been known to commit any crime. Contributions Bertillon made to the development of biometrics and authentication include:

- Developing a photographic method using a camera on a high tripod to capture the details of a crime scene before it could be disturbed by investigators, for later evaluation and study
- Developing the practice of gridding off and measuring features in the scene for later analysis and classification.
- Developing a physical measurements system to be used for identification of unique human characteristics belonging to specific individuals [1, 4, 5].

Further, the development and adoption of various biometric techniques (handwriting analysis, galvanoplastic compounds to preserve footprints and other impressions, ballistics, and a dynamometer for breaking force measurement) [4, 5]

His contemporary, '[Sir Frances] Galton, attempted, in the same period, to create accurate yet abstract images of such entities as "the criminal" and "the lunatic".' Francis Galton (1822-1911) pursued a variety of other topics and interest in what is now known as biometric science. He:

- Created the statistical concept of correlation (e.g., the usage of line regression lines R or R^2 statistic) [136, 137].
- was the first to apply statistical methods to the study of human differences,
- introduced the use of questionnaires and surveys for collecting data on human communities, and
- introduced the use of line regression line as well as the concept of the "r" correlation coefficient (R or R^2 statistic) [136, 137], and
- Contributed to the body of knowledge in psychology and the science of differences [3].

Globalisation – the interdependence of world views, products, ideas, and culture, coupled with easy accessibility of transportation and communications - is an ever-expanding phenomenon, which confers both positive and negative benefits on the functioning of civilization. One of these is the ease with which global crime may now be committed. There are four aspects of globalization which make this possible: trade and transactions, capital and investment flow, human migration and travel, and dissemination of knowledge [2], which facilitate the globalization culture, politics, commerce, poverty, and increasing inequality, all of which can also contribute to globalization of crime. Also, environmental challenges such as global warming/climate change, deforestation, resource pollution and degradation (cross-boundary water and air contamination, for example), and overfishing of the ocean are also linked as consequences of globalization [138, 139].

Appendix B

The final process step, prior to speaker recognition processing, is the speaker change detection (SCD) process. The audio document obtained under uncontrolled conditions is also likely to have been obtained under conditions where there is more than one speaker present. For this reason, it is necessary to pass the audio document through an SCD system. Because in many cases the audio document contains more than one speaker, it is essential to determine with a high degree of certainty, when a speaker change occurs in the audio document [48]. To identify the point of change in speaker, one task requires the segregating of parts of the document corresponding to homogenous speakers, which results in different segments classified as belonging to different speakers. This is an essential stage in speaker recognition and it is very crucial to correctly identify the sub-segment belonging to each speaker. Missed points around a speaker change or false detection of speaker change points where there has been none can adversely affect the performance of the system. Thus, Speaker change detection (SCD) is an essential stage in the speaker recognition process. It is also used in the areas of speaker diarization and automatic transcription of audio recordings [48, 140].

Speaker diarization is the process of partitioning an input audio stream, such as a speaker document, into homogeneous segments, as a function of the speaker's identity. Speaker diarization can be used to optimize or enhance an automatic speech transcription by structuring the audio stream into speaker turns [48]. When diarization is used together with speaker recognition systems, it can potentially answer the question "who spoke when?" and describes a process that combines the speaker segmentation task and the speaker clustering task in speaker recognition. The first task aims at finding speaker change points in an audio stream. The second task aims at grouping together speech segments on the basis of speaker characteristics [48].

Existing SCD approaches are based on the exploition of dissimilarities detected in the distribution of the data signal before and after a speaker change point. How that determination depends upon the classification method. Patterns may be detected and extracted from the data around the speaker change point and used to represent confirmed examples for recognition of the change point [111]. The recognized patterns extracted from data between recognized speaker change points may represent confirmed negative examples. The experimentally defined positive and negative example utterances, once collected, are subsequently used with the Support Vector Machine (SVM) to build the speaker change detection (SCD) model. Finally, the trained SVM is used to scan and analyse the continuous speech signal in a multi-speaker data document, and process it to find statistically likely points of speaker change and extract statistically homogenous, fixed-length samples of their speech [111]. These are, in turn, input into the SVM after extraction. The SVM uses them to classify the speaker change points and no-change points, refining on speaker features. In order to optimally perform this analysis two separate speaker conversations are required; however, in practice, these won't be available, more often than not.

There are a number of other modelling techniques proposed to detect points of speaker changes in a given audio document, which involve attempting to measure the dissimilarities between two consecutive segments of a parameterized signal to decide if these segments correspond to the same speaker or to two different speakers. The initial approach to this process involves sliding an analysis window through the audio stream and measuring the similarity between the adjacent subsets of the data within it, at each window position [141]. One of the most popular of the modelling techniques is the Bayesian Information Criteria (BIC) method. Its popularity lies in how well it performs in identifying acoustic change as well as speaker change. Further attempts to enhance the SCD performance has involved using a combination of distance measurement and

BIC [48, 140].such as XBIC. XBIC is a measure derived from comparing BIC with a distance measure of HMM [111]. Inverse Gaussian analysis(IGA) in conjunction with BIC is also used to reduce computational cost [48, 140]. BIC has been applied with a 'Divide and Conquer' strategy, which was shown to improve the vocal segmentation [111].In recent years as an alternative to the afore-mentioned methods, bilateral scoring-based speaker change detection (BLS-SCD), has been used. It is based on employing a probabilistic pattern matching approach, and has been shown to out-perform BIC and XBIC. BLS-SCD is an improvement on the Unilateral Scoring Method [48, 140]. It is also a more suitable method for SCD as it offers reciprocity of speakers. It is further improved by changing the statistical speaker representation from a single Gaussian model to a Gaussian Mixture Model (GMM) using a single-step Bayesian adaptation of a Universal Background Model (UBM). In recent years, there has been an attempt at modelling the segments using SVM. This approach claims better performance in handling the data insufficiency, when compared to the regular use of GMM [111].