Analysis of Voiceprint and Other Biometrics for Criminological and Security Applications

Farbod Hossyndoust Foomany

A thesis submitted to the University of Huddersfield in partial fulfillment of the requirements for the degree of Doctor of Philosophy

The University of Huddersfield
July 2010

Contents

Word count (excluding references and appendices but including footnotes and endnotes): 76000 words

Abstract

This Thesis examines the role and limitations of voice biometrics in the contexts of security and for crime reduction. The main thrust of the Thesis is that despite the technical and non-technical hurdles that this research has identified and sought to overcome, voice can be an effective and sustainable biometric if used in the manner proposed here. It is contended that focused and continuous evaluation of the strength of systems within a solid framework is essential to the development and application of voice biometrics and that special attention needs to be paid to human dimensions in system design and prior to deployment.

Through an interdisciplinary approach towards the theme reflected in the title several scenarios are presented of the use of voice in security / crime reduction, crime investigation, forensics and surveillance contexts together with issues surrounding their development and implementation.

With a greater emphasis on security-oriented voice verification (due to the diversity of the usage scenarios and prospect of use) a new framework is presented for analysis of the reliability and security of voice verification.

This research calls not only for a standard evaluation scheme and analytical framework but also takes active steps to evaluate the prototype system within the framework under various conditions. Spoof attacks, noises, coding, distance and channel effects are among the factors that are studied. Moreover, an additional under-researched area, the detection of counterfeit signals, is also explored.

While numerous technical and design contributions made in this project are summarised in chapter 2, the research mainly aims to provide solid answers to the high-level strategic questions. The Thesis culminates in a synthesis chapter in which realistic expectations, design requirements and technical limitations of the use of voice for criminological and security applications are outlined and areas for further research are defined.

Copyright Statement

Acknowledgments

It is a great pleasure to thank those who made writing this Thesis possible.

I am heartily thankful to my supervisors, Professor Alex Hirschfield and Dr. Michael Ingleby at the University of Huddersfield for accepting to supervise this project in the first place, for their invaluable guidance and advice over the past few years, and for generously sharing their expertise, knowledge and insight with me throughout this research.

I feel indebted to all my previous professors and supervisors without whom the completion of this work was impossible. Especially I would like to thank Dr. Umapathy, Dr. Naghian, Dr. Teshnehlab, Dr. Shams, Dr. Setayeshi and Dr. Vosoughi.

I would like to show my gratitude to my supervisors, project managers and colleagues at Pooya Company and Basamad-negar Company for providing me with the opportunity to gain insight into and obtain an in-depth knowledge of signal processing of speech and security concepts in practice. From many whom I can name I would like to especially thank Dr. Barari, Dr Eftekhar, Mr. Mohammad-Hossein, Mr. Mortazavi, Mr. Mirbaha, Ms. Zamani, Mr Nasir-Zadeh, and Mr. Farahbakht.

I am grateful to Dr. Pitchford, Head of Research Administration at the University of Huddersfield for his precious and continuous guidance and support throughout this project.

I would like to thank Dr. Alina Haines, Michelle Rogerson, Dr. Vicky Heap, Dr. Rachel Armitage and Kris Christmann at the Applied Criminology Center, University of Huddersfield, for their help and especially for offering me guidance on a range of social science topics, in the past years.

I always find it impossible to thank my parents as much as they deserve it. Like ever, as heartily as possible and as modest as words could be I would like to thank my parents sincerely for everything that they have done for me throughout my life.

Finally, I would like to offer my regards and blessings to all of who supported me in any way during the completion of this project.

List of Tables

List of Figures

Chapter 1 : Introduction

1.1 Introduction

It would be impossible to maintain that the fields of criminology and security ever will become liberated from perpetual quest for new measures to establish security, reduce crime and facilitate investigation of various types of fraud and crime. In this never-ending quest, in recent years a fairly new concept has emerged that has attracted strong opposition while gaining its own advocates. Use of human biological traits has migrated from science fiction novels and movies to the heart of modern societies where everyday transactions are taking place.

Biometric identification, verification and surveillance are scientific topics in nature, but more profoundly and subtly they are about people and this fact causes them to stretch out in several dimensions which have to be thoroughly researched prior to putting these concepts into practice. While it is true that, on a practical level, the development and effective deployment of new security measures requires technical and scientific knowledge to develop robust solutions and an awareness of the ways in which offenders are able to circumvent and breach defensive measures to commit crime (the so-called 'arms race'), on another level the non-technical issues are in constant collaboration with technical demands and they continually redefine the shape and scope of the technical problems.

For the above reasons and others detailed in this chapter and later throughout the thesis, research on biometrics, unless confined to optimisation of algorithms, is inevitably multi-dimensional and interdisciplinary. This research, therefore, not only scrutinises the topics related to the development of robust voice recognition technologies but looks at the human and criminological contexts for their deployment as well.

This chapter offers a succinct introduction to biometrics, this research and this thesis. All the concepts briefly touched upon here are described in greater detail in respective chapters in the thesis.

1.2 Introduction to Biometrics and Voice

Maltoni et al (2003) define Biometric recognition as the use of distinctive physiological and behavioral characteristics for automatically recognizing individuals. The recognition is facilitated by measurement of those characteristics that are called biometric identifiers (or simply biometrics).

The most important questions about Biometrics are simple ones. Do they provide enough distinctive information about the person they belong to? Could they be reliably captured? Could they be of any help in the process of deterring identity fraud, investigating the normal crimes (as well as those associated with identity theft), establishing security and more demandingly could they be relied upon in a forensic context? Does using biometrics have ramifications which should be studied just like any initiative which involves people?

Those questions are there for all biometrics and will be dealt with in an organised fashion in this piece of work.

Voice, on the other hand, conveys a large amount of identity information and we know this from our own experience. Voice is the main modality of communications and while our main goal of speaking is not communicating our identity, we send out our 'identity information' as meta-data through our conversations.

While the intuitive belief is that voice is not as distinctive, reliable and permanent as biometric identifiers such as fingerprints and iris, numerous reasons have attracted attention to the investigation of the suitability of voice verification for security, crime reduction and crime investigation. The first reason is that, in some cases, speech samples, whether we like it or not, are the only sources of evidence that we have in relation to a crime. The examples include when a threatening message is sent to a victim or the voice of a criminal is recorded at the scene of a crime or even when it has been heard by an uninvolved bystander. The second reason is that there is little stigma associated with use of voice, mainly because it has not been used for crime investigation purposes in the past. But, most importantly, voice has a great advantage over other biometrics in a civil context and that is its collectibility. Voice is easily collectible through everywhere-available devices and could be transferred easily in several ways. This qualifies voice as a good choice for long distance identification.

Just as it is true for any biometric identifier, there are numerous scenarios in which voice could be used as a security measure and there are numerous dimensions on which the appropriateness of use of this identifier needs to be assessed.

The above facts guide us to the next topic presented here which is how this research assists us in the direction of analysis of voice as a biometric, in other words, what are the goals, scope and contributions of this research in this domain?

1.3 Introduction to this Research

This research represents an interdisciplinary approach to crime reduction. It is partly concerned with the development of robust voice recognition technologies and evaluation of their reliability and security. Such systems could be used to protect credit card and other commercial transactions. However, it also examines the criminological context for their deployment.

The scientific elements that are presented in depth in the thesis are firmly rooted in mathematics and computer science. Through the exploration of stochastic models to interpret human voice data and the development of software to develop secure products, the feasibility of voice verification is analysed. The research is targeted towards identifying how the technology might be used effectively to block off opportunities that are available to motivated offenders to gain unlawful access to property and resources and to steal identities. The applied criminology component in the research is in the understanding of how vulnerabilities of the system give rise to the creation of criminal opportunities and how best we can block these off from the motivated offender. Whilst computing, engineering and the physical sciences point the way to what it is possible to do technically, the study of non-technical dimensions provides crucial intelligence on where, when and how to target technology, and by and for whom, in order to maximise its impact on the detection, prevention and reduction of crime.

The innovation in this project is in the development and optimization of an accurate voice biometric system, evaluation of its security / reliability and in defining the interaction that needs to take place between the development of the technology and the human factors that maximise its effectiveness in preventing crime. These factors are relevance to the nature of the crime problem, acceptability, ease of use.

The research has explored contextual issues involved in a deployment of new technology, in particular, the ways in which security breaches may occur when voice-enabled security measures are in place, the opportunities granted to the offenders, the best approaches to the blocking off of fraud opportunities and the practical and ethical implications of technological solutions to crime.

On the application level this research has sought to:

- Identify and classify the scenarios in which voice could be used
- Explore the social sensitivities and ethical concerns about surveillance and the implications of these for the implementation of technical solutions
- Find all quantitative and qualitative measures affecting choice of a biometric identifier in those applications
- Figure out reliability of voice in those applications
- Develop methodologies and approaches for facilitating the design of biometric systems, the evaluation of performance and the reduction of opportunities for fraud

On the technical level this research has been aimed primarily at:

- Finding, proposing and optimising speaker specific features
- Improving speaker identification accuracy by the fusion of features
- Analysing the feasibility of speaker verification on smart cards in accordance with available standards
- Researching the effect of noise, channel, mobile communication and a variety of factors on speaker verification accuracy
- Analysing the problem of spoof (counterfeit biometric samples) which encompasses analysis of: possibility of spoof for voice, variation of spoof algorithms and methods to detect spoof

Although the intention is not to use voice surveillance for the identification and tracking of suspects and offenders or for monitoring the behaviour of the general public, there are some potentially sensitive issues which have been looked into in this research. The focus of this research will be on automatic methods in security applications.

While this section only aimed to present an introduction to research questions and goals, the exact scope of this research is defined in chapter 2, where topics in need of further investigation are elaborated upon. The next section offers a roadmap to the thesis.

1.4 Introduction to this Thesis

Following this short introduction, chapter 2 presents a roadmap to the thesis, which aims at providing a detailed description of questions in need of investigation, describing research approaches / methods and highlighting the contributions of this research.

Chapter 3 offers a comprehensive review of the literature. A survey of the principles of biometrics, non-technical issues (human concerns about misuse of biometrics as well as legal, ethical and privacy related issues surrounding biometrics) and technical issues (legacy of speech science for speech and speaker recognition) is presented in that chapter. Based on the review of literature and opportunities for the use of voice, a number of scenarios for the use of voice are suggested.

Chapter 4 addresses the most significant problem of the use of voice for security and crime investigation which is the lack of a comprehensive framework for evaluation of the reliability of voice-based verification algorithms under various conditions, especially when spoof attacks are brought into the picture. The characteristics of such a framework and the architecture for automatic evaluation in accordance with that framework are discussed there.

In chapter 5, the prototype voice verification system implemented for this research is described and it is shown that the system works suitably with a variety of parameters under normal conditions. System parameters are optimised and a number of questions about the nature of mathematical methods are answered.

Chapter 6 is devoted to the development of two spoof modules in the two categories of synthesis and conversion outlined in chapter 4.

Chapter 7 discusses the other side of the evaluation: reliability analysis. Several mismatch factors, such as distance and channel, coding, style and noise are analysed and in each case, some suggestions for improvement are made. In addition, the classification and categorization of conditions in which the verification system should be tested is undertaken. This is the basis for the evolution of the evaluation system on two dimensions of security (chapter 4 and 6) and reliability (chapter 7).

Chapter 8 tackles another research goal which is not adequately investigated in the literature: distinguishing between genuine and counterfeit / manipulated speech. Several features, methods and algorithms are proposed and tested in this chapter on spoof detection.

Finally, chapter 9 summarises the main propositions, ideas and findings discussed in the thesis and offers design 'blue-prints' that describe robust and reliable semi-supervised voice verification systems (sensitive and respectful to human concerns) that could be developed. Conclusions are drawn about the current and future status and problems of each group of scenarios presented previously, in light of the findings. The chapter is written with the aim of placing the findings in the broader context of the analysis of voice as a biometric for criminological and security applications.

Chapter 2 : Questions in need of further investigation: Research Hypotheses, data sets and methodologies

2.1 Purpose of This Chapter

The main purpose of this chapter is drawing a detailed roadmap for the rest of the thesis with the greater aim of putting each piece of the presented results and investigations into perspective. Due to the interdisciplinary nature of the research one runs a high risk of deviating from the main purpose of this study, which is a high level analysis of problems and issues surrounding use of voice verification in the context of criminology and security. The best approach to avoid this is to highlight the purpose, significance and the role of each portion of the research in filling the gaps in our knowledge about accuracy and security of speaker identification, and for each portion describes its contribution to answering the question of 'how reliable and useful voice verification could be in different contexts'. In other words, throughout this chapter we will try to:

- Classify different contexts and applications in which voice verification is used
- Summarise gaps in our existing knowledge about making use of voice verification in those contexts
- State the hypotheses that need to be tested and problems that have to be tackled
- Outline the new work and contribution of this research in each of those areas

A quick reference to available research on the subjects ends in this chapter and leads into the literature review (chapter 3), or later in the thesis as specified.

The rest of this chapter is organised as follows:

The second section describes different contexts in which voice as a biometric contributes to security enforcement, crime investigation or crime reduction. Within each context, the practical applications for voice verification are introduced. Then we will list existing questions, problems and gaps in our knowledge in 2.3. It can be seen from the list of problems that they fall into three major categories which are elaborated in greater detail with a reference to where these questions are dealt with in the thesis in sections 2.4 to 2.6. Section 2.7 elaborates on research methodology and datasets and section 2.8 summarises research contributions.

2.2 Overview of Categories and Contexts of Voice Verification

It is evident from the literature that voice verification, as an example of biometric recognition, has numerous applications in security enforcement (access control), crime investigation (post-incidence identification and content verification) and crime reduction by means of surveillance. Rose (2002) enumerates a number of those applications for forensic phonetics such as construction of voice line-ups, speaker profiling (determination of the regional or socioeconomic accent of the offender's voice(s)), content identification and tape authentication (determining whether a tape has been tampered with). Despite the variety of applications, the questions in need of further investigation are largely similar. For this thesis, it is necessary to build a more solid understanding of the categories of those applications. Afterwards we will be able to elicit and classify existing questions in each category.

In figure 2.1 the categories of voice verification's applications in the context of crime reduction, investigation and security are portrayed and summarised. These categories are:

1. Security

Voice can serve as a biometric to secure access to assets, services and information. Owing to widespread use of new channels of communication such as mobile and voice over the internet, this topic seems to be the main driver for the development of voice verification systems in the future. Voice verification serves as an additional security measure beside the current methods which examine the possession (e.g. card) and knowledge (e.g. PIN) of the claimant.

In the realm of security, and regardless of what is meant to be secured, we should deal with the problem of reliable content verification (what is being said) and reliable speaker verification for text-prompting voice-verification systems (which is a must to avoid voice replay attack). These two categories bring in several problems which are discussed in the next section.

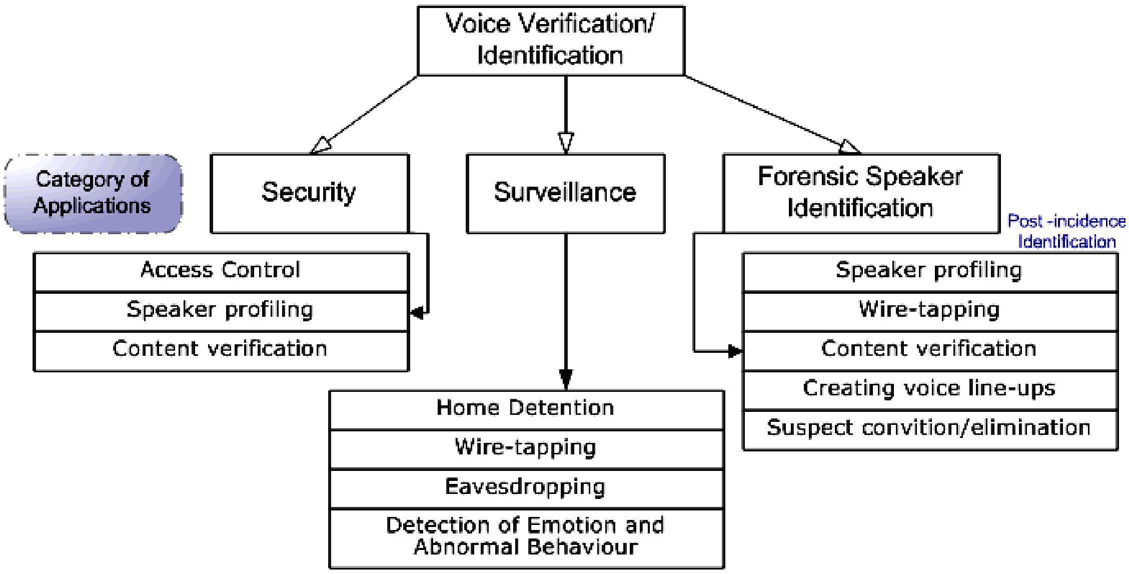


Figure 2-1 Contexts and categories in which voice as a biometric is useful

2. Surveillance

While surveillance could literally be defined as the process of close observation of a person or group, especially one under suspicion[1], it normally involves monitoring the behavior of people, objects or processes within systems for conformity to *expected or desired norms* in trusted systems, for *security* or *social control*[2].

One of the novel applications of voice verification is in the monitoring of curfewees. This type of supervision, as a rival to bracelet and tags, poses its own risks. A threat named ‘relay’ threat here is discussed in chapter 3 and 4 among the types of spoof attacks[3].

Eavesdropping falls in the category of content recognition posing the question of impact of ambience noises and distance to microphone on speech recognition.

Problems around use of voice for wire and telephone tapping are either technical (e.g. accuracy of speech recognition) or non-technical (surveillance ethics and privacy issues).

Finally voice could be used in public places to produce a notifying signal is case of violation of a desired norm. This would include detection of screaming, sound of gunfire or collision in public places. This application similarly demands accurate content recognition (where content here is not necessarily speech).

[1] The American Heritage Dictionary of the English Language, Fourth Edition (2000)

[2] “Surveillance”, Wikipedia.

[3] Spoof attack is a type of attack on biometric system in which fake biometric data is presented to the sensors (Nixon et al. 2008).

3. Forensic Speaker Identification

Forensics (or forensic science) aims at answering questions of interest to the legal system[1]. Generally it encompasses a broad range of sciences[2]. In particular, forensic speaker recognition involves legal and judiciary use of voice cues (e.g. voice-prints) to assign a recording obtained through anonymous calls, wire-tapping or direct listening to a group of suspects. The reliability of forensic voice verification is greatly disputed (see chapter 2 and appendix E). Due to the sensitivity of this application, little trade-off or compromise in this field is tolerable.

The questions before forensic speaker recognition are largely in common with those present in the area of voice-enabled access control. The differences are that:

1. Forensic speaker recognition normally involves post-incidence identification. Therefore some supplementary data is available to investigators such as transcription of the voice, conditions under which voice is recorded, history of the speakers, extra phonetic and linguistic data and hypotheses about the content of voice.
2. Forensic speaker recognition can be based on linguistic information about dialect, speech habits, etc.
3. Automatic modeling and scoring of similarity is less possible and desirable in forensic speaker identification. In cases where a score is provided, because of the many possible mismatch factors, a decision reliability estimation (normally based on Bayesian inference) should also be provided (Richiardi et al., 2006).
4. Reliability and confidence measures should be provided in addition to decision result. Unlike in access control applications, a 'pretty good' system is not utilizable.

Although my investigation about the effect of mismatched conditions, compensation methods, verification reliability and resistance to tampering would also be useful and interpreted in the realm of forensics, my main focus will be on the less controversial application of voice in access control.

[1] Forensics, Wikipedia. Similar definitions with the same theme exist in the literature for example in Wilson (2008).

[2] The American academy of forensic sciences includes the following specific areas of expertise: pathology and biology, toxicology, criminological, questioned documents, forensic odontology, anthropology, jurisprudence, psychiatry, and a general section (Eckert, 1997)

2.3 Overview of Questions Surrounding Voice Verification

A look over the contexts and applications described above reveals that the main questions faced by voice verification are:

1. Reliability of content verification

This is the question of speech recognition which is less dealt with in this research. Our interest in speech recognition is restricted to whether converted or synthesised voice could still deceive an automatic speaker recognition system with both speaker and content verification modules or content verification is successful in blocking these types of attacks to some extent.

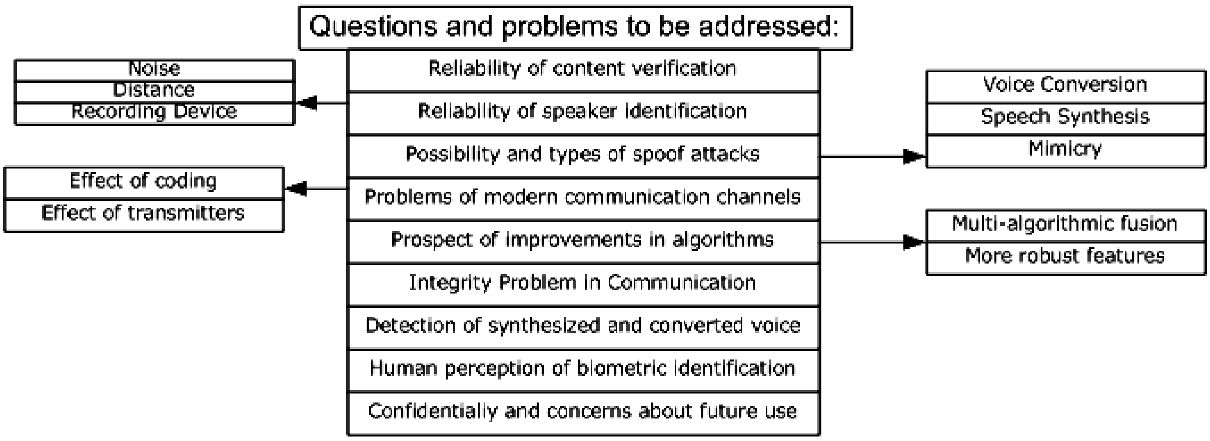

Questions and problems to be addressed:
Reliability of content verification
Reliability of speaker identification
Possibility and types of spoof attacks
Problems of modern communication channels
Prospect of improvements in algorithms
Integrity Problem in Communication
Detection of synthesized and converted voice
Human perception of biometric identification
Confidentially and concerns about future use
Noise
Distance
Recording Device
Effect of coding
Effect of transmitters
Voice Conversion
Speech Synthesis
Mimicry
Multi-algorithmic fusion
More robust features

Figure 2-2 Question and problems extracted from the study of various contexts and applications of voice verification

2. Reliability of speaker verification/identification

The major questions in this area concern the impact of mismatched conditions (between enrolment and test sessions, or between two circumstances in which voices are collected) on the accuracy of voice verification. Several factors account for the variation in the accuracy of speaker recognition such as: background noises, microphone mismatch, channel distortion (mobile, telephone, coding and compression effect), speaker's age, style of speaking, distance to recording device, room's acoustics, emotional state, health state and temporary effects (such as eating, drinking, etc.).

The goal of this research is not to repeat the experiments carried out in previous studies. Whenever published results of earlier investigations provide adequate ground for determining the extent of reliability in the examined conditions these have been mentioned and form the basis of the discussion on these matters. It is notable however since the datasets, conditions of experiments and evaluation scenarios are quite varied in the literature this work provides a higher level of comparability, since it examines various factors in one study and in the same set-up.

3. Possibility of voice forgery and types of spoof attacks

Different categories of spoof attacks are identified and classified in the thesis. In addition the need for a standard scheme for evaluation of the security of voice verification systems especially for access control is highlighted in Chapter 4.

4. Problems of modern communication channels

In telephony, mobile and over-the-internet transmission of voice spectral characteristics of speech are distorted. This phenomenon has severe impact on the reliability of voice verification for both forensic identification and commercial access control. The problem falls into the category of the reliability of voice verification and is described in greater detail in the next section.

5. Prospects of improvement in algorithms

This is the main research question dealt with throughout this thesis. Having a clear idea about the present and the prospect for the future improvement in accuracy, reliability and security of speaker and speech recognition helps us to place voice in the spectrum of biometrics. In chapters 6 and 7 this question is divided into more specific ones concerning security and accuracy of speaker identification.

6. Problem of Communication Integrity

Integrity, along with confidentiality and availability are three pillars of security (known as CIA). Communication integrity ensures us that data has not been manipulated inappropriately, whether accidentally or deliberately in the process of traveling from source to destination. There is a whole range of standards for biometric record exchange, and secure transfer protocols (such as secure socket layer, SSL) which guarantee integrity in communication (see Appendix F on security concepts). Two other related topics in this area are watermarking for upholding authenticity and steganography for communicating a hidden message. Watermarking is the process of embedding information into a digital signal for example voice or image, which is normally undetectable by a human. This can help us verify authenticity of a signal later. Faundez-Zanuy et al. (2006) suggested a watermarking technique for improving voice verification security. Watermarking cannot be helpful in fighting various types of spoof attacks such as synthesis and conversion. Steganography is an application of watermarking, in which a secret message is communicated through embedded information[1].

Since the communication integrity is a less significant security problem in case of voice verification compared to spoof attacks and it is largely studied before it is assumed that communication integrity (regardless of the contents of message which in our case is speech) can be guaranteed in the process of remote data collection, and consequently, problems in this area will not be covered specifically in this Thesis.

7. Detection of synthesised and converted voice

Following the recognition of speaker verification vulnerability to speech synthesis and voice conversion, the immediate question that springs to mind is whether the manipulation of the signal can be detected or not. This is in the end an impossible question to answer, since advances in signal processing can benefit both sides of this 'arms race', and there will be no particular point in time that we may be able to declare a winner. Nevertheless we will be able to assess the strength of each side over the other in a specific time. This involves having adequate knowledge about a wide range of voice forgery techniques and a framework for evaluation of security. We will formulate this need and discuss the requirements of such a framework in Chapter 4.

8. Human perception of biometric identification and human factors

Since the final users of biometric verification systems are people, their perception of the new verification methods and their convenience using the system are crucial factors in the success of the whole initiative. Human factors including right to privacy, confidentiality and concerns about future uses are discussed in Section 2.6.

A look over these 8 topics extracted from the application of voice verification shows that they can be dealt with in three groups, concerning reliability, security and human dimension of voice verification which are detailed in three separate sections below.

2.4 Problems around Reliability of Voice Verification

In this section different factors affecting the reliability of voice verification are discussed, previous work on the effect of those factors is reviewed and the role of this thesis in filling the gaps in our knowledge in these areas is outlined. Throughout the research if adequate study has

[1] Uludag (2006) in his thesis has addressed several problems around the security of biometrics, including steganography and watermarking for finger-print templates.

been carried out in the specified field, we will just mention the results, and will base our analysis on those studies.

Important factors to be studied with regard to reliability of voice verification are:

1. Noise

Several types of noises may degrade the results of speaker recognition. Having some knowledge about the nature of signal and noise can help us make use of fixed or adaptive noise cancellation filters such as Wiener (1949) filters for stationary and Kalman (1960) filters for non-stationary noises. These methods however rely on some hypotheses about the model of signal and noise or a source of noise for adaptation of the noise cancellation models.

In the context of speech processing the destructive effect of noises has long been known. This topic will be discussed in detail in chapter 7 and Appendix I.

The main aims of this research in this area are classification of types of noises, classification of noise compensation methods and identification of their underlying assumptions / requirements. Chapter 7 is entirely devoted to this mission.

2. Channel and Mismatched Recording

Channel mismatch is one of the most distinctive sources of performance degradation in voice verification. A channel normally consists of air, the recording device and the communications channel proper. Each of these elements distorts the speech signal and attenuates or amplifies spectral bands selectively.

There is a fair amount of research dedicated to channel mismatch and channel compensation in speech processing which has been closely examined in this work (Chapter 7 and appendices). A table consisting of procedures used to reduce channel and noise effect is compiled in chapter 7.

A close investigation of literature in this area leads to identification of following questions and topics in need of further investigation:

- Range of expected accuracy of voice verification in mismatched conditions with and without compensation needs to be determined.
- Type of suitable compensation methods in different conditions and extra data needed for compensation needs to be specified in different conditions. Such data may include knowledge about the handset parameters or some previous data recorded using the same microphone.

- Analysis of the availability of those types of data in each context is necessary in different cases.
- Effect of mismatched recording on confidence measures especially in forensic speaker recognition needs to be determined.

The specific contributions of this research are as follows:

- Giving a summary of previous research in this area including a comprehensive account of proposed compensation methods
- Conducting re-recording experiments for analysis of the effect of distance and channel combined
- Development of a new multi-algorithmic fusion technique based on fusion of scores from features with different spectral focus

3. Mobile and Landline

Distortion in speech characteristics caused by telephone or mobile line is one of the examples of channel distortion. However because of its importance a separate discussion is devoted to it here

Works by Moreni and Stern (1994) on analysis of the sources of degradation in telephone network, Lamel and Gauvain (2000) on the telephone speaker recognition, Kunzel (2001) on the effect of mobile and telephony transmission on the measurement of the vowel formants are among the many studies conducted on the effect of mobile / phone channel effect. In an interesting work, Byrne and Foulkes (2004) divided the effects observed in the telephony speech transmission into three types: environmental effects (noises), speaker effect (change in speaker's register and speaking style: 'telephone voice') and technical effects (effect of channel and selective transmission of frequencies). They examined both mobile and landline recorded speech and, based on the calculated formants, concluded that mobile phones cause considerable distortion in the acoustic properties of speech (place of formants). They confirmed results of the experiments carried out earlier by Künzel (2001) showing that the effects of landline and mobile phones are quite different.

The effect of channel is not limited to physical distortion. New transmission protocols used in mobile and internet communication distort the speech signal at the expense of using the bandwidth optimally. Guillemin and Watson studied the effect of Adaptive Multi-Rate (AMR) coding[1] on the measurement of formants (2006). They concluded that 'the GSM AMR codec used in these networks can in some cases have a major, and often unpredictable, impact upon the measurement of formant frequencies'.

In this research the effect of speaker's register is measured through the comparison of verification results for different speaking styles (chapter 7). On the other hand the accuracy of voice verification on AMR coded speech in different bandwidths which is missing in the literature is assessed and suggestions are made for bring the coding effect under control (chapter 7).

4. Distance

The accuracy of voice verification is a decreasing function of the distance to the recording device. Possible solution to this type of degradation is not fully investigated in the literature.

The theory of sound absorption suggests that sound attenuation in the fluids is dependent on the frequency of the sound and the amount of attenuation is not identical for all the spectral bands[2] (Kinser et al., 2000).

The attenuation of speech signal in the air has a side-effect too. As speech wave attenuates in the air by distance, the signal-to-noise ratio decreases and it becomes more likely for a speaker to be misidentified based on his noise-smeared and distorted recording. Saastamoinen et al. studied the effect of many factors such as sampling rate, distance to microphone, additive noises, etc. (2005) on speaker recognition. In their work however they used vector quantization (and not Gaussian mixtures as used in this work) and did not quantify the error rates as a function of distance. The effect of attenuation, signal to noise ratio, and channel (speaker, microphone, and media) are studied through the experiments reported in chapter 7 in this thesis. Having an accurate picture of the effect of these factors is extremely crucial for offering conclusion about the suitability of voice based surveillance, where distance to recording device is a deciding factor.

5. Speaking Style

[1] which is standardised algorithm for the Global System Mobile Communication (GSM).

[2] As elaborated on in Kinser et al. (2000) according to the first classical theories proposed for explanation and prediction of sound dispersion in the air by Stokes and Kirchhoff two types of losses account for dispersion of sound waves: viscous losses and heat conduction. The empirical results by Sivian in 1947 however showed that the classical theory and these two sources of loss do not accurately predict the attenuation in air and most liquids. A third category of attenuation known as molecular thermal relaxation could account for the rest of observed losses.

A detailed review of works on speaking style is offered in chapter 7. Previous research reported in that chapter, along with the experiments carried out on various speaking styles, presents a clear picture of the effect of mismatch of the style on verification results.

2.5 Security of Voice Verification: an Urge for a Standard Evaluation Scheme

It is demonstrated in chapter 4 that despite various experiments on the vulnerability of speaker verification to speech manipulation techniques, the available works do not address the problem of the security of voice verification, nor provide a solid ground for development of robust verification systems.

Works by Lindberg et al. (1999), Masuko et al. (1999), Masuko et al. (2000), Pellom, et al. (1999), Lau et al. (2004) and many similar works detailed in chapter 4 show that mimicry and spoof are significant threat to the security of voice verification system.

While those studies offer some insight into the security problems, it could be argued (as detailed in chapter 4) that these types of study suffer from a number of drawbacks:

1. It is not possible to ascertain how the results of previous studies apply to a new voice verification system, which necessarily does not use the same algorithm utilised in one of those studies

2. Neither the spoof attack, nor the defence algorithms are implemented as re-usable blocks for future experiments.

3. The studies do not provide an evaluation framework for analysis of the success of spoof detection or any other defence mechanisms.

Available biometric standards are briefly introduced in Chapter 4. Some of those standards are useful for detection of vulnerable points in a biometric system for example Common Criteria's Biometric Evaluation Methodology (BEM) identifies 15 points of vulnerability and categories of threat in biometric verification systems. Many of these threats are not specific to biometric systems and would still be there if, instead of biometric records, a password was used for authentication. Likewise, ISO/IEC JTC-1 SC-37, which is responsible for the international standardisation of biometrics, has published 16 biometric related standards up to April 2007. Among these standards, four of them relate to application programming interfaces (BioAPIs), two of them specify the framework and methods of conformance tests to BioAPI, two of them concern biometric testing and reporting criteria including the evaluation types and error rates, and eight others define biometric data exchange format for fingerprint, face image, iris image and vascular image data.

Neither of the aforementioned standards addresses the problem of strength of the function of biometrics against spoof attacks in the way outlined in this work.

The need for a flexible framework which realizes at least two objectives is highlighted in Chapter 4. These two objectives are: assessment of the vulnerability of the system and assessment of the success of fighting back mechanisms.

There are two key attributes expected from such a framework: comprehensiveness and extensibility.

1. Comprehensiveness: A comprehensive study of the strength of voice verification systems against spoof attacks, stipulates identification of all types of attacks. In the context of security, leaving one back-door unattended is as dangerous as leaving the entire house unprotected. Therefore thoroughness should be a key attribute of such a study. Nonetheless it is not evidently possible to test all types of vulnerabilities in this research but the framework and how the results are presented are clearly documented in this work. One type of attacks in each category of conversion and synthesis are proposed and studied in chapter 5.

A real evaluation scenario should embrace all types of attacks in a well defined framework.

2. Extensibility: Another significant point made earlier is that due to the rapid changes in the technology and algorithms contributing to the voice forgery side, there is no absolute guarantee of security for any verification system. However, the system should be easily augmented and expanded with new blocks of counterfeit detection.

Furthermore identification of a new type of attack should become legacy knowledge in the domain of security. In other words, spoof techniques should be simulated as independent blocks which are usable without any dependency on the verification system.

These requirements are expanded on in Chapter 4 along with the study of attacks, but in anticipation of this discussion the characteristics of an evaluation system should at least encompass following items:

1. It has to be independent of voice verification system to minimise the time required for adaptation to the new system
2. It should be capable of accommodating new threats, especially spoof attacks (independence of attack simulation layer from voice verification system)

3. It should allow changes in the design of framework and introduction of new metrics for automated tests which translates to independence of voice verification system from evaluation system (a controller layer)
4. The success of algorithms for detection and prevention of each type of attack should be easily testable and the system should allow effort-free addition of modules implementing those algorithms

In parallel to the presentation of the framework and the investigation of types of attacks chapter 6 will develop mathematical grounds for one type of voice forgery in each category of automated spoofing. The answer to these questions is sought through simulation of the attacks:

1. How vulnerable is voice verification to automated spoof?
2. Does reliability translate into security? Is an accurate enough system secure enough?
3. How much knowledge of system parameters and algorithms is necessary for conducting a successful spoof attack?

The goal of determining the success rate of spoof detection by inconsistency and discontinuity detection is pursued in chapter 8.

2.6 Human Factors and System Design Considerations

Biometric-based surveillance and control systems are different from other forms of security enforcement systems in many subtle ways. They closely examine a person's body (which raises the concerns about violation of the integrity of human body), they have capacity to exploit body information in numerous ways (which was not expected at the time of data collection) and they hold unique information about the users which if stolen could not be replaced easily or at all (unlike a password or a key that can be regenerated).

Biometrics expands along several human dimensions and raises questions and concerns in the following domains:

1. Right to privacy

In many cases the validity of the question about the invasion of privacy as a result of the introduction of a security or surveillance measure is extremely tough to evaluate. Our judgment is constructed by the legal and social values of the society and we own a 'juridified intuitions that reflect our knowledge of and commitment to, the basic legal values of our culture'

(Whitman, 2004). The right to privacy is also closely related to right to anonymity. Biometric identification can infringe the right of anonymity and pseudo-anonymity[1].

There is a great deal of philosophical and practical discussion in the literature on the extent of right to privacy and right to be anonymous. This is not a subject that we can conclude about and this is not the main focus of this research. Nevertheless as an indispensable part of a biometric related research it is imperative to summarise the current viewpoints on this subject and the legal status of the protection of right to privacy. With regard to voice verification, these view points and laws translate to limitations on speaker recognition and surveillance initiatives and affect the system design (concerning the ownership and access to voice data). These are discussed in Chapters 3 and 9 in greater detail.

2. New concerns raised by biometric recognition

As mentioned before biometric systems raise new concerns which are partly investigated in previous works. These concerns and the extent to which they apply to voice verification are summarised in chapter 3 and the synthesis chapter (9). A blue-print for biometric verification systems is proposed for reducing some of these concerns in chapter 9.

3. Human perception of biometrics

Many surveys have been conducted on the public perception of the necessity and advantages of introduction of biometric identification for various applications from which a few good examples are Furnell and Evangelatos (2007), Westin (2002) and Biometric Identification Technology Ethics project survey of 2006.

Current surveys include questions about participants' perception of the reliability of different biometric systems (e.g. Furnell and Evangelatos (2007)). These questions are not enough to determine the public's perception of biometrics and their fears when real biometric systems are deployed. In addition to those questions new public opinion surveys are required to investigate the satisfaction and fears of 'informed users'. That is because in light of security concerns, such as those investigated for speaker recognition in this thesis, new questions about perception of biometric identification and human concerns after gaining knowledge of security risks are shaped. Any party undertaking biometric verification should seek to address those concerns within the community for which biometrics is going to be used.

[1] Use of a fictitious distinguishing mark with the desire to remain anonymous (Goemans, 2001).

It chapter 3 the results of a few surveys are presented and compared to determine the extent of concerns and that how different is voice verification from other biometric recognition methods. The literature shows that public opinion tends to shift in time, and that it favours use of biometrics in applications which are considered important. Major questions here concern comparison of voice with other biometrics, applications in which public is more willing to use voice verification, public acceptance of voice verification before and after knowing about threats and finally new fears introduced by likelihood of identity theft and biometric spoof.

4. Shift in the type of the fear of identity theft and concerns about vulnerabilities

Abovementioned surveys already show that current security measures are not completely satisfactory for the users. Nevertheless, it seems that deployment of biometric systems introduces new concerns (as described earlier) and creates a generation of habits directed towards safeguarding biometric information. In a biometric observed society people should protect their biometric information such as fingerprints, voice, iris-scan and other clues that could be used to forge a biometric template. This new type of fear could be considered as a severe disadvantage for biometric security and surveillance

Assuming that having phonetically rich enough samples of someone's speech allows intruders to produce any desired sentence with the person's voice, leads us to feel more unprotected than before with voice enabled security measures and more careful when talking to strangers over the phone or in public places. This is true about and applies to some extent to all biometrics with the possibility of counterfeit.

5. Issues around collection and use of biometric data

Due to the two categories of concerns: function creep and sensitivity of biometric data, the collection and use of biometric data remains a controversial issue.

A blue-print is presented in chapter 9 which is extremely helpful in regulating the use of biometric data by public and private sectors and controlling the access to biometric data through specifying certified channels and services.

6. System design and policy recommendation

The outcome of various studies in the security and access control domain culminates in development of an identity management system which comprises many modules including biometric authentication modules. Issues around use of identity management systems are introduced in chapter 3 and Appendix C. In the Synthesis Chapter the lessons learnt from all the previous studies are applied to the problem of design of a biometric-enabled scheme. A blueprint for collaboration of service-providers and (biometric) identity providers with emphasis on separation of access permits and certification of biometric services is presented in chapter 9.

2.7 Research Methodology and Datasets

This section aims at description of research methodology, datasets and phases.

The purpose of this research, analysis of suitability and limitations of voice as a biometric in the context of security and criminology stipulates two types of research: exploratory (identification of applications, new problems and current solutions) and constructive (development of a system for assessment of hypotheses, suggestion of methods for tackling the problems and re-evaluation of success of those methods).

The technical investigations in the research follow the model of empirical and scientific research. Scientific research undertaken is based on employment of mathematical models, available datasets of speech, simulation of hypothesised conditions and generalisation over the results of experiments.

In contrast, human research portion of this study is built upon secondary data including surveys, case studies and questionnaires which follows classification and interpretation in the field of concern.

Although there is room for field research under this title e.g. examination of human perception of voice verification and concerns about biometric theft this work makes use of secondary data in those fields. As described in chapter 9 this research identifies the topics in relation to which human perception should be gauged and lays a foundation for conducting surveys prior to deployment of the verification system.

The results of empirical investigations are also used for predication of security volition scenarios, identification of future human concerns and providing suggestion for system design.

The majority of technical experiments are carried out on the speech data available in two corpora of IVIE and CHAIN. These two corpora are introduced in detail in chapter 5.

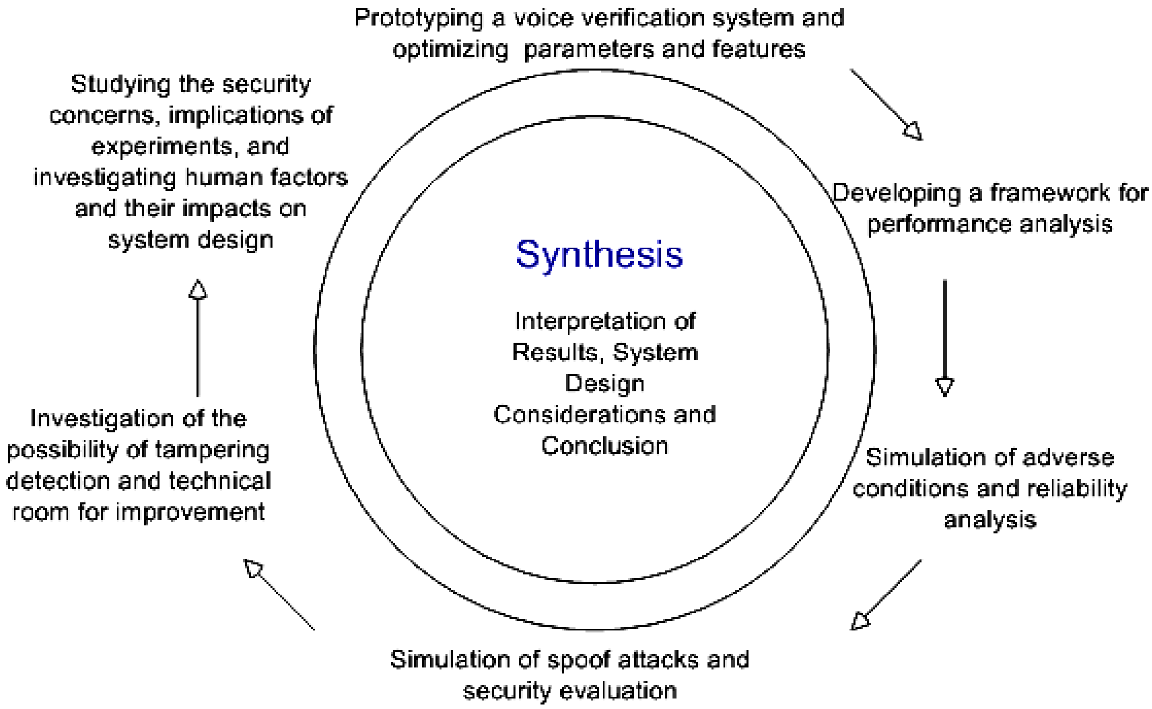

Prototyping a voice verification system and optimizing parameters and features
Developing a framework for performance analysis
Simulation of adverse conditions and reliability analysis
Simulation of spoof attacks and security evaluation
Investigation of the possibility of tampering detection and technical room for improvement
Studying the security concerns, implications of experiments, and investigating human factors and their impacts on system design
Synthesis
Interpretation of Results, System Design Considerations and Conclusion

Figure 2-3 Research phases and activities

Figure 2.3 depicts the research stages and actions taken in each stage. These stages are briefly described here:

1. Prototyping a voice verification system and optimizing its parameters and features

This phase involves development of a voice verification system and optimizing several system parameters. This is undertaken in chapter 5.

2. Developing a framework for performance analysis

This step lays the foundation for the rest of the research which allows us to carry out the security and reliability related experiments. It is however notable that the framework offers more than just a foundation for conducting experiment in different conditions. It presents a design which provides agility in coping with the attacks and extensibility to accommodate further threats menacing a voice verification system. In Chapter 4 spoof scenarios are investigated and the requirements for a security evaluation framework are identified. The framework constructs a significant part of the contribution of this research and the other chapters could be seen as enhancements of this framework.

3. Simulation of adverse conditions and reliability analysis

This phase includes study of the impact of noises, AMR coding, speaking styles, and distance to recording device on the accuracy of voice verification. Chapter 6 reports on the results of this phase.

4. Simulation of spoof attacks and security evaluation

Two types of attacks one in the category of synthesis and the other in the category of voice conversion are simulated and tested in chapter 6. Various experimental investigations about spoof attacks are undertaken with the aim of providing answers to these questions:

- Does difference in parameters of the system and those hypothesised by intruders render an attack unsuccessful? In other words, should the impostors have detailed knowledge about the system to conduct a successful attack?
- How much data (data about the system parameters, samples of speech from target speaker, and models parameters of the target speaker) should be at the intruder's disposal to enable them to carry out the attacks?
- Does system precision guarantee protection against attacks? In other words what is the relation between the reliability and performance of the system, and its vulnerability to spoof attacks?
- How is the magnitude of errors due to spoof attacks comparable to the errors imposed by mismatch conditions?

The results of experiments carried out in this phase are reported in chapter 7.

5. Investigation of the possibility of spoof detection and technical room for improvement

This research explores the prospects of improvement in speaker verification's accuracy by means of various methods especially the proposed multi-algorithmic fusion technique (chapter 7).

On the other hand, in chapter 8, the results of experiments carried out on the possibility of spoof detection is reported. The results are based on the analysis of the signals gathered from automatic speech synthesis systems in addition to the ones generated by the spoofing modules developed in this work. The range of spoof detection success by applying different mathematical methods is reported in chapter 8.

Chapter 7 and 8 allow us conclude about the prospect of improvement in the reliability and security of voice verification.

6. Studying the security concerns, implications of experiments, and investigating human factors and their impacts on system design

The ninth chapter entitled ‘synthesis’ is written with the aim of summing up findings about security issues, human factors, privacy issues and their impact on the system design as summarised above. It provides a system level interpretation of the experiments on the security and reliability of voice verification.

2.8 Body of Research Covered in Literature Review

The body of research reviewed in chapter three is very much an expansion of the topics covered in this chapter and is inline with the research goals and phases described so far.

An introduction to biometric recognition (including methods, terms and principles), identity fraud as motivation for use of biometrics, privacy / legal / human related dimensions of deployment of biometric systems, available solutions to human concerns, summary of previous research in the area of speech science and previous research on forensic speaker recognition is presented in chapter 3. I have organised the applications of voice for ‘crime reduction’ (including security applications) and ‘crime investigation’ (including forensic applications) in 4 categories and 9 groups of scenarios in chapter 3.

2.9 Summary of Research Contribution and Next Chapters

Table 2-1 summarises the contributions of this research and specifies where they are presented in the thesis.

This chapter introduced the problems, gaps and weaknesses in our knowledge about voice verification mechanism and posed a number of research questions in different dimensions in need of close investigation. The next chapter reviews the previous work on each topic in greater detail and lays the ground for introducing a framework that will be used for the empirical research component of this Thesis. The experimental research carried out in the next chapters followed by the interpretations and discussions help us determine the shortcomings and strengths of voice as a security biometric in its potential applications and locate voice within the range of biometric identifiers used for security enforcement and surveillance.

Table 2-1 Summary of research contributions

Chapter	Contribution Made or Problem Addressed
3	Classification of voice verification contexts, scenarios and categories
3, 6	Classification of human factors and identification of two new human concerns in light of spoof possibility
4	Classification of types of speech spoof
4	Identifying the requirements and proposing a security evaluation framework
5	Suggestion of distance measure for two utterances which is used in cluster analysis and Dendrograms
5	Analysis of the clusters in speech feature space
6	Formulating and simulating a synthesis attack based on Hidden Markov Models
6	Proposing and simulating a new voice conversion attack based on Gaussian Mixture Models
6	Conducting spoof simulation experiments based on the suggestions of the evaluation framework through use of the abovementioned spoof modules
7	Analysis of the effect of attenuation in air and distance to microphone on reliability of speaker verification
7	Carrying out experiments on the effect of AMR coding and mobile channel distortion in different bit rates on voice verification
7	Isolation and analysis of the effect of style of speaking on speaker verification
7	Comparing and making contribution to the studies on the selection of most speaker specific spectral components of speech
7	Suggestion of a new feature discrimination analysis measure in addition to F-Ratio and Mutual Entropy
7	Proposing and testing a new successful multi-algorithmic fusion technique for speech in various set-ups
7	Augmenting the evaluation framework by specifying adverse conditions for which evaluation should be carried out
8	Analysing the prospects and possibility of detection of discontinuity in synthesised voice for spoof detection
8	Proposing novel methods for discontinuity and inconsistency detection in speech
8	Proposing wavelet features for spoof detection along with the analysis of suitability of a large pool of features for this task
9	Proposition of a voice signature scheme
9	Analysis of the issues around use of speech in all categories especially for surveillance
9	Proposition of a blueprint for collaboration of service-providers and (biometric) identity providers with emphasis on separation of access permits and certification of biometric services
9	Proposition of a system design based on a semi-supervised approach towards voice verification which enables taking advantage of spoof detection blocks

Chapter 3 : Literature Review

3.1 Goal and Organisation of This Chapter

Analysis of voice print and other biometrics for criminological and security applications is a multidimensional undertaking. This chapter reviews the literature of voice as a biometric from different perspectives: roughly termed non-technical and technical.

The goal here is to identify and elaborate on the specific problems and research questions which have been partly presented in the previous chapter with more emphasis on non-technical issues.

The ideas are presented both in the chapter's text and the accompanying appendices. The appendices introduce the complexity to each problem, by comprehensive collection and presentation of thoughts and reflections on each topic.

In the interest of unifying the terminology in the thesis, and explanation of the required concepts, a short introduction to biometrics is presented in 3.2 and Appendix A (section 1). The emphasis here is on fusion and spoofing. A major portion of the contribution of this thesis relates to analysis of voice-spoofing and its detection. I have tried to briefly show here that spoofing is not only a threat to voice-verification. Biometric fusion has a technical dimension and a non-technical one which are investigated in this chapter.

I have tried to show in 3.3 that identity fraud justifies the use of voice verification in many applications in which security measures reduce the 'opportunity for fraud'.

The discussions in 3.4 culminate in the production of a list of all the concerns about the use of biometrics that have been expressed irrespective of their validity. It is shown in chapter 9 that voice is one of the best biometrics in view of non-technical concerns and with the assistance of integration of certain ideas into system design within the current body of regulations the majority of the concerns will be alleviated.

In 3.5 the legacy of speech science is analysed. Following a historical and analytical discussion about the forensic speaker verification and highlighting the current problems, I have organised the applications of voice for 'crime reduction' (including security applications) and 'crime investigation' (including forensic applications) in 4 categories and 9 group of scenarios.

A broad summary of how the next chapters contribute to the development of the argument and answering the research questions will be presented at the end of this chapter.

3.2 Overview of Biometrics and Performance Measures

3.2.1 Biometric Principles

This section provides a quick review of biometric concepts while the detailed description of relevant biometric principles is presented in Appendix A.

For the purpose of this analysis, it should be pointed out that while there are already several "biometric identifiers" such as DNA, Ear, Face, Facial Thermogram, Fingerprint, Gait, Hand-Geometry, Hand-Vein, Iris, Keystroke, Odor, Retina, Signature and Voice available the list is always growing with identifiers that their verification may seem to enter our personal space.

Biometric systems can be implemented with or without people support, can be overt or covert can involve direct contact with an individual or operated from a distance and can be open/closed[1].

Biometrics can also be used in commercial/civil, governmental or forensic applications.

Biometric Identifiers are usually categorised based on a number of factors such as universality, distinctiveness, permanence, collectability, performance, acceptability and the possibility of circumvention. Biometric recognition could be of either types of verification (one to one) or identification (one to many).

Type I and Type II statistical errors are usually reported in performance evaluation. If the null hypothesis is that the biometric record belongs to the claiming user, Type I or the error of rejecting the null hypothesis while it is true is the same as false rejection error. Likewise type II error (failing to reject the null hypothesis when it is not true) is the same as false acceptance error. These errors vary as a result of changing the system threshold. The errors are plotted in a variety of diagrams such as false acceptance rate/ false rejection rate (FAR/FRR) plot, receiver operating characteristic (ROC) curve and detection error tradeoff (DET) plot on logarithmic, semi logarithmic or linear scales.

The classification of applications and the systems presented above is somehow rough. A more discriminating and helpful classification in the context of voice verification (which allows us

[1] For more details and the references see Appendix A.

analyse the problems and shortcomings) will be offered in this chapter and scenarios of use of voice as a biometric will be analysed in chapter 9.

3.2.2 Biometric Fusion

A review of biometric fusion is necessary for conducting further discussions about the technical and non-technical dimension of the use of biometrics, especially voice. In section 5 of Appendix A the literature on biometric fusion techniques is reviewed in greater detail.

Multi-biometric systems can be of any type of multi-sensor, multi-algorithm, multi-instance, multi-sample, multi-modal and hybrid (Ross et al., 2008)[1]. The fusion could occur at sensor level, feature level, score level and decision level.

The most important points in the appendix should be re-iterated here. First, biometric fusion allows us to eliminate some of the technical and non-technical obstacles before large-scale usage of biometrics. The non-technical problems such as age and disability are of great importance since they immediately affect the perception of biometrics and alleviate some ethical concerns (discussed later).

The fusion techniques are various. Fusion by sum/product of raw/normalised scores, linear weighted sum, product of FARs, min/max of FARs, product of likelihood ratios, logistic regression, AND/OR decision fusion, discriminant classification fusion, product fusion score (PFS) and Biometric Gain against Impostors (BGI) fusion are among the ones mentioned in Appendix A (section A.5).

In chapter 7 (and appendices) simple product rule (or sum rule in log domain) will be used as a bench-mark for fusion. A variety of new fusion techniques as well as the well known product fusion score (PFS) which is the same as BGI-based fusion will be used and compared.

3.2.3 Biometrics and Spoofing

A spoof is a counterfeit speech sample designed to imitate a legitimate biometric submission (IBG, 2006) and Spoof attack is a type of attack on a biometric system in which fake biometric data is presented to the sensors (Nixon et al. 2008). Section 6 of Appendix A, offers a detailed account of how the spoof attack is a significant threat to the effectiveness of biometrics and this is invariantly true about the more reliable biometrics such as iris and fingerprints

[1] See A.5 for details.

It is mentioned in Appendix A (section A.6) that Schuckers (2002) in a review of spoof and countermeasures for various biometrics especially fingerprint explained that the possibility of stealth and making a copy of a key would not discredit the use of keys. Schuckers suggested other means of reducing spoof risks such as supervising the verification/identification process, enrolling several biometric samples (e.g. several fingers), multi-modal biometrics, and live-ness detection (deciding if the subject of identification is authentic and not e.g. a counterfeit object).

For voice verification she cited work by Broun et al. (2002) in which they incorporated a person's lip characteristics into the speaker verification task. A reference to Broun et al. (2002) and similar studies is made in Appendix A (section 6).

Despite Schuckers (2002) suggestion, adding modalities does not easily solve the spoof problem especially in the case of voice. It is evident that use of other modalities adds several dimensions to the study of voice verification for security purposes. On one hand use of other cues eliminates some of the advantages of speech-based authentication such as the low bandwidth necessary for transformation of speech data, simple and ubiquitously-available devices for speech collection and non-intrusiveness of speech collection. On the other hand, the same robustness and vulnerability analyses should be carried out for any new modality. While the aforementioned studies on the combination of lip-reading and face recognition with voice verification have not been overlooked, they are beyond the remit of this Thesis.

Spoofing is one of the main themes of this study and is extensively analysed and treated in chapter 4, 6 and 8 and the results of the studies carries a considerable weight of arguments offered in the final chapter.

3.3 Identity Fraud: Motivation for the Use of Biometrics

A biometric study is incomplete without an investigation of ‘for what reason’, ‘where (for what application)’ and ‘how’ biometric verification is needed. Identity fraud is repeatedly cited as the justification for the use of biometrics. While identity fraud is a broad topic the modest goal of this section (along with Appendix B) is to define identity theft/fraud, to give a picture of their magnitude and types of identity fraud, to define identity verification and, finally, to ascertain the characteristics of biometric identification which can reduce identity fraud.

Identity theft and fraud are defined in Appendix B and a break down of losses in different organizations due to identity fraud is presented in the appendix.

The most important points are worth re-emphasizing here. First as Jones and Levi describe (2000, p. 11): "The modern thinking on identity is that two separate equations should be satisfied. The first is to show that the individual actually exists. The second is to show that the applicant is or is not the individual they say they are". Biometric identifiers are only helpful in association of our physical body to the previously stored records. Identity fraud can arise from the loss/theft of physical identity documents, from their improper taking from existing official/commercial files, and from impersonation (Jones & Levi, 2000). Biometrics can not eliminate all of the possible scenarios for identity fraud especially when it comes to the misuse of data.

This section inevitably only scratches the surface of the numerous components of identity fraud. Nevertheless, through the information presented here and in Appendix B we can infer that:

1. There is huge room for the use of biometrics, especially voice, in financial sector. Some of the fraudulent activities carried out in these types of applications, for example, in the case of the 'card-not-present' transactions, stolen cards or mail-non-receipt category are mostly conducted by opportunists. Use of biometrics limits the opportunity for fraud/crime even if the biometric measures are not very strong.

2. Not withstanding this, in many applications of biometrics, in which authentication is ordinarily "in person" and "supervised" verifying the identity of the person is more crucial. The fraud (organised crime) in these applications is mainly carried out by resourceful criminals. Hence use of a weak biometric will be of no value. Besides, the crime may be carried out with the assistance of insiders such as officials or system administrators. This may happen despite the deployment of the strongest biometric identification system.

3. Usually in the second type of application, biometrics could not be of any help if the criminals make use of a real person with a clean record, e.g. for carrying out an act of terrorism.

4. In some of the above mentioned applications biometrics with the 'negative authentication' power are required. An example of such applications is iris-based deportation tracking system which prevents re-entry of deportees into the country after they had been expelled (such a system is already installed in Dubai and is analysed by Lazarick & Cambier, 2008).

Of further interest in this regard is whether or not voice verification can be used for applications calling for negative authentication.

While the information presented above sheds light on some aspects of identity fraud and the need for better identity evaluation methods, many other questions are left unanswered. We certainly need better ways for authentication from a distance. One question is where voice is located in this regard. What are the implications of use of voice for security especially in distant and unsupervised conditions? Is voice better than other biometrics, for example, due to greater acceptance by the public? Are we in the position to list all the scenarios in which voice as a biometric can combat identity fraud? Answering these questions is undertaken in the remaining parts of this chapter.

3.4 Privacy Issues and Human Concerns around Use of Biometrics

3.4.1 Investigating Human Dimension of Biometrics

In this section, non-technical aspects of biometrics will be scrutinised. Non-technical issues are divided into a several groups: human concerns, ethical concerns, issues related to privacy invasion and legal aspects. These titles are only chosen for logical division of the topics. There is a lot of overlap among all these areas and the main aim of this analysis is to identify the human concerns, why they are raised and how we could address them.

The text in this sub-section and its relevant appendix (Appendix C) is the result of critical analysis and extensive collection of concerns expressed in the previous research reports, opinion articles and surveys. The study has culminated in a table of non-technical biometric issues which is presented in chapter 9. Discussions have ensued in that chapter on the relevance of voice to each item. The approach adopted is not trying to refute or accept arguments in favour or against severity of an issue. Instead, here and especially in chapter 9, I have specified 'how' the voice-enabled security system should be designed to minimise the concerns. I have also tried to specify how the target population should express its opinion and raise awareness (for example concerning the risk of spoofing) in the light of discussions presented here.

The text is organised in this way. The chapter text in 3.4.2 only contains pointers to the concepts and discussions expanded on in Appendix C. While it is advisable to read Appendix C before this text, for a reader familiar with these concepts the appendix could be consulted only when needed.

3.4.2 Ethical, Legal and Privacy Related Aspects of Biometrics

Based on the thorough analysis of the literature in medical, ethical, privacy related and legal contexts and also based on the surveys of biometrics, I believe that the concerns expressed about the misuse of biometrics or harms inflicted as a result of their use could be placed into one or more of these categories:

- Concerns about stealing parts of body for carrying out a fraud (e.g. cut of a finger)
- Ethical aspects of fear reduction (without reducing its underlying cause)
- Concerns about the invasion of privacy and the undermining of human integrity
- Concerns about possibility of function creep in the use of biometrics
- Fear of biometric tracking and over pervasive surveillance
- Ethical concerns about value of human life and human dignity (branding argument)
- Concerns about biometrics being used beyond their primary purpose (e.g. to spy on people)
- Disclosure of sensitive information
- Fear about covert surveillance
- Privacy questions
- Depriving people of anonymity
- Possibility of permanent ID theft
- Religious concerns
- Fear of misuse of data
- Social stigmatization
- Costs (deployment, updating, maintenance as well as costs of fraud and wrong decisions)
- Direct medical concerns
- Indirect medical implications
- Power accumulation and weakening of democracy
- Law enforcement concerns including false interpretation of biometric evidence in courts
- Disability and age problems
- Impact on social interactions
- Fear of construction of full profile of actions from partial identities (similar to tracking)

Appendix C demonstrates in greater detail how this list was identified.

Apart from the concerns expressed in previous studies I tried to draw attention to two ethics-related aspects of the use of biometrics. First I pointed out that biometrics may give a false sense of security without offering the real security. In view of arguments by Rogerson and Christmann (2007) reduction of fear of crime without changing underlying dangers, which may be taken on by biometric developers, is unethical.

Secondly, communications and social interactions in a biometric diffused society are largely affected by the fear of biometric spoofing in the future. As we will see in the case of voice, having more voice samples from a person facilitates the generation of a counterfeit speech signal that resembles his/her actual voice. This may lead to us limiting our communication with strangers generating concern about the real intention of a person trying to contact us[1]. This may be exacerbated where the use of biometrics becomes pervasive (creating the opportunity for misuse of biometric data) and data collected through daily social interactions could be used for the purpose of spoofing (which causes deterioration in the quality of life and increase in the fear of crime).

Section 3 of Appendix C views biometrics from the privacy point of view. It could be noted that biometric authentication has some relevance to information and physical privacy.

After the review of literature on privacy, we reach a point at which we will have to decide whether or not we define and respect the “right to privacy” as an undeniable right of the person, which, even within a democracy can be lost, or we attribute a ‘limited value’ to this right in balancing privacy against other social goals (e.g. security).

A study of the legal aspects of biometrics (presented in C.4) proves that the second view has been taken by legislators. EU Directive 95/46/EC, UK Data protection Act (DPA) 1998 and US privacy act of 1974 are reviewed in Appendix C.4. While the details are of less importance to us, the information presented on legal and privacy issues demonstrates that:

1. It is evident that surveillance affects our behaviour and intrudes upon our private space (even if we are in the public space). No-one is unaware of the effect that monitoring can have on privacy in the psychological, sociological, economic and political dimensions described by Clarke (2006). Consent is the key when it comes to data collection about people. Nevertheless

[1] In C.2 the analogy with the email threats was given which shows how severity of threats affects the usage norms.

there are always ‘exceptions’ in the laws which justify the use of surveillance in the interest of the society[1].

2. Privacy has been treated as an “interest” rather than an “undeniable right” of which a person can not be deprived. Through the democratic processes the public’s interest can outweigh the person’s privacy interests or even rights. These democratic processes lead to (and have led to) different outcomes in different societies and cultures. Instead of global preferences, the surveys and polls should focus on the ‘target population’ and their preferences.

In Appendix C (C.6) issues of identity management systems (IDMs) are examined. Going over the issues presented in the related articles demonstrates that there is a huge fear about the misuse of IDMs. While the concerns are valid, it is arguable that biometrics only adds little to those concerns on two conditions: if used only for authentication and biometric data is not transferred in forms of authentication tokens or any form to service providers. On the other hand the identity providers should be certified and trustworthy. These requirements are laid out in chapter 9. A blueprint and conceptual framework in which these requirements could be addressed will be presented in that chapter.

3.4.3 Proposed Solutions to Human Concerns

Several suggestions for alleviating concerns about the use of biometrics and data collection have already been mentioned in the legal discussions. In fact, the requirements stipulated by laws and directives (such as those concerning notice, consent and disclosure) analysed hitherto are types of solutions to the presented problems. In addition to those, explicit suggestions are summarised here.

Clarke (2001) made a few recommendations as the safeguards for bringing the impact of biometrics under control. The recommendations included: self regulation by biometric providers, compulsory social impact assessment, generic privacy laws and specific regulations[2].

[1] The European convention for human rights (1950) for example in article 8, while declares that “everyone has the right to respect for his private and family life, his home and his correspondence” states that a public authority could interfere the exercise of this right in accordance with the law when necessary in a democratic society and in the interests of national security and public safety, prevention of crime, and for similar reasons mentioned in the article. Many examples were given are given in the appendix, e.g. recall use of the word, ‘unreasonable searches’ in the Fourth amendment of the US constitution.

[2] Such as prohibiting storage of raw biometrics; prohibiting central storage but instead storing data only on a device under the person's control, and subject to security features; creating design standards for biometric measuring

European Group on Ethics in Science and New Technologies (EGE) made a number of proposals in the draft Charter of Fundamental Rights of the European Union (2000). EGE suggested that the charter should highlight two concepts of freedom and dignity as the hallmarks of European society. On the data protection article, EGE proposed the need for respecting an individual's right to protection of personal data, confidentiality of personal data, right to determine which of the data are processed and when and by whom, and, finally, the right to have access to one's own data (for any modification). They referred to three fundamental principles of confidentiality (personal data is part of a person's identity), autonomy (linked with the principle of consent) and right to 'information' and proposed that "No person shall be subject to surveillance technologies which aim at or result in the violation of their rights or liberties" (EGE, 2000, p. 26)[1].

Woodward (2008) with a mention of the fact that the actions of private sector is not regulated under the US Privacy Act 1974, calls for blueprints for Biometric Code of Fair Information Practices (CFIP) and suggest a number of items that have to be included in such code which includes:

1. Notice: The capture of the biometric must be accompanied by a prominent notice. No hidden data collection should be allowed.
2. Access: individuals have right to access their information and should know how the data collector is using data. The data collector should disclose its privacy practices.
3. Correction Mechanism: users should have an opportunity to correct or change their data.
4. Informed Consent: before any disclosure to the third party, the individual must give his/her consent.

Woodward (2008) contends that the user should knowingly and voluntarily provide his biometrics to the data collector in the primary market. The 'use limit principle' should limit access to biometrics which means that the information should only be used for the purpose defined by the data collector and known to the individual. Also reliability and safe guarding of data is among the issues that should be taken into consideration.

Grijpink (2001), in an article on biometrics and privacy, makes a few suggestions of 'rules for the use of biometrics'. Some of these rules that can be added to the points made so far are:

devices; prohibition of the manufacture or import of biometric-measuring-devices that do not comply with the design standards.

[1] The document (EGE, 2000) states that the provision suggested does not define an individual right but it is a principle that has to be respected in a democratic society.

- Sectoral boundaries: Data from one sector can not be used in another sector[1]
- Proportionality: use of biometrics should be proportional with what they are used for
- Subsidiarity: What can be done at sectoral level should not be tackled at the government level. If private storage suffices, no central database should be used.
- Precise delineation of the target group (important for determining the permissibility of the application and communication with the target group)
- External audit of data management profile by independent parties

While it is possible to conclude the non-technical discussions presented so far, right here, and specify how this research, except by collection and formulation of the issues, contributes to the current study, to avoid repetition this task is deferred until the end of this chapter.

The next part of this chapter looks at 'voice', in particular, from a closer distance and makes a quick-yet comprehensive-review of voice-specific knowledge present in the field and problems ahead of voice verification in various contexts.

3.5 Speech Science: A Summary of Previous Research

3.5.1 Body of Literature Covered on Speech Science

Speech processing, whether by human or machine, has a long history. Furui has given an account of five decades of progress in two areas of speech processing which are speaker and speech recognition (2005). However, speech processing spans a wider range of subjects and as it will be revealed throughout this thesis, successful speech-enabled security systems need components depending on all those subjects. Those related areas include speech recognition, speaker recognition, speech coding/compression, speech enhancement, speech synthesis and speech watermarking. This sub-section, therefore, very selectively and briefly, surveys the main ideas, which are later used in the thesis.

The main concepts covered here are:

1. Models of speech production: the speech production models, especially source-filter models, are the basis for feature extraction techniques. Automatically extracted features, as well as those

[1] Grijpink later elaborated on the concepts of chain and supra-chain and transferring the use of biometrics from one level to another in a two part article on barriers to realizing the benefits of biometrics (2005).

calculated by experts such as acoustic features, formants and fundamental frequency are built upon the theory of speech production.

2. Features for speech processing: the emphasis in this chapter will be on Mel-frequency cepstral coefficients which are used in the experimental parts of the thesis. Nevertheless, a short introduction to some other features will be presented.

3. Gaussian Mixture Models (GMM): GMMs are used for speaker verification in the majority of verification experiments in this research as well as for voice conversion for analysis of spoof attacks. Modeling using GMM is the main technique used in the area of voice verification.

4. Hidden Markov Models (HMM): HMMs are likewise, the main tool for speech processing and in this research will be used both for content verification and speech synthesis.

3.5.2 Speech Production and Units of Speech

Speech is our most natural and first modality of communication. The main goal of speaking is putting our messages across (and not communicating our identity). Our auditory and speech production organs are, in parallel, optimised towards this goal. Therefore, there are many connections between our physiology of hearing and speaking, some of which are captured and used, for example, in the area of speech recognition.

With the emphasis on meaning and contents, the phonemes, which are the smallest units in speech[1], have been modeled and studied for speech recognition and synthesis. Words and sentences are built from phonemes. Speech recognition aims at recognition of phonemes from acoustic features extracted from (usually) fixed length frames of samples. The recognition is not only based on acoustic models. Linguistic models are also used for the optimization of recognition outcome (see Jurafsky & Martin, 2000).

Depending on the vibration of the vocal cords, the phonemes could be voiced or voiceless (unvoiced). There are also many manifestations of phonemes which are called allophones (Holmes & Holmes, 2001). The International Phonetic Alphabet (IPA) chart uses 20 symbols for representation of vowels (7 short vowels, 13 long vowels and diphthongs) and 24 for consonants in the BBC Accent of English or Received Pronunciation (RP) (Roach 2004). IPA symbols and categories are broader than those covered in this section; for example supra-segmental symbols

[1] where substitution of one unit for another might make a distinction in meaning (Holmes and Holmes, 2001)

such as stresses and different types of consonants such as pulmonic (bilabial, dental, etc.) and non-pulmonic are differentiated in the IPA symbols which are outside the scope of this Chapter.

The recognition of allophones and choice of phonemes used in the construction of words gives information about the speaker's social and ethnic background. We, therefore, need to model and recognise units of speech for content recognition and in forensic sociolinguistics.

On the other hand automatic speaker recognition, which has to be carried out in a text-independent (TI) and text-prompting mode to withstand replay attacks, does not need any information about the units of speech. Experiments mentioned in chapter 5 (by Yu et al. 1995) show that modeling the transition between speech units does not improve the TI speaker recognition results.

For automatic security applications, we need speech recognition for content verification (and prompting new phrases in each verification attempt) in parallel with a speaker recognition module.

On the contrary, in forensic speaker verification, mostly in post-incidence studies by forensic experts, close examination of the duration of speech units, formants (covered hereafter), and fundamental frequency is helpful for suspect conviction or elimination.

3.5.3 Speech Production Model and Features of Speech

In the main model of speech production, speech is the outcome of the passage of a source signal through a linear filter (which models the human vocal tract). This model is elaborated on in Appendix D (D.1) and a simulation involving the production of a vowel is presented which shows that they verification system based only on the measurement of formants (not the main verification method in this study) will be vulnerable to simple spoofing techniques.

The main acoustic feature used in the experiments conducted in this thesis is Mel-Frequency Cepstral Coefficient. The reason behind this choice is that, first this feature worked better than other features in the tests on IVIE datasets[1]. The second reason is that due to the common usage of this feature for speaker recognition, the generalization and interpretation of results will be

[1] Kinnunen experiments indicate that cepstral coefficients based on LPC estimated spectrum (LPCC) outperform usual MFCC coefficients based on FFT spectrum (on TIMIT database). In my experiments the results of both were similar on IVIE corpus, and for the test set the equal error rates for LPCC was 1.9% while for the MFCC, EER was 1.4% which justified proceeding with the more common method of feature extraction.

more reliable as well as comparable with findings with other similar studies. Many optimization techniques, however, are introduced and implemented in chapters 5 and 7.

In part 2 of Appendix D, Mel-frequency cepstral coefficients (MFCC) and Formants are introduced in detail. The critical analysis of the MFCC calculation steps-presented in D.2-gives an insight into how each step contributes to the robustness of the feature. Quick examination of D.2 is necessary for understanding the technical parts of the next chapters.

Over thousands of years, the main functionality of speech has been communication and conveying the messages, not recognition and providing proof of our identity. For this reason it sounds reasonable that the distinct speech units have been shaped by our auditory characteristics and the use of non-linear Mel scaled filters that will produce good results for speech recognition. For speaker recognition, however, there are no experimental or theoretical grounds to prove this. This topic (possibility of optimizing filter-banks for speaker recognition and its potential for improving the recognition results) is dealt with in chapter 7.

3.5.4 Content Verification

Regardless of the context, the main objectives of speech analysis in security, surveillance and forensic applications are to find the speaker of a piece of speech (who); and the contents of speech (what). For the latter, we simply need to produce a sequence of symbols representing the units of speech for a recording under investigation. In this section, the techniques for speech recognition -especially hidden Markov modeling-is studied.

There are two problems to be resolved before being able to recognise speech. Firstly we need to characterise the events in speech (for example, phonemes) in order to be able to recognise them. Secondly we need to translate a piece of speech into those events. The problem is that each time we produce a phrase, it has different characteristics (Furui, 2001), not only, in terms of acoustic features of each speech unit, but also in terms of duration and place of occurrence of those units.

For the first problem (characterizing units of speech) feature extraction, followed by modeling techniques such as vector quantization, neural networks and Gaussian mixture modeling (GMM) could be used (for a review of the history of development of these techniques see Furui, 2005).

For the second problem (matching the events of speech) two promising techniques are dynamic-time-warping (DTW) and hidden Markov modeling (HMM). Dynamic time warping uses dynamic programming techniques to expand and contract the time axis to match the phonemes between a piece of speech under analysis and a reference template (Furui, 2001).

A newer modeling technique which answers both questions simultaneously is the hidden Markov model. It allows modeling each unit of speech independently and concatenating the models to produce more complex models for longer phrases.

Hidden Markov models are introduced in section 3 of Appendix D (D.3). In short, three questions need to be considered in advance of HMM modeling. These are:

1. **Evaluation/Recognition:** Having a known HMM model we need to calculate the probability of generation of a series of observations.
2. **Decoding:** Finding the most probable state sequence for a given observation sequence (Viterbi Algorithm).
3. **Training:** Optimizing the model parameters to obtain the best model that represents the set of observations (Baum-Welch Algorithm).

In the interest of simplicity we can first suppose that each model is trained on one phoneme. In this condition, HMM modeling allows concatenation of these blocks to build larger models representing words and sentences. Models for virtually any phrase can be built by joining phoneme HMMs. By using evaluation/recognition algorithm (task 1) we can calculate and compare the probability of seeing a set of observations given various models (word A, word B, etc.) and choose the best model which matches the observation.

If the HMMs are trained for allophones and by using task 2 we can build larger models for one phrase and decide which realization of the phrase (in terms of accent and choice of allophones) is chosen by the speaker (this will be demonstrated by an example shortly).

Since the phonemes exhibit different characteristics in different contexts because of co-articulation[1] effects, training one HMM for each phoneme does not necessarily bring about good recognition results. Therefore, these context-independent models of phones (or monophones) are replaced with context-dependent models which are called tri-phones. A tri-phone model is a phoneme in its context, for example, tri-phone model of the vowel ≅Y in 'coat' is different from the one in 'mole' as depicted in Figure 3-1. Both words are modeled by three phonemes (sil denotes silence). Each three state HMM model represented in the figure, models a phoneme with

[1] The movement of articulators to anticipate the next sound or perseverate from the last sound (Jurafsky and Martin, 2009).

regard to its prior and subsequent phoneme. Therefore, there are no overlaps between the models used for phonemes in these two words. Since the number of models will be extremely large (around M^3 for M phonemes[1]) the context is modeled based on the 'category of phonemes' instead of 'particular phonemes' for example, one model for ≅Y situated between all nasal consonants (such as n or m) and all fricative consonants (such as s or ch) can be trained (see Jurafsky & Martin, 2000 and 2009).

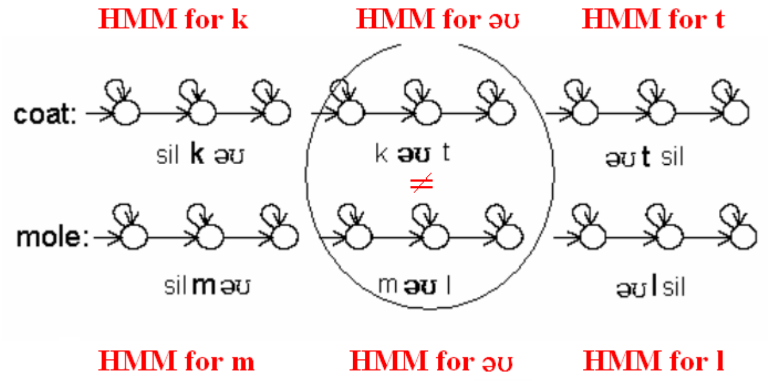


Figure 3-1 Demonstration of Use of Triphones in Two Words

Despite many suggestions for optimization of HMM for speech recognition, the HMM modeling is the dominant content verification technique in the field.

Content verification, not only is helpful in speaker recognition systems or for unveiling the contents of a difficult piece of speech, but also, could be used to give some information about the speaker's background. In such cases HMM is used for determining the most probable realization of a word given an observation sequence (by Viterbi algorithm). Figure 3-2 shows possible realizations of the word 'water' pronounced in different accents or by people with different social/cultural backgrounds. A piece of speech can help us determine which of those classes is more likely to account for the speaker's background.

[1] Some of the combinations are not possible.

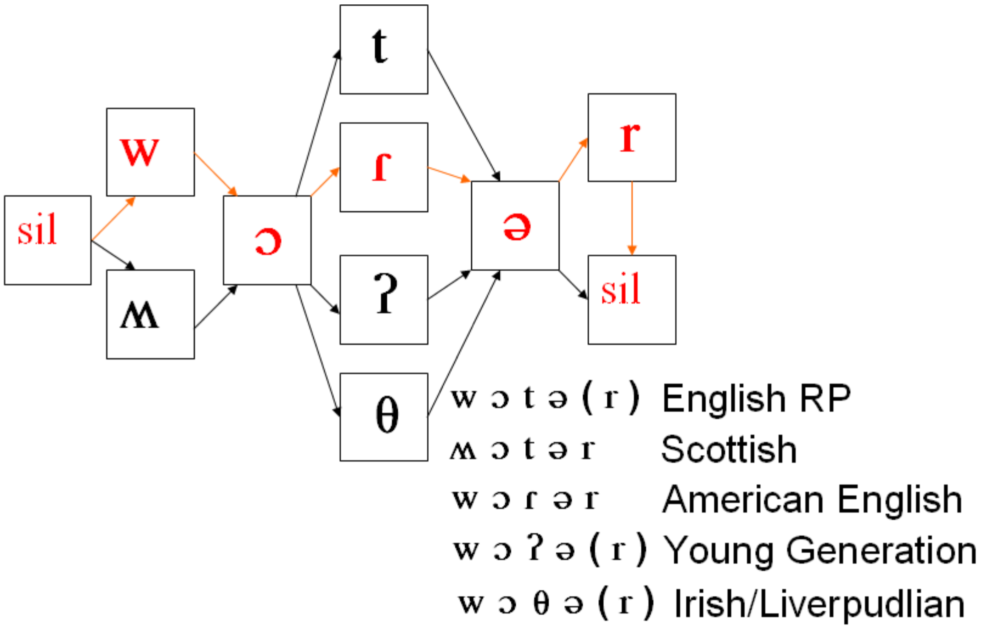

sil
w
ʍ
ɔ
t
ɾ
ʔ
θ
ə
r
sil
w ɔ t ə (r) English RP
ʍ ɔ t ə r Scottish
w ɔ ɾ ə r American English
w ɔ ʔ ə (r) Young Generation
w ɔ θ ə (r) Irish/Liverpudlian

Figure 3-2 Various Phonetic Realizations of the Word 'Water'

3.5.5 Speaker Recognition

Speaker recognition is the process of identifying a person through his/her speech. The first attempts to build an automatic recognition system dates back to 1960s, one decade after the introduction of speech recognition (Furui, 2005). Speaker recognition systems can be text-dependent or text-independent[1]. It is interesting that the early recognition systems were text-independent but due to their poor performance, researchers switched into text-dependent methods. It was not until 1990s, when new HMM-based systems appeared, that text-independent recognition with rotating passwords regained proper attention (Furui, 2005).

In a text-dependent system, the speaker is asked to say a fixed phrase, such as a password consisting of a number of digits, several times in an enrollment session. The recorded speech data is used to adjust model parameters for that fixed phrase. Such systems are highly prone to impostor attack (obtaining and replaying of a recorded piece of speech from the true speaker) compared to text-prompting systems, which could prompt a speaker for a new phrase each time. A short description of Multivariate Gaussian mixture modeling, which is widely used for speaker recognition, is given in section 4 of Appendix D (D.4).

[1] A text-prompting system consists of a speaker recognizer in text-independent mode and a speech recognizer in speaker-independent mode.

3.5.6 Overview of Forensic Applications of Voice Verification

Broeders (2001) undertook a review of forensic applications of audio analysis in categories such as speaker identification by ear-witness, speaker identification by experts, intelligibility enhancement, integrity and authenticity analysis, disputed utterance (speech to text) and linguistic authorship studies.

Similarly Foulkes and French (2001) categorised the applications in the legal context of phonetics and sociolinguistics into 4 categories:

1. De-ciphering the content of a piece of speech, for example, from a plane’s black box or in a threatening message;
2. Determining the speakers’ profile including sex, age, social background, region and idiosyncrasies. This category may include figuring out the likelihood of intoxication and use of alcohol;
3. Speaker Identification based on the data from various sources e.g. a threatening message, a criminal recording or secret surveillance recording;
4. Voice parade or speaker verification by a lay person for cases in which a crime is committed or has been attempted but the only piece of evidence available is the voice heard by ordinary people, for example, in a phone call or a suspect’s voice heard in the dark or under a mask.

The first two categories involve content verification which was discussed before. This thesis does not concentrate on speech recognition. Nevertheless, almost all problems and techniques are in common to both speech and speaker recognition and the experiments carried out here have relevance for both fields. The following two sections analyse the questions in need of investigation for the construction of voice parades and those arising from previous research in the field of forensic speaker verification.

3.5.7 Recollection of Voices and Construction of Voice Parades

Key questions in relation to speaker verification by lay person revolve around how the voice parade should be constructed and what factors affect the reliability of such types of identification, especially how much the passage of time affects the memory of voice and identification results.

Foulkes and French (2001) cite Kunzel (1994) who revealed that witnesses' performance in recollection tasks depends on various factors including hearing ability, the degree of familiarity with the voice in question and the fact that exposure to the voice has been active or passive.

Does user knowledge of phonetics affect the reliability of decisions about the source of voice in auditory identification? Schiller and Koster conducted an experiment in which two groups of participants (experts in phonetics and untrained) were asked to identify the voice of a target speaker among five other speakers (foils). They showed that when phonetically trained and untrained listeners were exposed to the same speech materials, the listeners in the first group performed significantly better in identifying a speaker (1998).

The memory of voice has not been left unexamined. Saslove and Yarmey (1980) and Clifford et al. (1981) conducted a number of experiments in this regard[1]. Hollien and Schwartz (2000) showed that a speaker's memory for aural identification decays over time. They also showed that non-contemporary samples receive lower identification rates but reported that for latencies from four weeks to six years the effect on the results was only about 15–25 %. For 20 years the drop was more substantial (31 %).

The number of voices that we can remember is another important factor in constructing the voice parades. Watt in the same context points out that the number of distinct voices that we can, in general terms, hold in memory is still unknown (2009). There are various design considerations posed by different groups for shaping voice line-ups or parades. The Home Office in Britain has published 'advice on the use of voice identification parades' (Home Office Circular, 2003) which is based on the procedures devised by DS McFarlane (Metropolitan Police) for a case brought to the central criminal court in 2002. The recommendations rule out the use of 'live' parades. The guideline suggests collection of at least 20 samples from people within the same ethnic, social and age group as the suspect. It also recommends undertaking the procedure within 4-6 weeks of the incidence under question. The guidelines state that one sample from the suspect with the length of one minute and 8 samples from other people should be chosen and recorded on tapes.

[1] Saslove and Yarmey (1980) reported no significant difference between voice-parade tests immediately after hearing someone's voice and those carried out after a delay of 24 hours on students. Clifford et al. (1981) also evaluated the effect of time delay on aural identification with two experiments. The first one involved a maximum delay of 130 minutes and the second one involved a maximum delay of 14 days with various delay groups within them. In experiment 1, delay had no overall effect (though best results were produced by the shortest delay) while in the second one, the delay had such an effect.

This thesis does not focus on the verification of voice samples by humans and is unable to draw conclusions about the reliability of voice parades considering all the possible questions that exist in this field. Nevertheless, it is proposed that human decisions should be quantified and undergo the same process of reliability analysis as that outlined in chapter 4, regardless of where the scores come from: an automatic system, an untrained person or an expert. A discussion on this appears in chapter 9.

3.5.8 Forensic Speaker Verification

A thorough account of historical developments in forensic speaker recognition has been given in many texts such as (Lindh, 2004) and (Eriksson, 2005). Here, the key points are summarised along with questions in need of further investigation under separate headings:

1. Voice-print and Earliest Attempts at Spectrographic Recognition

The chronological events in the use of the term 'voice-print' which were simply the "patterns in the spectrogram" are presented in section 1 of Appendix E. Previous research presented there shows that the context has a great effect on the spectrograms (voice-prints). Further questions need pondering such as: Are forensic features more robust to channel distortion than acoustic features (introduced in Appendix D)[1]? Are they robust in terms of independence from factors such as age? Are they robust against disguise?

2. Questions about the Reliability of Forensic Speaker Recognition

The starting point in this discussion is to acknowledge the fact that that when we talk about the reliability of forensic speaker verification it is not absolutely clear to which approach we are referring. A mention of Broeders's classification of approaches is necessary here.

Broeders (2001) divided forensic identification approaches, by experts, into three groups. In one group, the experts use the language-specific combination of auditory phonetic analysis and a variety of acoustic measures. Those experts only work on one particular language in this group. Broeders refers to a strong 'subjective element' in this approach. In the second group are the experts who base their decisions on semi-automatic measurements of particular acoustic parameters such as formants and articulation rates. The third group consists of an approach based

[1] A quick answer in the context of text-independent voice verification based on Tinnuen's experiments (2004) is that formants are not more reliable features than cepstral coefficients in normal conditions.

on automatic speaker recognition. This approach is global i.e. it does not use any specific acoustic speech parameters and treats the signal as a physical phenomenon. GMM based speaker recognition with acoustic features could be considered as an example of this approach. Broeders identifies the within and between speaker variations as a common problem in all approaches, and the channel effect as being specific to the third approach.

Research carried out by Nolan and Grigoras (2007), McDougall (2007), and Morrison (2008) is summarised in Appendix E (E.2) to demonstrate that the formants (whether their values or dynamics) have a pivotal role in contemporary speaker recognition.

Experiments conducted by Endres et al. (1971), Kunzel (2001), Byrne and Foulkes (2004) and Guillemin and Watson (2006) show that formants are prone to change over the time and more importantly, they are affected by channel.

3. Current Problems of Forensic Speaker Recognition

Two serious problems with forensic speaker recognition have already been mentioned. The first one is the lack of standard, objective and quantifiable measures for recognition. The second problem is the unreliability of the parameters used for forensic recognition in various contexts.

Bonastre et al. (2003) after reviewing methods of aural recognition, spectrogram recognition, forensic phonetic recognition and automatic speaker recognition, concluded that for many reasons, including the need for user co-operation, possibility of synthesis, control over the linguistic contents and control over the recording conditions and equipment “at the present time, there is no scientific process that enables one to uniquely characterise a person’s voice or to identify with absolute certainty an individual from his or her voice” (p. 3)[1].

Eriksson’s remark is mentioned in Appendix E in which he concludes that “voice-printing is still done by private detectives and other non-academic experts but nobody in the speech science community believes in its usefulness for forensic purposes any more.” (Eriksson, 2005, p. 5).

A third problem that should be added to this list arises from the fact that normally, forensic evidence is rare and is not subjected to any great extent to statistical analysis.

Koolwaaij and Boves (1999), after an analysis of a harassment case investigated in a text-independent speaker verification mode, stated that the forensic speaker verification values can

[1] The problems they mention for forensic-phonetics-based speaker recognition however were limited availability of qualified phoneticians, complications of different languages, inadequate resources to determine how typical voices are.

replace the subjective values in the field of forensic study only if large amount of data, that match the case, is available for training the impostor and world/global models.

To revive forensic speaker verification, it is necessary to show that statistics support its use under various conditions. The practice of methods in the field should be underpinned by the unambiguous proof of concepts. This is only possible through quantification of forensic scores and by undertaking the tests presented in chapters 4 through 8 of this thesis.

3.5.9 Interpreting Results in Forensic Speaker Recognition

In addition to the previously discussed problems with finding reliable features for speaker recognition, the questions recently dealt with in the area of forensic speaker identification revolve around statistical interpretation of forensic analysis regardless of its underlying process. In this new wave of studies, the same acoustic features and techniques, which are used for large-scale automatic speaker verification, could be used as the building-blocks of analysis.

Chapmod and Meuwly (2000) tried to propose methods and guidelines for interpreting the evidence in the field of speaker recognition. First, they argued that setting a threshold for verification by scientists is interfering with the judicial process. The threshold relates to the concept of reasonable doubt which can not be decided by the scientists. They stated that a probabilistic framework based on Bayes theorem can be helpful to assist scientists to assess the value of scientific evidence, to help jurists to interpret such evidence and to clarify the roles of scientists and of members of the court.

If we define *O(E,I)* as the odds that the suspect has produced the recording *S* (formulated as evidence *E* given by the scientists), given the circumstances of the case (*I*) and if H_1 is the hypothesis that the suspect has produced the speech and H_2 the hypothesis that someone else in the population has made the suspicious speech, then:

Equation 3-1

$$O(E,I) = \frac{P(H_1 \mid I,E)}{P(H_2 \mid I,E)}$$

Based on the conditional probability and Bayes theorems:

Equation 3-2

$$O(E,I) = \frac{P(E \mid H_1, I)}{P(E \mid H_2, I)} . \frac{P(H_1 \mid I)}{P(H_2 \mid I)} = LR(E).O(I)$$

On the left of the equation 3-2 we have the posterior odd which, according to Chapmod and Meuwly, is what the court needs (the odd that the suspicious speech belongs to the suspect given the circumstances and the observations). On the right of the equation, there are two terms, *LR(E)* which is the likelihood ratio estimated by the scientists and the *O(I)* ,prior odds, which is independent of the evidence and is decided by the court and jury.

Champod and Meuwly state that:

"The scientist is generally not in a position to assess the odds in favour of an issue, because a complete assessment must combine both the forensic statement (*E*) and background information (*I*). The scientist does not usually have access to the background information that is available to a member of a jury or a judge." (Champod & Meuwly, 2000, p. 8).

I would like to challenge this argument by reminding the reader that:

Equation 3-3

$$LR(E) = \frac{P(E \mid H_1, I)}{P(E \mid H_2, I)}$$

It could be easily noted that the scientist should be aware of background information (*I*) to estimate the likelihood ratio. In practice what a scientist may do is divide the circumstances into two classes; relevant and irrelevant to the estimation of a likelihood ratio. For example, the fact that the speech is obtained in a phone conversation affects how the probabilities are calculated. Some other information used by the judicial system may be irrelevant to the calculation of likelihood ratio.

Despite the separation reflected in equation 3-2 and the attempt to distinguish between the role of two groups (scientists and member of judicial system), one may reasonably ask how the scientist can decide one aspect of circumstances is irrelevant to the calculation of likelihood ratios. Currently available techniques, allow estimation of:

Equation 3-4

$$LR(E) = \frac{P(E \mid H_1)}{P(E \mid H_2)}$$

and not equation 3-3, in predefined circumstances, such as noisy and channel affected conditions. The effect of a variety of dimensions in circumstances is unknown on the voice. For practical purposes, however, we assume that they have no effect on the evidence. This issue is elaborated on to a greater extent in chapter 4. The circumstances also affect the relevant population reflected in the hypothesis H_2.

The framework, described above, is used by Drygajlo et al. (2003) to introduce a two-stage statistical approach. They used a GMM based probability estimation technique for assigning the scores to the observations and a univariate statistical analysis for estimation of $P(s \mid H)$ (for both hypotheses) based on the scores obtained from GMM models where:

Equation 3-5

$$s = P(O \mid \Pi)$$

Π is the GMM model of the suspect for hypothesis H_1, and the GMM model of rival population for hypothesis H_2 and *O* is the observation e.g. the speech features.

To estimate the $P(s \mid H)$ therefore Drygajlo et al. (2003) needed another 'control set' and described the use of three databases: the potential population database (*P*) used for training GMM models for the population's speakers; the suspected speaker reference database (*R*) used for training GMM models for the suspected speaker; and the suspected speaker control database (*C*) used for estimation of second stage statistical models.

Botti et al. (2004) presented an adaptation of T-Norm normalization in the same framework for compensating the mismatched recordings. Not surprisingly, normalization based on the use of the mean and standard deviation when the control database in available in mismatched conditions can reduce the errors. Alexander et al. (2004)[1] analysed the effect of mismatched conditions on aural and automatic recognition when automatic recognition was based on GMM modeling and RASTA-PLP features. They found out that automatic speaker recognition outperforms aural speaker recognition in matched conditions but in mismatched conditions their performance becomes similar. They also tried to identify the features that human subjects attend

[1] Each pair in the last three cited publications has at least two authors in common.

to when trying to recognise a voice by asking the subjects 'what factors they considered in recognizing the questioned recording'[1].

Gonzalez-Rodriguez et al. (2006) introduced the same Bayesian framework for the interpretation of evidence with the same set-up described above for control, reference and population databases. They proposed a leave-one-out procedure for the cases in which there is only one recording available from the suspect[2].

3.6 Applications of Voice in Crime-Reduction and Investigation

The application of voice in crime investigations and for crime reduction, can best be summarised by grouping the approaches into four categories and illustrating them through 9 groups of scenarios.

Category One: Automatic Speaker Recognition for Crime Reduction

G-Scenario 1 (Group of Scenarios): Voice Verification by Smart Cards

Appendix F (F.2) shows that current smart card standards allow for the easy substitution of PIN data with biometric data. In this scenario, user specific models are placed on a smart card. In various situations a person could use his or her voice instead of a PIN number or passwords. The verification by voice is just one of the several security checks (see F.2). The verification process could be supervised e.g. at shops or unsupervised e.g. at ATMs. The existence of noise is a determining factor in such applications.

G-Scenario 2: Use of Voice Verification in Long-Distance Authentication (over the internet, mobile phone and landline telephone)

A user calls the customer service of a communications company and tries to change his usage plan. The user may talk to an agent (supervised) or go through the menus (unsupervised). The security implications are definitely different and are discussed in chapter 9 in light of the

[1] It is worth mentioning that what we think we do in many cases is different from what we intuitively do. We are not aware of many of our internal processes. Therefore asking people about the features they use for recognition (although a valuable initiative) may not be a reliable source of identifying those features.

[2] The procedure involved splitting the recording into N segments, using N-1 of them in each attempt for training models (P) and the other one for scoring (C). The approach will give us N scores representing the within speaker variation. Based on this procedure they suggested a within source degradation prediction technique to re-estimate the suspect's probability distribution function. They also proposed a likelihood ratio adjustment technique with the objective of minimizing the number of suspected non-perpetrators speakers obtaining likelihood ratio above 1.

experiments. Also, many scenarios could be pictured in which a person uses his/her voice along with credit card information for placing an order or finalizing a purchase.

An application of voice verification for crime reduction and security enforcement is in electronic monitoring of curfew sentenced offenders through random calls.

G-Scenario 3: Voice Signature

The security principles which are needed for presentation of a voice signature scheme are summarised in appendix F. My proposition for a voice based signature is that the user with a voice certificate from a biometric certification authority (I call them BCA, in contrast with CA), can read a portion of the contract (or any document that is being signed). This raw voice data, along with the person's certificate, could be encrypted by the person's private key and be used as a voice signature. The detail will be presented in chapter 9 and the related Appendix (L).

Category Two: Automatic Speaker Recognition in a Forensic Context

G-Scenario 4: Conviction or Elimination of Crime Suspects

The question of forensic suspect elimination/conviction is whether or not a piece of speech belongs to a person, or two pieces of speech belong to one person. Expert opinion, semi automatic approaches and automatic approaches could be used in such applications. The consequences of a wrong decision in this scenario are dreadful. However, a line should be drawn between the time the data is used for 'crime investigation' and when it is used for conviction in a court of law.

G-Scenario 5: Identification by Lay-Persons and through Voice Parades

A lay-person has to decide (based on his memory) if he can choose between a number of speech samples the one that is most likely to belong to a person at a crime scene (heard by the person).

Category Three: Speech for Crime Investigation

G-Scenario 6: Determining the speaker's profile

Any information about the speaker's race, gender and social background could be of assistance to crime investigators. Wrong information could be misleading but this becomes much more serious if socio-demographic characteristics are used as evidence in the court.

G-Scenario 7: Deciphering contents of a piece of speech

Many scenarios could be pictured in which a voice message is available but its content is unclear. Speech science could be of great help for noise removal and deciphering the speech contents.

Category Four: Speech for Surveillance

G-Scenario 8: Eavesdropping and Wire-tapping

A government agency would receive permission to control the phone conversations or try to covertly listen to suspects' conversations in special circumstances

G-Scenario 9: Surveillance for Public Security

This scenario could also be placed in category one. Use of speech has not yet been explored in many potential applications in which currently CCTV is in use. Simple examples are the detection of alarming sounds such as screams, gun shots, collisions or more complicated scenarios involving the detection of emotions, use of alcohol and abnormal behaviour. A distinction between surveillance by government and surveillance by the private sector should be made since use of voice in each class has different ramifications.

3.7 Broad summary and its bearing on subsequent chapters

In this chapter, several technical and non-technical aspects of use of voice for crime reduction and investigation were analysed. This final section of the chapter recapitulates the non-technical discussions (up until 3.5) and technical (voice specific) issues, respectively.

An exhaustive survey of concerns about use of biometrics was presented in 3.4. The outcome of such a critical examination of ideas expressed in this context was a list of issues which may or may not be relevant in the case of voice verification but have to be addressed for any biometric system. Going through the list will demonstrate that voice is one of the less controversial biometrics from human and ethical perspective (This will be done in chapter 9).

The analysis of legal frameworks demonstrated that they allow 'use of biometrics contingent upon user consent' and 'use of biometrics without consent' if there is a sound 'rationale' for it. Some of the questions around the use of biometrics, especially for surveillance and 'in the interest of public', are matters of opinion and could not be settled once and for all. The democratic processes might yield different outcomes in different cultures but reliable decisions are only achieved if the public are well-aware of the real ramifications of use of biometrics.

Chapter 9 will discuss how the design of a biometric system can eliminate hitherto unaddressed concerns or reduce their seriousness.

As for voice surveillance, which may capture voice data for which consent has not been granted there is a need for a clear legal framework and detailed guidance on its use. It can be argued that if voice overcomes its technical difficulties and presents itself as a reliable means of monitoring in public places, there will be a strong case, in circumstances where voice surveillance is deemed necessary, for developing unambiguous guidance for its ethical use bearing in mind current EU directives, human rights legislation and other legal instruments. To draw an analogy, the Information Commissioner's Office (ICO) in the UK[1] has revised and published a CCTV code of practice (2008)[2] in accordance with DPA. Based on the concepts presented in EU Directive 95 and DPA, voice surveillance is no different from video surveillance. The reason that such a code of conduct is not released for voice surveillance is probably that the need for it has so far not been identified. Chapter 9 touches upon the suggestions which could be incorporated into such a code of conduct for voice surveillance.

To recapitulate the technical issues, what we had here was an exposition of how speech science has developed over the decades to offer features and models for speech and speaker recognition. The forensic discussion, however, showed that there is a huge discrepancy in the reported confidence level of voice verification.

The main focus of this Thesis will be on security based speaker recognition i.e. scenario one and scenario two. This is because the majority of applications which call for voice-verification fit into these scenarios. Reliability of voice verification can hardly allow its use in the applications where decisions are vital. Nevertheless, even for more crucial uses of voice, what we need is a proof of concept which could not be guaranteed except through an evaluation framework.

Four facts should be taken into account about voice: Verification methods are various (different modes e.g. automatic or expert-based, different features and different models). The reliability of

[1] Which is set up to uphold information rights in the public interest and to promote data privacy for individuals

[2] Information Commissioner's Office (2008), 'CCTV Code of Practice', Last Retrieved 2010-04-27,<http://www.ico.gov.uk/upload/documents/library/data_protection/detailed_specialist_guides/ico_cctvfinal_2301.pdf>

methods in different ‘contexts’ is not tested. The contexts are not well defined. The possibility of spoofing and spoof detection is not studied.

These facts compel us to seek an evaluation framework which can ‘prove’ that a voice-based security system, or even an expert-dependent forensic speaker verification procedure, produces reliable decisions in its usage conditions. The framework should determine reliability under a wide range of conditions and should be independent of the system and its scoring mechanism.

The next chapter elaborates such a framework and how it can build on our present knowledge of speech, and become progressively more inclusive. The next chapters evaluate a moderately optimised speaker verification system under various conditions. However it should be emphasised that while the discussions in chapters 5 through 8 are mainly technical, they have a serious bearing on the conclusions and design considerations presented in Chapter 9.

Chapter 4 : A Dynamic Framework for Security and Reliability Evaluation of Voice Verification Systems

4.1 Goal and Organisation of This Chapter

In this chapter the need for development and standardization of a comprehensive and dynamic framework for security evaluation of voice verification systems in face of spoof attacks is highlighted and important steps toward development of such a framework and its related evaluation system is taken.

Such a framework should be used for the development of evaluation systems. Therefore architectural requirements and issues and reporting details are discussed as well.

Research results partly reported in chapter 2 and mainly reviewed here show that spoof attacks pose severe security threats to biometric verification systems. Not only is voice biometry unexceptional, rather it is more vulnerable to a broader range of security risks because of its unique qualities such as ease of transformation, manipulation and transferring from one point to another. This fact calls for the development of security evaluation systems for various types of spoof-related threats and for defence mechanisms for protecting the systems.

The present literature lacks, nevertheless, a comprehensive, flexible and dynamic framework for security analysis of biometric and especially voice based verification systems when spoof attacks are taken into picture. This chapter aims at highlighting the vulnerabilities, classifying the threats and clarifying requirements for such a framework, with the capability of supporting the transfer of vulnerability evaluation results into legacy knowledge in the security domain. The proposed design and architecture for the framework facilitate-comparison, in an independent and isolated way, of various systems as well as definition of the threats.

In short, the three objectives pursued throughout this chapter are: clarifying the meaning and requirements of a dynamic framework for security evaluation and enhancement of speaker verification systems; spotlighting the threats to voice verification systems and their severity; and finally accommodating the requirements through a framework design and the development of an evaluation system.

The key point put forward here is that a comprehensive security evaluation mechanism in this context is missing at present. An important secondary point emphasised in the chapter is that, due to the evolutionary nature of defence enhancement in security systems, a dynamic framework is required – not only, to evaluate the system's robustness but also to facilitate further recognition of attacks and to raise alarms in event of probable attacks. Without such a framework, neither the documentation of spoof attacks, nor the 'fight-back algorithms' (counter measures) will become re-usable items in the security domain legacy.

After this introduction a quick description of elements of verification systems especially in an identity management system will be presented in 4.2. The meaning intended by dynamic framework is clarified in 4.3. Available standards are reviewed in 4.4 with the aim of demonstrating that such an evaluation framework is missing in the security domain. In 4.5, 4.6 and 4.7 the points and types of vulnerability based on the research review and close analysis of the system structure will be identified and listed. Then the architectural and specific reporting requirements will be discussed in the next part sub-section (4.8). Performance and security evaluation metrics are presented in 4.9. Attack detection blocks are elaborated on in 4.10. For evaluation of forensic speaker verification a detailed analysis and customization of the framework is presented in 4.11. Finally a summary conclusion describing how the analysis proceeds in the subsequent chapters is presented.

4.2 Outline of Voice Verification for Authentication and Access Control

For voice verification systems that prompt for new sentences (a requirement for being protected against replay attack) one promising design consists of two separate blocks: one for verification of speech contents; and the other for matching characteristics of speech against the speaker's statistical model. Speaker recognisers can use many techniques – such as template matching, neural networks, nearest neighbor search and modeling techniques based on Gaussian mixtures (Reynolds, 2002) or hidden Markov processes or combinations of these. A recommended choice for commercial and evaluation applications would be using hidden Markov models (HMMs) for speech recognition while Gaussian mixture models (GMMs) are preferable for speaker recognition (Heck, 2004).

The voice verification module is logically used in a larger system for identity management and access control. Resources and applications (separately, or through a portal) can delegate their authentication process to this authentication system (Figure 4-1). When a user requests access to a resource, the control asks for a proof of identity, such as a token or a web-cookie which will be checked with the identity management system. If this is not present the user is redirected to the identity management system to communicate directly with it and authenticate identity. One of the ways that the proof of identity can be provided is through submitting voice data. After that the voice verification module goes through the process of feature extraction and comparing the features with the user's model in the database (Figure 4-1).

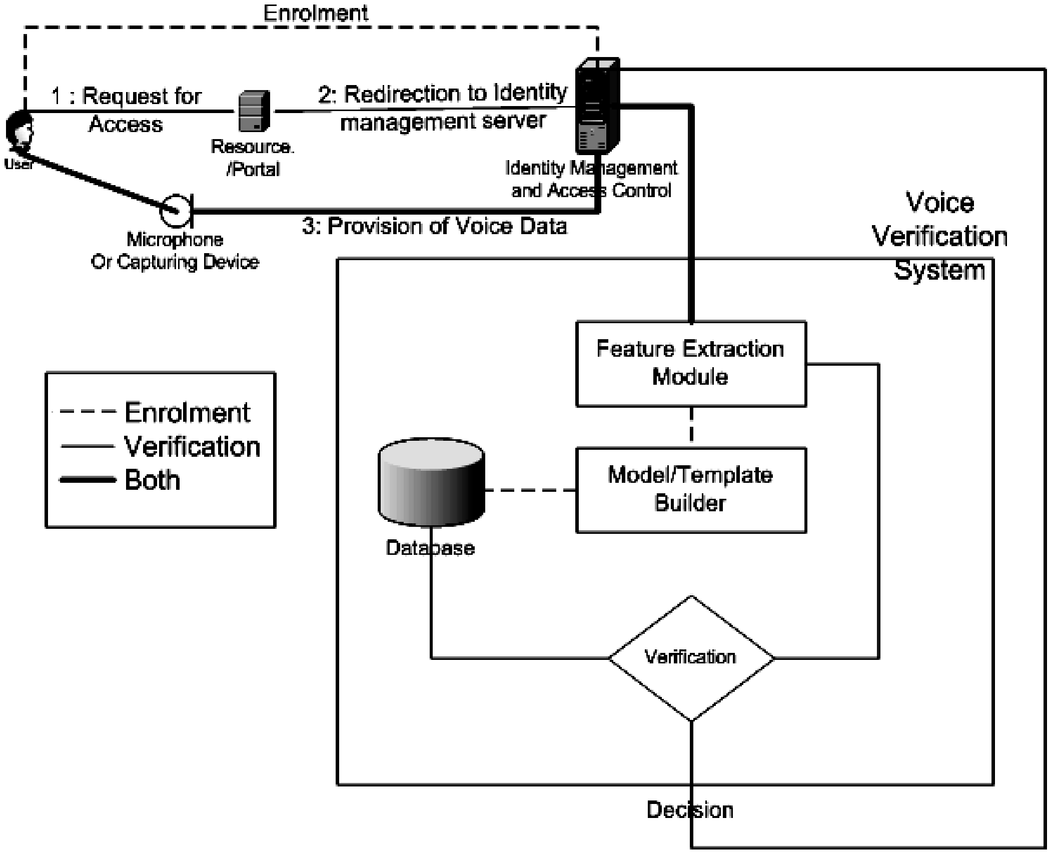

Enrolment
1 : Request for Access
2: Redirection to Identity management server
User
Resource. /Portal
Identity Management and Access Control
Microphone Or Capturing Device
3: Provision of Voice Data
Voice Verification System
Feature Extraction Module
Enrolment
Verification
Both
Model/Template Builder
Database
Verification
Decision

Figure 4-1 An Identity Management System with voice-enabled authentication module

Many vulnerable points could be identified in this scheme. Some of these points are common to all the authentication systems, and the fact that speech data is being used as the biometric does not change them. Spoof related vulnerabilities, however, as well as vulnerabilities related to use in adverse noise conditions, and the relation between these two, are specific to voice biometry.

The sketch presented in this sub-section helps us to define what is meant by the evaluation framework and to specify the components whose security/reliability needs further analysis.

4.3 Clarification of the Meaning of a Dynamic Framework

Security evaluation processes are criticised (Jackson, 2007) for being costly, ineffective in fighting the real threats (by primarily focusing on documentary procedures) and time-consuming (making the evaluation results obsolete by the time the tests are completed). On the other hand, it is evident from the study of security breaches that crime prevention, like other evolutionary struggles, needs to be adaptive to avoid obsolescence (Ekblom, 1999).

In this context, crime prevention involves perpetual monitoring of the processes, detecting the threats and mitigating the risks associated with these threats. In addition, the response of verification systems to a particular attack should be measurable and comparable ~~(for two system)~~ for the sake of evaluation purposes.

The dynamic framework described here, is a formally or semi-formally specified scheme independent of any particular verification system, which measures the robustness of verification system in terms of known performance rates in the face of extensible list of spoof attacks and in adverse conditions.

The concept of flexibility can be clarified by translation into requirements for a security evaluation system:

- A security evaluation system has to be independent of the voice verification system to be 'pluggable' into any system for comparison purposes;
- It should be capable of dealing with new threats and allow unambiguous definition of threats in pre-defined classes;
- It should allow modular changes in framework design to accommodate new metrics and facilitate test automation;
- The attack algorithms implemented by different perpetrators should be documented as legacy knowledge in the security domain.

Apart from flexibility requirements the evaluation system should provide hints for detection of spoofed signals by specifying the footprints of each spoofing technique and evaluate the success of spoof detection in relation to the voice verification system.

Available biometric standards and evaluation frameworks are reviewed in the next section with this question in mind: 'Are there standard sets of vulnerability tests available to reassure us that security in a voice verification system cannot be compromised?'

4.4 Overview of Biometric and Security Evaluation Standards

There are two committees of the International Organisation of Standardisation (ISO) which are involved in the development of biometric standards[1]. The first one, ISO/IEC JTC-1 SC-37, responsible for the international standardisation of biometrics, has published 16 biometric related standards up to April 2007. Among these standards: four of them relate to application programming interfaces (BioAPIs); two of them specify the framework and methods of conformance tests to BioAPI; two of them concern biometric testing and reporting criteria, including the evaluation types and error rates; and eight others define biometric data-exchange formats for fingerprints, facial images, iris images and vascular images. By the same token, the US INCIST M1 committee under the USA National Standards Bodies has published 17 biometric standards. Two of these standards concern BioAPIs and the interfaces between biometric systems and other systems, seven describe data-interchange formats, four define biometric profiles for specific applications (such as border management and access control), and finally four others relate to performance testing and reporting.

In addition, the Common Criteria (ISO/IEC 15408) information technology security evaluation standard provides “a common set of requirements for the security functionality of IT products and for assurance measures applied to these IT products during a security evaluation” (p. 9)[2] . The IT products could be in the form of software, hardware and firmware. By defining the assurance measure and the evaluation process Common Criteria aims at helping the users and consumers decide whether or not the product satisfies their security needs. This is the objective that is pursued in this chapter with regard to voice verification systems.

There are currently at least three sets of biometric security evaluation standards under Common Criteria, published by the governments of the UK, US and Canada. The British version, Biometric Evaluation Methodology (BEM)[3], describes possible threats to biometric verification systems and introduces some vulnerability tests.

Fig. 4.2 summarises the given information on the biometric standards and standardisation bodies.

[1] See ‘Biometric standards - An update’, 2006 and ‘2007 Annual Report on the State of Biometric Standards’, 2007.

[2] ‘Common Criteria for Information Technology Security Evaluation: Part 1: Introduction and general model’, Ver 3.1, 2009

[3] Common Criteria Biometric Evaluation Methodology Working Group, ‘Common Criteria, Common Methodology for Information Technology Security Evaluation’, Version 1.0, August 2002

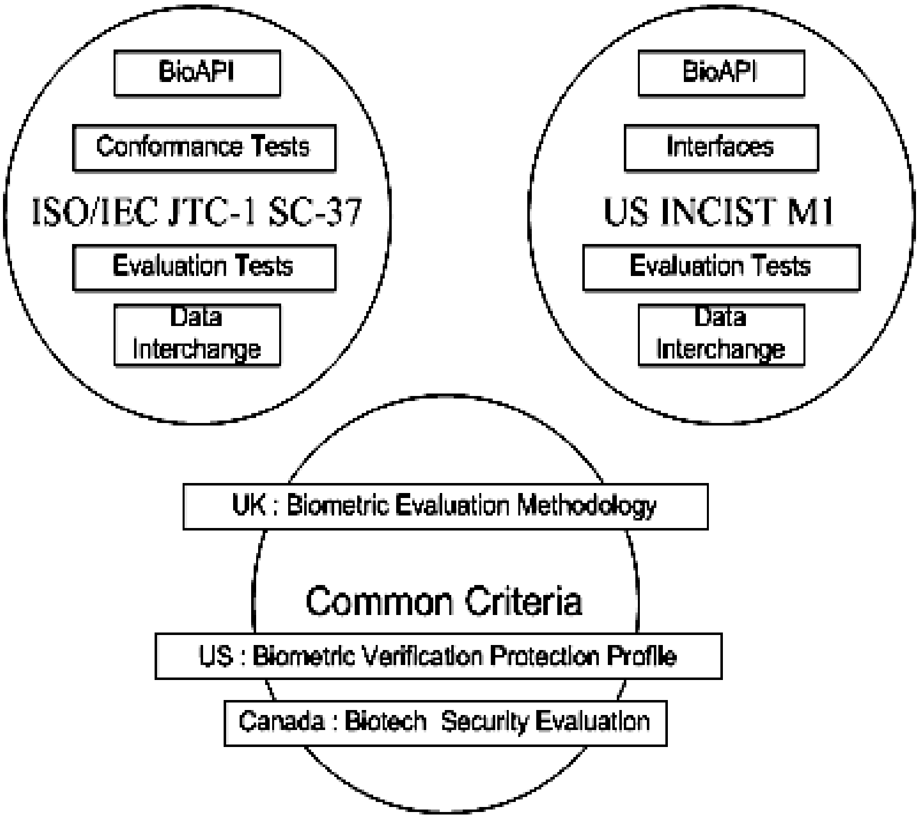


Figure 4-2 Summary of Biometric Standardisation Bodies and Published Standards

While all those documents and standards are valuable in their discipline, none of them defines a set of security evaluation tests for voice verification when spoof attacks are taken into consideration. They provide bases for performance evaluations but specific categories of security vulnerabilities for different biometrics still call for further investigation and standardization. In short a series of tests and a guideline for interpretation of their results for evaluation of security of voice verification against spoof attacks and reliability of its decisions in adverse conditions in the same context does not exist.

We start by a rough classification of the types of attacks and organizing the previous studies according to that classification. Then we proceed with the study of British version of Common Criteria Biometric Evaluation Methodology in the development of a voice verification security/reliability evaluation framework.

4.5 Previous Research on Vulnerabilities of Speaker Recognition to Spoof Attacks

4.5.1 Overview of the Types of Spoofing Algorithms

While there is no comprehensive and unitary research addressing all the aspects of voice verification vulnerabilities, many valuable works have covered some of the relevant threats in a piecemeal fashion.

Finding the categories of vulnerabilities could be based on matching the voice as a biometric with an already available standard on all biometrics (e.g. BEM) and extracting the voice-specific items, examination of case-studies and previous research results and/or a logical reasoning for classification of types of attacks. All these three will be carried out in this chapter for identification of types of spoof attacks.

To classify types of speech manipulation/spoofing a simple criterion is suggested here: speech is either produced by a human or a machine and it is either altered by a human or a machine. One of the four classes in which speech is produced by machine but is altered by human, is not tenable; therefore three types of attack on a voice verification system remain and will be studied here: Speech Synthesis (machine-machine), Voice Conversion (human-machine) and Mimicry (human-human). This is depicted in Figure 4-3.

The boundary line between conversion and synthesis is blurred. Since both methods use machine computation for generating desired data, many methods could equally be called conversion or synthesis. The difference intended here mostly pertains to the application of the method. For speech synthesis, the assumption is that the impostor inputs the desired phrase to a machine (for example, through a keyboard) and the machine based on the previously trained models and samples generates articulates the phrase artificially. For conversion, the same models and data could be available but the input is the true voice of an impostor which is modified to reflect the statistical characteristics of the target voice.

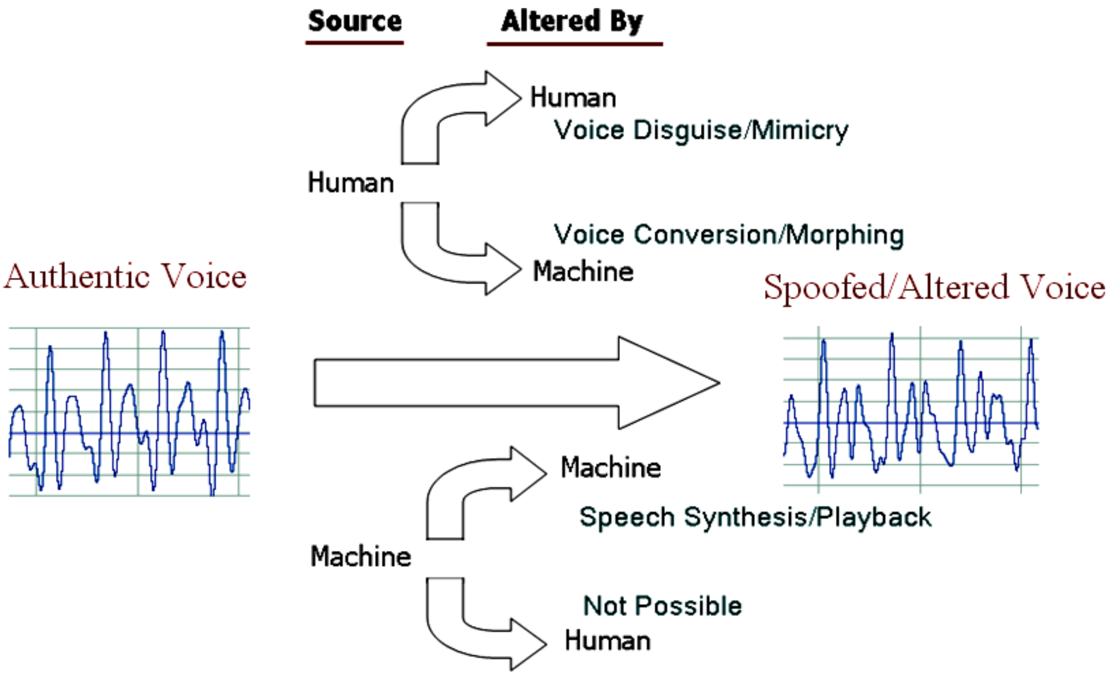


Figure 4-3 Classification of Possible Attacks Based on Source and Alteration Instrument

4.5.2. Speech Synthesis

Speech synthesis is the artificial production of human speech. Methods of speech synthesis can be divided into two classes: those based on human voice samples as raw input for concatenative machine processing (for example, triphone concatenation); *versus* those which create the speech based on pure mathematical methods such as formant synthesis. When the speech units or vocal characteristics for speech synthesis are obtained covertly from a target speaker, the synthesised voice will pose a significant threat to a voice verification system.

The equal error rate of a speaker recognition system based on hidden Markov models of spoken digits rises from 1.1% for baseline natural speech to 3.4% for formant synthesis and 27.3% for speech synthesis by concatenation of whole words (Lindberg and Blomberg 1999, male speakers). Their two other synthesis methods based on pitch and formant tracking didn't yield good results.

Masuko, et al. (1999) developed a synthesis system based on regenerative hidden Markov models (HMMs) using sampled MFCC coefficients from Mel Log Spectral Approximation (MLSA) filters. The reported results show that an HMM based speaker recognition system with the baseline EER of 0% against human mimicry suffers a false acceptance rate of over 70%

when attacked by the synthesised speech. They carried out further experiments trying to make use of samples' pitch to reject the synthesised voice. These attempts reduced the EER but were not completely successful especially when synthesis was based on pitch (Masuko, et al. 2000). Pellom and Hansen. (1999) used a synthesis method based on trajectory modeling of speech and Line Spectral Frequency (LSF) parameters for voice changing. They tested their algorithm on a voice verification system based on Gaussian Mixture Models (GMMs). The false acceptance rate was increased from baseline EER of 1.45% to 86.1% for morphed speech. They have also reported that the intelligibility of their speech morphed to human listeners has been 99.5%. Their results are alarming, even after noting that they didn't test that the content verification of their phrases and they used a large portion of common data in model-training and test samples.

4.5.3. Voice Conversion

Voice conversion or morphing is the process of transformation of an impostor's voice to that of a target victim's voice. It can be harmful to the security of voice-based authentication. Most feature extraction methods are based on the frequency components and a dynamic filter which allows transfer from the impostor's speech spectra to the victim's speech spectra.

The research carried out in this field does not necessarily target automatic speaker identification systems. Nevertheless the results and achievements reported indicate serious potential threats to the security of voice verification. Examples are Orphanidou, Moroz, and Roberts (2004) for wavelet-based voice morphing , Boccardi, and Drioli (2001) for morphing by Gaussian mixture models , Sundermann et al. (2006) on voice conversion by unit selection based on Euclidean distance of frames[1] and Shuang, Bakis, and Qin (2006) on formant mapping.

4.5.4. Mimicry

Despite common belief, spoofing by mimicry is not necessarily the greatest threat to voice verification systems. Nevertheless voice verification has shown a degree of vulnerability to mimicry even when performed by inexperienced impostors. Lau, Wagner, and Tran (2004) conducted a number of mimicry experiments by asking two amateur impostors to log-in into a voice enabled authentication system. At highest they reported false acceptance rate of up to 35%

[1] Despite the title of conversion the unit selection method is closer to speech synthesis but can be realized in both conditions from application point of view.

at EER threshold for a system with baseline equal error rate of zero. However, in their experiments at the security inclined thresholds and for some speakers as the target speaker the impostors couldn't break into the system. They regarded speakers with similar voices as 'close' and concluded that "a normal person can get a high chance to attack the system if that person knows the closest speaker... If that person does not know who is closest…that person can attack the system by using each speaker name at a time to log on" (p. 148). They repeated similar experiments in 2005 (Lau et al. 2005) with 6 impostors two of whom were professional imitators (linguists). They found the closest speaker in the database to each impostor based on GMM models. Their results show that, while both groups have high chance of deceiving the system for a voice close in nature to their own, the linguists were more successful for mimicking the person whose voice was not close to theirs. This however needs further investigation since the underlying notions of closeness of GMM parameters and closeness in professional human judgement of voice similarity are at some variance.

Based on GMM parameters,Farrus et al. (2008) showed that prosodic features except for the range of fundamental frequency are vulnerable to spoofing and reported that the identification error for fused features increased from 5% to 22% for imposture.

Masthoff (1996) classified forms of voice disguise but did not test mimicked voices against a voice verification system. Further research of this type can contribute to development of mimicry detection modules in verification systems.Among the recent works Perrto et al. (2007) studied four types of disguise which received higher number of answers when people were asked how they would change voices: hand over the mouth, a high pitch, a low pitch, and a pinched nostrils voice. The results (although based on small databases) showed that first the human can detect that the voice is disguised and even the type of disguise. In the worst case however (pinched nostrils) 19% of people decided that the voice was normal. The automatic detection based on MFCC features had worse results and was not successful for many cases especially considering the fact that it was based on 20 seconds of speech.

In a different study Kajarekar et al. (2006) tried to determine how much the false rejection rates for automatic and human recognition increases if voice disguise happens. This is applicable to the scenarios in which users try not to be recognised. They demonstrated that around 40% of the speakers can deceive the system by changing their voices if the system is trained without disguised voices.

It could be noted that based on the different experiments and for different applications four distinct questions are needed to be addressed for spoofing by conversion, synthesis or mimicry:

1 - Could the disguised/spoofed voice be detected by humans?

2 - Could the disguised/spoofed voice (truly belonging to the speaker) be verified by machine (forced authentication)?

3 - Could the disguised/spoofed voice (from an impostor trying to log into the system) be rejected by machine?

4 - Could the disguised/spoofed voice be detected as disguised/spoofed by machine (regardless of verification results)?

Depending on the type of application for which voice verification (forensic, commercial security, etc.) is being used and the verification conditions (over distance, supervised, unsupervised, etc.) the answers to these questions have important implications.

4.6 Analysis of Vulnerabilities Based on BEM

Common Criteria Biometric Evaluation Methodology (BEM) identifies 15 points of vulnerability in biometric verification systems, all could easily be shown in the Figure 4-1 [1].

It is noteworthy that a distinction should be made between system-level vulnerabilities and algorithm-level vulnerabilities. Algorithm-level vulnerabilities are related to the core-verification module and are those which exist even when the connections are secure and database contents are reliable – for instance, when there aren’t any weak or fraudulent models stored in the database. System-level vulnerabilities on the other hand are created by an insecure connection to the database, change of data before reaching the verification module or misuse of the system at the time of training or verification.

To focus on the idea of customising the BEM evaluation scheme for voice verification, we could consider the questions:

‘Which classes of threat are specific to voice verification?’ and

‘If the verification unit were merely a static module with fixed database (such as those used for password checking), what classes of vulnerabilities would disappear?’

[1] These categories are: User Threats, User/Capture Threats, Capture/Evaluation Threats, Extraction/Comparison Threats during Enrolment and Verification, Extraction/Template Storage Threats, Template Storage Threats, Template Retrieval Threats, Administrator Threats, User/Policy Management Threats, Policy Management Threats, Threats to Portal, Portal Threats, Threats to all Hardware Components, Threats to all Software Components, Threats to all connections. See BEM for more details

These rather generic questions could be asked for any other biometric identifier needing a spoof evaluation scheme.

The following focused and voice biometric-specific list of threats emerges from our extension of BEM:

- User Threat: Authorised user knowingly and provides speech to impostors either willingly (collusion) or unwillingly (coercion) (mainly applicable to fixed-phrase speaker recognition systems)
- User Threat: Impostor covertly captures voice data from authorised user (applicable to fixed-phrase speaker recognition systems)
- User/Capture Threat: Impostor presents own voice or modifies own voice (*mimicry*) to impersonate (a) a randomly selected user (b) a selected indistinctive model (c) a user with similar voice
- User/Capture Threat: Impostor presents an artificial voice sample (*synthesis*) to impersonate (a) an indistinctive model (b) a target user
- User/Capture Threat: Impostor presents a noisy, poor quality or null voice recording in an effort to match an indistinctive model or regular model (voice or sound synthesis)
- Extraction/Model Storage Threat during Enrolment: Authorised user presents a noisy, poor quality, highly varying, artificial, modified or null voice record to create an indistinctive model in the database

Another type of attack, the ‘relay attack’, could be added to the list in which an attacker relays his voice (by mobile or similar device) to imply that he is at an expected place. This type of attack can be used in applications such as electronic monitoring of offenders by random telephone calls.

Figure 4-4 illustrates the classes of threats including spoof attacks extended after the review of BEM.

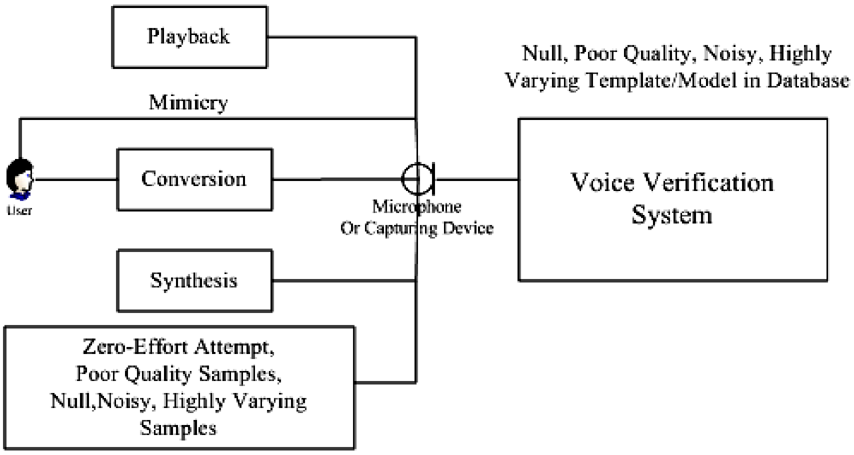

Playback
Mimicry
Conversion
User
Synthesis
Zero-Effort Attempt, Poor Quality Samples, Null,Noisy, Highly Varying Samples
Microphone Or Capturing Device
Null, Poor Quality, Noisy, Highly Varying Template/Model in Database
Voice Verification System

Figure 4-4 Summary of threats against the voice verification process

4.7 Categories of Threats and List of Possible Attacks

Table 4-1[1] categorises all the types of vulnerability from the previous section, [2] and specifies known attacks under each category. Major attacks against which the system should be tested and vulnerability categories include-but are not limited to-those listed.

Most of the vulnerabilities and attacks are related to algorithm-vulnerabilities. In addition, some of them – especially in the category of artificial/irregular voice playback – depend on a flaw the database contents. Any new record added to the database (speaker models) should be tested to make sure that it does not create an opportunity to break into system by means of irregular input signals. A number of such irregular records should be at disposal of the security evaluation system (in a separate subset of irregular signals). Replay attack threat can be removed by means of varying pass phrases for text-prompting speaker verification. Relay attack is one of the categories which show that vulnerabilities may arise due to the wrong expectation from the system.

[1] Based on the review of previous works disguise techniques (not specifically appeared in the table) include: accent change, hand over the mouth, low pitch, high pitch, slow, whisper, nasal speaking, clenched teeth, pinched nostrils, tongue out, hoarse sound, murmuring.

[2] The reasoning presented in section 4 shows that there are only three classes of attacks and even artificial voice (e.g. white noise or zero length voice) generation techniques could be counted as forms of speech synthesis. On the contrary the list of variations of attacks would never cease to expand.

Table 4-1 Classes and Variations of Known Threats to Voice Verification Systems

Category	Method	Reference
Artificial/Irregular Voice Playback	Highly Varying Sample Playback (White-noise)	Not Available (yet requires simple considerations)
	Zero Length Sample Playback	Not Available (yet requires simple considerations)
	Weak model/Template Targeting	Close to Lau et al. (2005)
	Empty Content (Silence) Playing	Not Available (yet requires simple considerations)
Conversion	GMM based Linear Transform	Boccardi and Drioli (2001), Ye and Young (2003) and (2004)
	GMM Based Conversion Based on Gradient Accent	This work
	Wavelet (DWT) Based Sound Morphing	Orphanidou et al. (2004)
	Morphing by Spectral Envelope Transformation	Orphanidou et al. (2003)
	Conversion Based on Formant Mapping	Shuang et al. (2006)
	Audio Morphing by Pitch, Spectrogram and MFCCs	Slaney et al. 1996
Conversion/ Synthesis	Unit Selection Based on Euclidean Distance on MFCCs	Sundermann et al. (2006)
	Formant and Pitch Tracking and Resynthesis/ or Conversion	Lindberg and Blomberg (1999)
Mimicry	Verification Evasion by Disguise	Kajarekar et al. (2006)
	Mimicry by Gifted People	
	Mimicry Against Closest Speaker by Nature	Lau et al. (2004, 2005)
	Random Mimicry Trying	Farrus et al. (2008), Perrto et al. (2007)
Synthesis	Concatenation of Digits of Raw Parts of Speech	Lindberg and Blomberg (1999)
	HMM based Speech Synthesis with MLSA filter	Masuko, et al. (1999)
	HMM based Segment Selection	This work
	Trajectory Modeling of Speech on Line Spectral Frequency (LSF) Parameters	Pellom and Hansen. (1999)
Playback Attack	Same / Similar / Different Phrase Playback	Not Available (yet requires simple considerations)
Relay Threat	Using any Mechanism for Voice Transfer	Not Available (yet requires simple considerations)

Whereas random calls to the offender's place of curfew in passive systems[1] (Richardson, 2002) has been considered as an alternative to tag-based monitoring (see e.g. Aungles and Cook (2004) and Mair (2005)) and such calls have been used manually in home detention programs (Baumer et al. 1993), with the advances in the methods of transferring the speech signal, the assumption that the voice is being made at the intended place is now disputable. The offenders can use a handheld device to return the calls with the requested information while they are not actually at the designated place. This type of vulnerability arises not because of a flaw in the voice verification system but as a result of having wrong expectation from the system.

4.8 Formal Specification of Evaluation System and Architectural Notes

Although several formal specification languages could be used to describe system resistance to attack in a global way, and ultimately deliver formal proof of a specified level of resistance, here the view is taken that the system resistance has a more architectural than logico-mathematical aspect. The approach that is opted is specifying the architectural aspect by setting up a 'Spoof Attacks Layer' to proxy how a real attacker might compromise the voice verification process. The relation of this layer to the evaluation core functions is shown graphically in Fig 4-5. It supports the following semi-formal representation of the spoofing process, and communicates with the core Controller Layer via an Experiment Manager class. The Experiment Manager class dispatches to the Spoof Attacks class a sample of genuine records of speech. To each genuine 'record' (speech sample $R_{Genuine}$), the Spoofer class applies a spoof algorithm to produce a 'counterfeit' voice record R_X aimed at spoofing a target speaker's voice. The Experiment Manager class uses R_X as input to the verification system to calculate a new false acceptance score as follows:

1. Using its model of the target speaker's voice, the verification system assigns a score S to the record using a verification process f

Equation 4-1

$$S = f_M(R_X), \qquad where \qquad M = u(R_{Tr})$$

[1] As opposed to use of active system in which a personal identification device is attached to the wrinkle or wrist of offender which sends electronic signals which are collected by a home monitoring unit.

in which S is the verification score, M is the target speaker's model parameter set obtained by a modeling process u from training records R_{Tr} of the target speaker's voice.

2. A spoofer module from the Spoofer class has used its own modeling process v and training data R_{spoof} to make the counterfeit voices (R_X)

Equation 4-2

$$R_X = h_\Lambda(R_{Genuine}), \quad where \quad \Lambda = v(R_{spoof})$$

in which h denotes the spoofing algorithm, Λ is the spoofing model extracted by process v from speaker data (R_{spoof}) in the spoofing module.

v represent the model training function.

The benefits of specification of these modules include: precise evaluation of vulnerability of a specific verification process to a spoofing technique, via an experimental test of the parametric modeling which defines the secrecy levels described in sub-section 4-9; and the possibility of automation of the experimental evaluation.

In addition, the semiformal approach facilitates the translation into a software architecture of the requirements outlined in section 4-3 – using well known design patterns (Gamma et al., 1995) of software engineering , to arrange, for example:

1. Independence and Loose Coupling (through Adapter/Proxy Pattern)
2. Accommodating New Threats (through Strategy Pattern)

The first requirement of section 4-3 is for decoupling the voice verification system from the evaluation system. The combination of proxy pattern (for providing connection to the remote voice verification system) and adapter pattern can serve this purpose. Adapter patterns allow converting the interface of a class into another interface that the client expects.

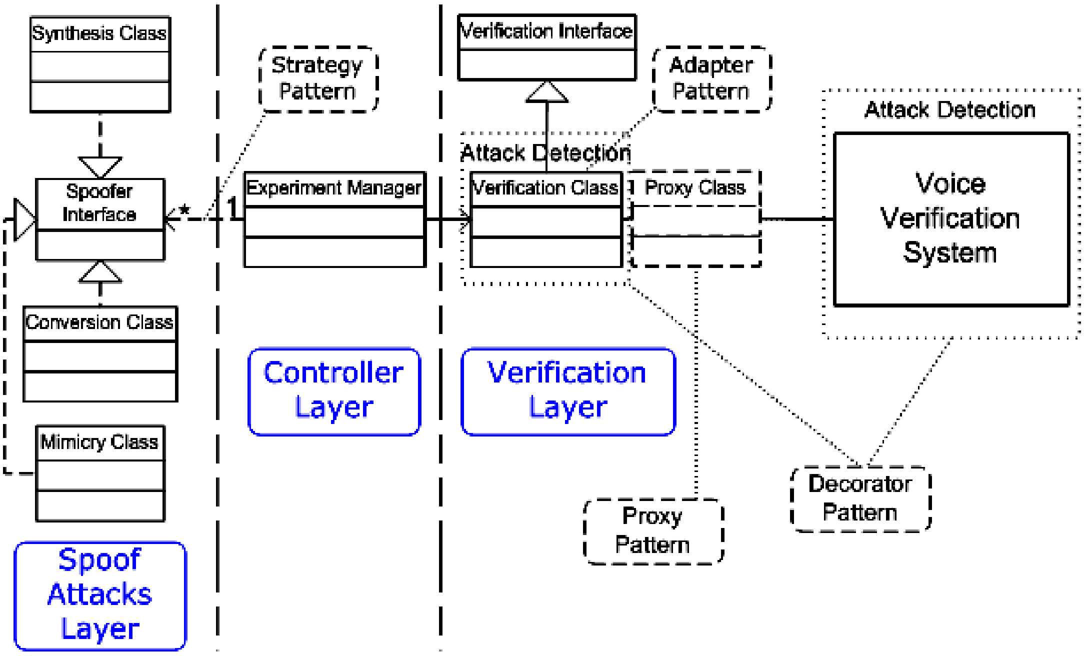


Figure 4-5 Software architecture proposed for automating security

The requirement of accommodating new threats is covered strategically because "Strategy pattern defines a family of algorithms, encapsulates each one, and makes them interchangeable. Strategy lets the algorithm vary independently from clients that use it" (Jones, 2007). In the architecture of Figure 4-5, 'Experiment Manager' class will have a copy of each spoofer class. Each spoofer class represents one instance of an attack and complies with the interface definition made by the spoofer interface. This interface sets a standard for the implementation of spoofer classes, allowing implementation by different parties. As soon as a new threat is discovered, a spoofer class which simulates this attack is developed and a copy of that class is registered with the 'Experiment Manager' without any obligation to change the evaluation code.

Further design patterns that are of interest include:

3. Effortless Change in Design and Outputs of Experiments;

4. Implementing and Testing Defence Mechanisms (through Decorator Pattern)

The first is achieved through separating the spoof attacks layer from the verification layer. The verification layer plays the role of the interface between the 'Experiment Manager' class and any voice verification system while the spoof attacks layer allows various possible experimental design expedients in the sampling of possible attacks. In the second case, Decorator Pattern is used to facilitate implementation of spoof detection blocks whose role is to reject counterfeit records. The pattern allows safe augmentation or extension of the behavior of the verification core system. The original detection object is wrapped in a decorator that implements the same interface but adds extra functionalities to the behaviour of the object. This is the recommended way for adding security checks to the services before and after the main functionality of the service (Jones, 2007). No change is necessary in the verification or client's system as the functionality for spoof detection is augmented by the rejection of counterfeit records.

4.9 Reporting on Performance and on the Security Evaluation Process

4.9.1 Parts of a Report

This section deals with three crucial topics: the methods used for evaluation; what the evaluation reports consist of; and how the results should be interpreted.

The controller layer should evaluate the vulnerabilities and calculate the error rates for each of the threats under analysis. Therefore an evaluation report provides test scores for all the categories of threat.

Each of the next sub-sections sheds light on one aspect of the compiling of test scores.

4.9.2 Assumed Secrecy Level and Type of Data Available to Intruders

An important factor affecting the evaluation results is the degree of knowledge impostors have about the system. In general, the security evaluations are either black-box or white-box. Black-box tests do not explicitly use knowledge of the internal structure and algorithms while in white-box tests some knowledge of the system is incorporated into the test plan. The systems which pass only black-box tests usually rely on security through obscurity. The evaluation results should be interpreted based on the adopted secrecy level. We can define 5 types of data (ToD) available from a voice verification system[1]:

ToD1: The same data used for training the models is available to intruders (R_{spoof} has some overlap with R_{Tr} in Equation 4-1 and Equation 4-2)

ToD2: Speaker models stored in the database are available to intruders (M in Equation 4-1 is available to intruders)

[1] There are two independent factors which could be used for defining secrecy level: the data from the speaker, and the data from the system. Several other combinations could be imagined but the ones listed above are more practical. In case of an unlisted assumption about the knowledge about the system it could be directly specified.

ToD3: Knowledge of both algorithms and the algorithms' parameters are available to intruders (such as the number of states in hidden Markov models, number of MFCC filter-banks, structure and number of layers in neural networks, etc.)

ToD4: Some samples of the target speaker's voice are available to intruders (but R_{spoof} does not contain data from R_{Tr})

ToD5: General knowledge of the algorithms employed in verification is available to intruders (e.g. whether the neural network, hidden Markov model or template matching technique is used)

4.9.3 Subset Based Evaluation

The view taken in this chapter is that the evaluations should be conducted and reported on a subset basis. The result of the test specifies the false acceptance or false rejection errors (based on the types of subset) occurred when verifying subset *A*. For evaluation of the reliability of different aspects of voice verification system different subsets should be prepared and used, as these encapsulate the experimental design of different evaluation tests.

For analysis of the reliability of voice verification in adverse conditions, the subsets should include: subsets with signal contaminated with environmental noise, subsets with human speech interference (so-called 'babble-noise'), subsets based on data passed through different channels (such as telephone, mobile, different microphones), subsets with encoded/decoded signals, subsets with different styles of speaking and emotional states. A close examination of important factors related to adverse conditions is presented in chapter 7. Each of the above subsets can have several variations which are discussed in chapter 7.

For analysis of the security of voice verification, the several subsets should be tested against all the categories of spoof attack. One problem in the area of speech processing is that the studies mostly report performance optimization for one type of test data. For example an algorithm is shown to improve the verification of a type of noisy signal. At the same time the same algorithm may drastically worsen the verification performance for a variety of channel-distorted or codec-manipulated signals. Unless we can show that the intended type of signal for which the study is conducted is recognizable and it is possible to restrict the use of that method to the specified signal the improvement is not of pragmatic value. That is why for practical applications the comprehensive study of all conditions (reflected in the suggested subset-based approach) is necessary.

4.9.4 Reliability and Performance Measures

Any performance measure related to false acceptance error rates could be used as a partial security evaluation measure, and the set of all such performance measures provides an over-arching security measure when threats are taken into the picture. These error rates should be calculated and reported as part of an evaluation report: False Accept Rate (FAR), False Reject Rate (FRR), and Equal Error Rate (EER). These depend on a decision threshold, and since the thresholds are normally set to the EER threshold in many conditions, the FARs for the counterfeit speech should be calculated at this threshold to measure system vulnerability in face of the different types of spoof attack.

Receiver operating characteristic (ROC) curves are widely used for demonstrating the results of voice verification tests. These plot hit rate, (100-FAR%), against false alarm (FRR%) rates for a given type of spoof as decision threshold is varied as described in Appendix A.

As we will see throughout this Thesis especially in chapter 6 and 8, the difficulty of setting a global threshold for the system is the most crucial obstacle to the practical use of voice verification system. ROC curves, by concealing threshold values, do not overcome this difficulty: two ROC curves could be geometrically congruent, yet require two different set of thresholds to realise a specific error. As illustrated in Figure 4-6, if one of two congruent curves refers to verification in normal conditions and the other to the noisy conditions, the common ROC curve may wrongly imply that the noise has had no effect on the verification, whereas in fact the thresholds needed to make rates equal are very different.

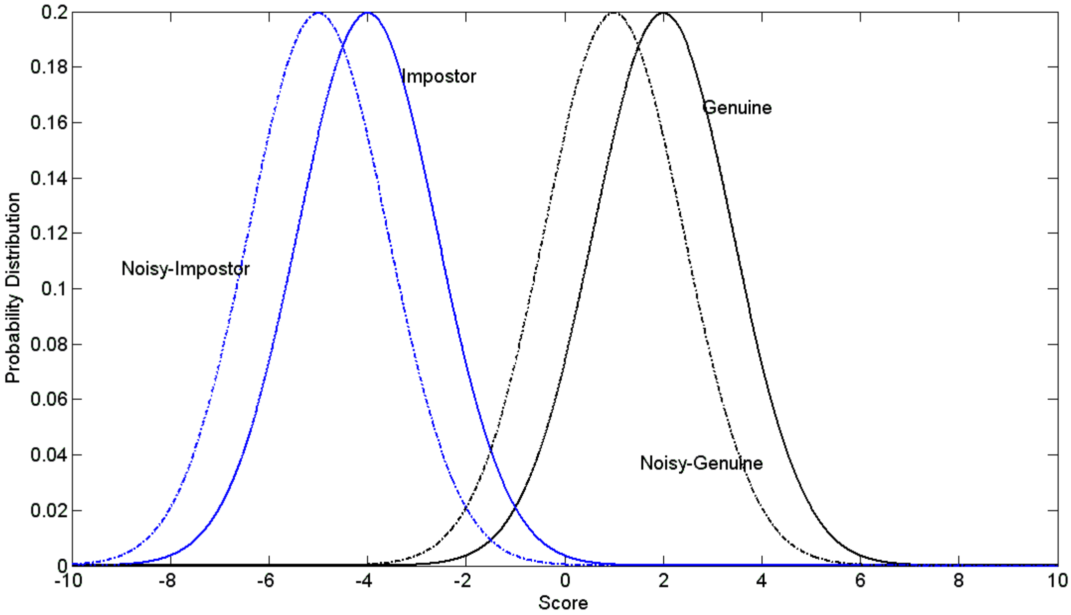


Figure 4-6 Two Different Pairs of PDFs with the Same Resulting ROC Curves

A decision threshold could favour system security (high FAR) or user convenience (high FRR). A way of reporting the error rates for the system under attack that makes this trade-off clear is to use the three thresholds:

- Security Inclined Threshold: Where FRR=a.FAR[1]
- Equal Error Rate Threshold: Where FAR=FRR
- Convenience Inclined Threshold: Where FRR=1/a.FAR

When the verification is based on the results of two or more corpus-trained speech-recognition modules (say HMM or GMM), each module would have its own modifiable threshold. In this case, it is recommended that the above three thresholds for each of the speech-recognition modules be reported.

With the same corpus-based approach, the reliability tests will be conducted in the same way that the security tests are carried out and the same form of error reporting is recommended. It is only the test subsets that differentiate the reliability analysis (chapter 7) from security analysis (chapters 6 and 8).

Although reporting the results of each category of attack is not the aim of this chapter a hypothetical plot for demonstration purposes is shown Figure 4-7. The baseline EER of the system and the baseline secure point are shown in the plot. The severity of the threat could be

[1] The factor of a is determined by the designer of the evaluation system. A factor of 3 is used in this thesis.

recognised by assuming that the system threshold is set at its previous value, for example at the EER threshold.

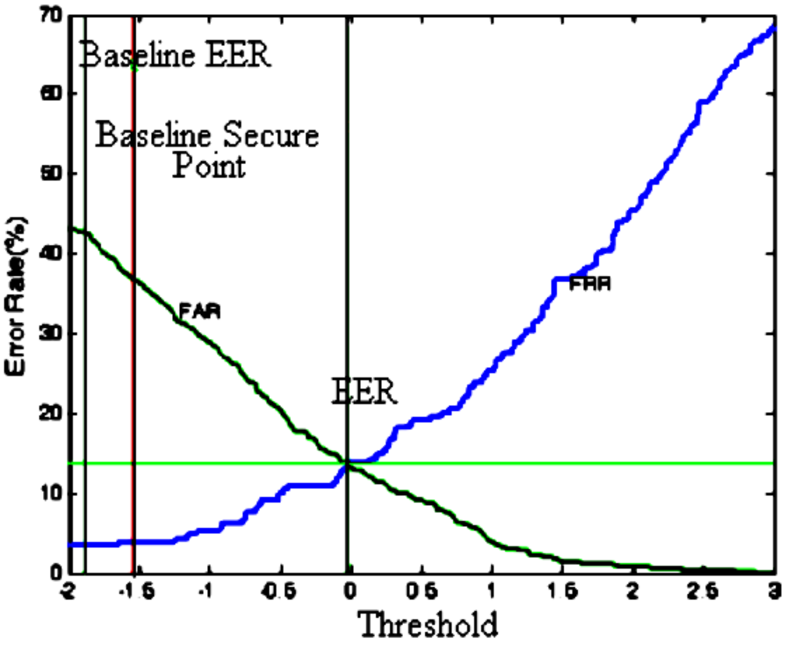


Figure 4-7 Illustration of a Hypothetical Security Evaluation Plot

4.9.5. An Evaluation Checklist

For each type of spoof attack, a reasonably complete evaluation should include the following items to figure in an evaluation checklist:

- Category of spoof attack (Conversion, Mimicry, Synthesis, etc.)
- ToD (i.e. types and amount of data from system or individual needed for realizing the attack)
- Vulnerability (Experimental error rates, reported at different thresholds or through FAR and FRR curves)
- Rate of Detection by Human Supervisors (Subjective Tests of whether a human listener can detect that the signal is counterfeit – has implications for supervised and semi-supervised verification discussed in chapter 9)
- Scenario Implications
  - Remote (Applicable to Phone, Mobile, VOIP) / In Person
  - Supervised (Human Detection, Expert/Inexpert) / Unsupervised
- Automatic Detection of Attack

- Footprints of the spoof attack
    - Signal Analysis, Verification Scores, Speech Anomalies
    - Usage traces: e.g. delay in response
- Rate of spoof detection (elaborated on in chapter 8)
- Time to detect (When will a detection alarm be issued?)

4.9.6 Notes on the Interpretation of the Results

The key point in interpretation of the evaluation results (when reported per subsets) is that the overall system error is decided by the probability or frequency of the input signals in each subset. In the case of FAR of the system at a given global threshold will be:

Equation 4-3

$$FAR_{th} = \sum_i FAR_{th}^i . P_i$$

where FAR_{th}^i is the FAR for that particular subset *i* used in validation experiments at the specified threshold *th* and P_i is the probability that impostors will employ signals similar to those used in validation subset *i* .

Similarly, in the case of FRR:

Equation 4-4

$$FRR_{th} = \sum_i FRR_{th}^i . P_i'$$

where FRR_{th}^i is the FRR for the particular subset of genuine users used in validation experiments at the specified threshold and P_i' is the probability that a genuine user will employ signals similar to those in the validation subset *i*.

In both cases, the estimated probabilities should sum to 1, but in the case of equation 4-3 the frequencies are to be obtained from user logs of the system, whereas in the case of equation 4-4 the frequencies are to be obtained from criminological data.

4.10 Spoof and Attack Detection Blocks

The vulnerability results reported in this Chapter as well as those which will be presented in the entire Thesis especially in Chapter 6, demonstrate the necessity of using spoof detection modules along with the verification system. As illustrated in the Figure 4-5 these blocks act as shields to the main system and should also be evaluated in the same fashion as the verification system.

It is notable that architectural independence of spoof detection and verification does not guarantee statistical independence of detection and verification decisions. While the thorough discussion has appeared in chapter 8 it could be said that three types of reports for success of spoof detection blocks could be suggested in this framework:

1. Global false positive error and false negative errors at certain thresholds (such as equal error rates at false alarm of 5%): These are only helpful approximate hints for how in reality the spoof detection block will work. The real combined result is only determined (due to the correlation between decisions) when the subsets are tested with the verification system and spoof detection block working together and at the thresholds set for both (expanded upon in chapter 8).
2. False negative and false positive errors as a function of the score given by the verification system: When this method is chosen the evaluation system can calculate the FAR and FRR of combined systems as specified in chapter 8.
3. Case reports for subsets: This is a helpful report in which the results of decisions or authenticity scores (the likelihood that the recordings are authentic based on the score assigned by the detection module) are assigned by the spoof detection blocks and given to the evaluation system. It can be followed by a 'test' by the evaluation system to determine the FAR and FRR rates when the verification system and the detection modules are active.

Spoof detection blocks and the possibility of spoof detection are examined in depth in Chapter 8 in which a demonstration of how these tests can be carried out is presented.

4.11 Evaluation for Forensic Speaker Recognition

While most of the above points apply to forensic speaker verification types and subsets used for evaluation of reliability and security of voice verification in forensic applications are different from those used for security-based speaker verification.

In security-based speaker verification the study of ‘population’ is intended. It is possible that, within the population, certain individuals are not affected by the adverse conditions or spoofing techniques to the same degree as others. When false rejection rate of a system is 5% it does not necessarily mean that each individual is rejected in 5% of his attempts. Also when we study the effect of mismatched conditions this effect is not equal for every individual[1].

Figure 4-8 shows that speakers have their own score distribution functions which altogether shape the genuine (or impostor) curve for the population. In forensic speaker verification the tall (user specific) PDFs are used as well as impostor distribution.

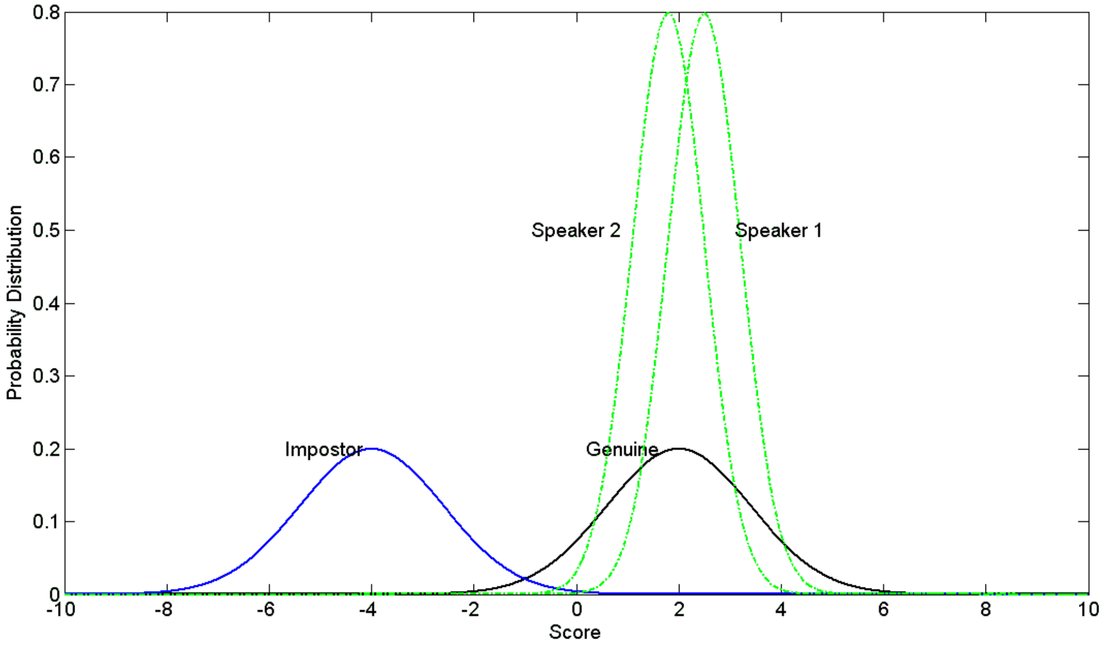


Figure 4-8 Comparison of Speaker-Specific and Population-Specific PDFs

For forensic speaker verification therefore we need a control database to evaluate the within-speaker variability as explained by Gonzalez-Rodriguez et al. (2006).

The rest of the evaluation process, regardless of the method used for assigning scores to the utterances in the forensic speaker verification, is very similar to security-oriented speaker verification.

For security-based speaker verification the sampling subsets that are needed are:

- *P* (speaker): Subsets consisting of speech from genuine speakers for training models. For noisy speech, channel affected speech, code affected speech, style-specific speech, etc. we need different subsets of *P* for training robust models. These categories are extensively detailed in Chapter 7.

[1] Zero normalization (Z-Norm) method introduced in chapter 3 aims to reduce this variation within each individual and bring the individual curves closer to the population curve. Test normalization (T-Norm) on the other hand tries to reduce the mismatch effect.

- *G* (global): Subsets consisting of speech data from impostors. This data is used for training the cohort models if the cohort normalization is used or for training global/world models otherwise.
- *C* (control): If normalization per speaker is applied for each user we need another subset to calculate the mean and standard deviation of scores obtained by user (for Z-Norm). This is where we try to remove the within speaker variation and bring the PDF curves for all users close to each other and to the population.
- TP (Test, Speaker): For test purposes *TP* includes test phrases which are going to be verified (speech data from genuine speakers in all abovementioned situations)
- *TG* (Test, Impostor): data from impostors.

Speech spoofing techniques may have their own subsets to train their models (detailed in chapter 5 and implemented in chapter 6).

In forensic speaker verification, *P* subset only consists of data from one speaker. All the other subsets are the same and having a control subset becomes necessary, as suggested by Gonzalez Rodriguez (2006).

Due to the importance of forensic speaker verification, statistical tests should be performed in order to demonstrate the reliability of decisions. Although these tests are applicable to the area of security-oriented speaker verification their use is not as vital.

In forensic speaker verification, as described in chapter 3, a likelihood ratio is assigned to the evidence which is:

Equation 4-5

$$LR = \frac{P(E \mid H_1)}{P(E \mid H_2)}$$

Where *E* is the evidence (speech here) H_1 is the hypothesis that the evidence is from the person under suspicion H_2 is the hypothesis that it is from someone else in the population. Regardless of how the *P(E|H)* is calculated (acoustic features and models, linguistic features and models, etc.) it is extremely important that this value is approximated on a reliable basis and is not biased by sampling artefacts.

For this reason it is recommended here that several control and global subsets ($C_1, C_2, ..., C_N, G_1, G_2, ..., G_N$) should be available and tested. The goal of tests is to prove that the values assigned to a piece of evidence is reliable.

Common statistical investigations such as T-test, F-Test, ANOVA and KS-test could be suggested here[1].

If the statistical tests reveal that two (two or more) distributions are not equal at a significance level it is fair to specify a range for LR values based on all subsets as illustrated in Figure 4-9. In this figure it is demonstrated that for two subsets the approximated distribution functions are different and based on the selection of each pair of *C* and *G* subsets we can calculate a different likelihood ratio. Therefore for all 4 specified points and for their corresponding curves, likelihood ratios should be calculated and a range for it should be reported. In this case the likelihood ratio is reported as a range such as: $|LR| + / - \delta_{LR}$ where $|LR|$ and δ_{LR} are mean and standard deviations of likelihood ratios.

Another hidden assumption behind reporting the likelihood ratios is that there is no mismatch between test utterance (evidence) and the control subsets. As we will largely investigate and observe in Chapter 7 there are numerous types of acoustic mismatch conditions which exist and each has a tremendous impact on the distributions. The results of the statistical tests (on a variety of *C* subsets) should EITHER show that the assigned probability values are reliable for a range of acoustic variations (which should include the conditions under which the utterance under investigation is collected) OR there should be good reasons to believe that there is no considerable acoustic mismatch between the *C* and the speech under investigation (test subset or evidence). Without either of two assumptions we can not make any judgment on the derived LR value (which is calculated for a different condition).

[1] 1. Under the assumptions of independence, equal standard deviation values and normal distribution of scores, we can conduct two-sample location-test (t-test) to show that means of each two normally distributed populations are equal (Miller and Freund, 1965).
2. Under the same assumptions of independence and normal distribution of scores, we can conduct a two sample F-test (Anderson et al. 1994) to show that the standard deviations of two populations are equal.
3. Without any assumption about the distribution of scores, we can carry out a two sample Kolmogorov–Smirnov test (KS-test) to show that the two populations have the same distributions. KS-tests are non parametric tests for differences between two cumulative distribution functions (Miller and Freund, 1965). The tables and methods for KS-tests could be found in the handbooks of statistics such as Beyer (1968).
For k population in addition to pair tests we can perform Analysis of Variance (ANOVA) tests on the k population means (Anderson et al. 1994).

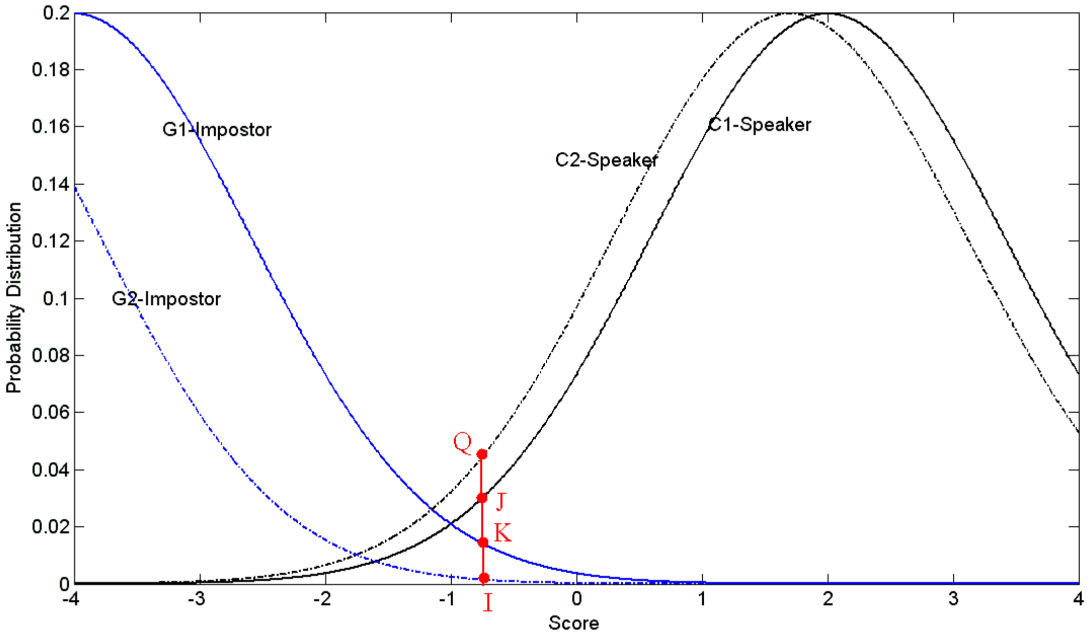


Figure 4-9 Illustration of the Possible Changes in the Forensic Speaker Verification Likelihood Ratio from Different Estimations

4.12 Broad summary and its bearing on subsequent chapters

In this chapter the importance of security evaluation of voice verification system was emphasised and it was shown that none of the current standards can address all the security concerns in this area, especially as emerging spoofing techniques emerge in the arms race between criminals and defenders of the system.

We also defined the broad requirements for a dynamic evaluation framework that is independent of the type of verification system. These requirements led us to consider a generic voice verification evaluation structure, enabling easy experimentation with new forms of attack and allowing empirical testing of new defensive countermeasures.

In addition to the security evaluation, with the 'subset based approach' proposed in this chapter and in the same architecture (involving the independence of the controller layer) the effect of different types of degradation in the quality of speech signal ,for example, by noise and channel is also opened up to empirical testing.

With these possibilities, any 'claims' reliability and robustness in face of spoofing techniques could be 'proved' by conducting above-mentioned tests and by showing that the system withstands the impostor attacks introduced and classified in this chapter .

The evaluation, however, is not a one-time process and the system should have pre-defined mechanisms for identifying and coping with new security threats. These call for continual monitoring of security breaches and assessment of reliability of defence mechanisms. The results should be added to the evaluation system as new spoof modules or as input data for increasing coverage and assurance of decisions made by detection modules.

In addition, it was emphasised that since employing spoof detection blocks and reliability improvement suggestions has consequences for other conditions for which they are not tested, the comprehensive evaluation of 'security' and 'performance' of voice verification (in various conditions) together is essential.

The rest of this Thesis deals with more detailed analysis of each element introduced in this broad framework.

In chapter 5 the main voice verification system will be developed and it will be shown that the system is robust in normal conditions and has very low error rates.

In chapter 6 two new spoofing algorithms will be implemented and the security of the developed voice verification system against these exemplar spoofing techniques will be gauged.

Chapter 7 reports the results of a comprehensive study on the effect of various adverse conditions on the reliability of speaker verification decisions.

Chapter 8 elaborates on the development; evaluation and use of separate sub-systems for detection of synthesised and morphed voice.

With all these details results in place, we can conclude on the design and requirements of voice verification systems for different areas of application (Chapter 9).

Chapter 5 : Implementation of Voice-Verification System and Baseline Experiments

5.1 Purpose of This Chapter

This chapter describes the development of the prototype voice verification system implemented for experimental components of the thesis. It precedes the experiments exposing this system to the security threats and adverse conditions, which are dealt with in the next chapters. The present chapter also aims at justifying the choice of parameters and architectural issues presented in the evaluation framework in chapter 4, rationalizing the choices made, and describing the conditions under which the results are interpretable. Hence a significant portion of this chapter is devoted to the discussion of robustness, components and parameters of an automatic voice verification system.

The detailed list of topics and objectives covered in this chapter are as follows:

- Introducing the central voice-verification and evaluation system, its components and their linkages;
- Presenting the two corpora used in this research: IVIE and Chain;
- Justifying the design of samples from these corpora for reliability and vulnerability analyses;
- Introducing the adjustable parameters in a voice verification system, optimizing the parameters and justifying the choices;
- Exploring the nature of automatic speaker verification by examining the number of clusters in a person's speech as part of parameter optimization;
- Discussing the need for and demonstrating the development of a silence-removal component in the voice verification system.

5.2 Implementation of Verification and Evaluation Prototype in MATLAB and Java

The automatic voice verification system and all the core algorithms introduced later in the thesis are implemented in MATLAB. Nevertheless, in order to demonstrate the software architectural concepts put forward in chapter 4 a prototype java application is also implemented which acts as a shield over the MATLAB core functions.

For hidden Markov modeling and Gaussian mixture modeling, the Bayes Net Toolbox for MATLAB by Murphy (2004) is widely used in the thesis. For calling MATLAB functions from java the JMATLINK package by Müller (2003) is used[1]. Experimentation with cepstral coefficients and new features in later chapters is enabled by extracting them using the VOICE-BOX toolbox by Brookes (1997).

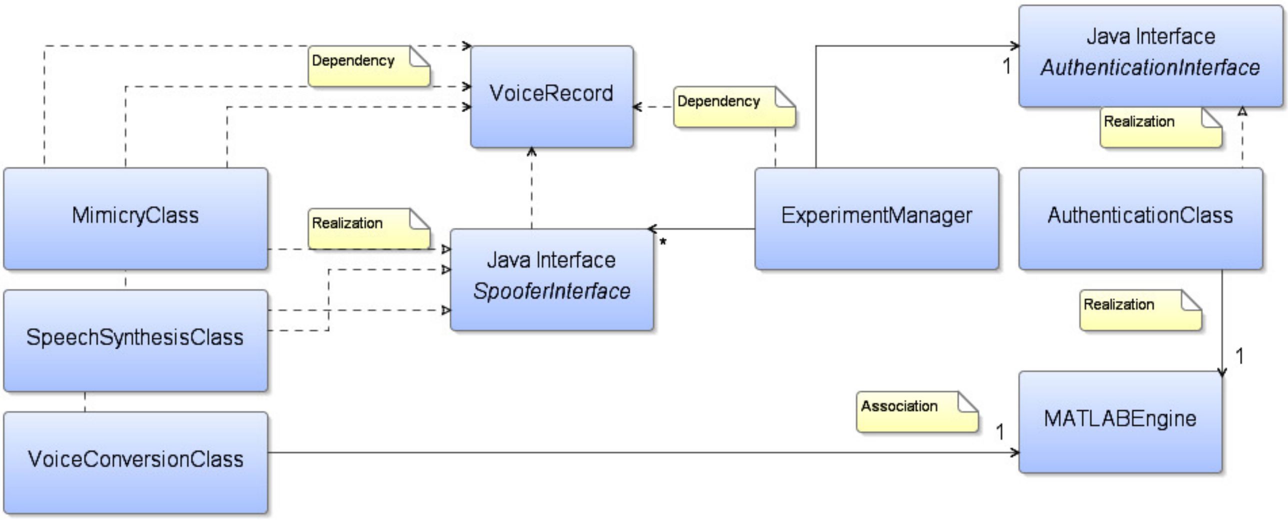


Figure 5-1 Class diagram of the implemented voice verification and evaluation system

Fig. 5-1 illustrates the elements of java implementation of the voice verification and security evaluation system in a class diagram. It is produced automatically from the application's code. There are two main interfaces in this diagram which define the input/output requirements and the methods that should be implemented: the *AuthenticationInterface* which defines the implementation rules of the authentication system; and the *SpooferInterface* which defines the input-output interface of spoofing modules, which may be of any type; mimicry, voice conversion or synthesis. The *ExperimentManager* class has a copy of these two types of interface and knows nothing about the internal structure of the classes which implement these interfaces. This class, and the objects instantiated from it, dispatch the voice records to the spoofer modules and send the spoofed records to the authentication system for verification. The *ExperimentManager* is responsible for recording the error rates, plotting the diagrams and

[1] This was done with the goal of the implementation of a stand-alone evaluation system. Due to the availability of better plotting tools in MATLAB the experiments were carried out and the plots were produced in that environment.

making evaluation reports. Since MATLAB provides better visual and reporting tools the same reporting processes are also reproduced in MATLAB.

5.3 Characteristics of the IVIE Corpus and design of experiments

The IVIE corpus contains 36 hours of speech data from 110 speakers covering nine dialects of speech from the British Isles (Grabe et al. 2001). The original recordings in the corpus fall into five categories: controlled sentences, read passages, retold passages, map tasks, and free conversations. Dialect varieties in the corpus include English accents from: Belfast, Bradford, Cardiff, Cambridge, Dublin, Leeds, Liverpool, London and Newcastle.

We focused on controlled sentences for training and test. There are either 21 or 22 controlled sentences for each of 110 donors in the corpus, including examples of the following types: 'S' = statement, 'Q' = question, 'W' = WH-question, 'I' = inversion questions, 'C' = coordination[1].

Figure 5-2 below summarises the design of training and test samples. For GMM training sentences recorded from 70 speakers have been selected (Subset A) and for verification a similar selection has been made (Subset B). The sampling of the speakers has been performed so that speakers from all dialects are almost equally present in both subsets. Six other sentences (Subset C) from the same 70 speakers are used for the GMM-based voice conversion experiments presented in chapter 7. It is assumed for the spoof experiments that conversion or synthesis modules do not have access to the same data upon which training models are built. With this in mind, subset C has no overlap with subsets A and B. For speech synthesis spoofing experiments, trained generic HMM models are needed to generate a prompted sentence, and these need training data. The data was selected as 3 sentences from another 20 speakers that are not in the training or test samples (Subset D). Note that the selected sentences are the same as test sentences but uttered by different speakers, reflecting a real-world scenario in which the perpetrators of spoofing attacks have access to the voices of accomplices but not the voices of authentic speakers. For training Global GMM Models, 8 sentences from 20 speakers were selected (Subset E), and for training HMM models, that are needed in content verification, a further 3 sentences from the same 20 speakers were selected (Subset F).

[1] An Example of the statements is ‘They are on the railings’; An example of the inversion questions is ‘May I lean on the railings?’; An example of the coordination sentences is ‘Are you growing limes or lemons?’.

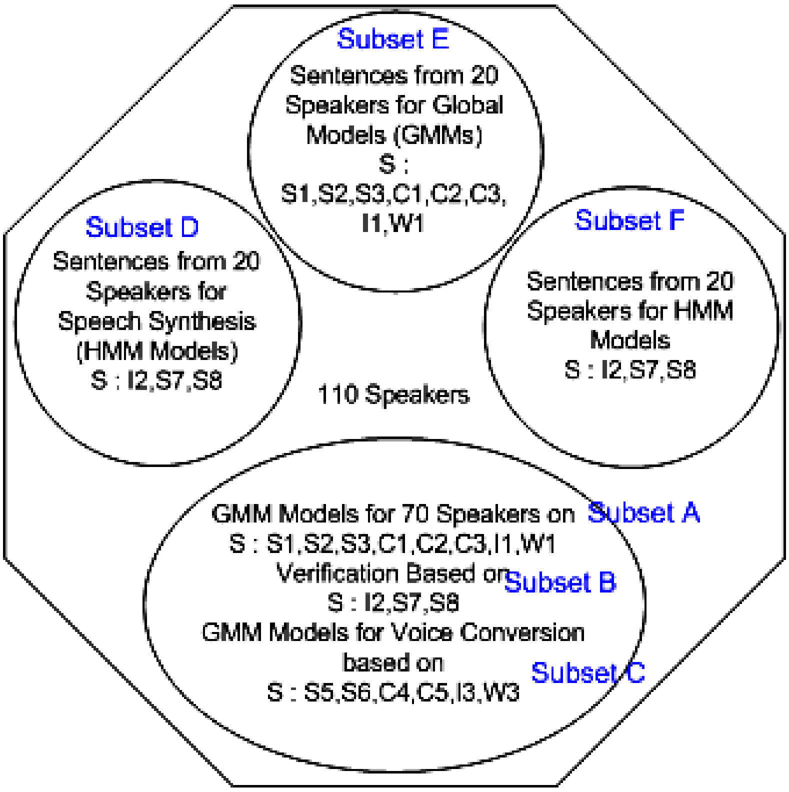

Subset E
Sentences from 20 Speakers for Global Models (GMMs)
S : S1,S2,S3,C1,C2,C3, I1,W1
Subset D
Sentences from 20 Speakers for Speech Synthesis (HMM Models)
S : I2,S7,S8
Subset F
Sentences from 20 Speakers for HMM Models
S : I2,S7,S8
110 Speakers
GMM Models for 70 Speakers on
Subset A
S : S1,S2,S3,C1,C2,C3,I1,W1
Verification Based on
Subset B
S : I2,S7,S8
GMM Models for Voice Conversion based on
S : S5,S6,C4,C5,I3,W3
Subset C

Figure 5-2 Classification of Speakers and Sentences Used for Verification and Spoof Attacks

Additional experimentation on the effect on speaker verification of change of register from reading to retelling and speech pauses is reported in chapter 7. Samples for these experiments were selected in the form of:

- Read Passages (Passage of Cinderella story read by 70 speakers, Subset G);
- Retold (Cinderella story retold spontaneously from memory by 70 speakers, Subset H);
- Hand-Trimmed Retold (A modification of Subset H from which silences and pauses have been removed manually, 70 speakers, Subset I).

5.4 Cross-Validation of Results with Chain Corpus

Despite the advantages of the IVIE corpus (variety of dialects and conditions of recording), the corpus has two drawbacks for speaker verification. The first one is the lack of diversity in speakers' age (all speakers are adolescents) and the second is the fact that all recordings are made in one session.

A more recent corpus, the CHAINS corpus, is concerned with intra-speaker variation, and contains the recordings of 36 speakers obtained in two different sessions with a time separation of about two months. According to the team of collectors at the University College Dublin, the design goal of the corpus has been providing a range of speaking styles and voice modifications for speakers sharing the same accent. Other existing corpora, in particular the CSLU Speaker Identification Corpus, the TIMIT corpus, and the IVIE corpus have served as referents in the selection of material (Cummins, et al., 2006). Across the two CHAINS sessions, recordings are collected in six different speaking styles:

- Solo reading
- Synchronous reading (read passages)
- Spontaneous speech (indicated in the followings as Retold)
- Repetitive Synchronous Imitation
- Whispered speech reading
- Fast speech reading

Only Solo, Fast, Read and Spontaneous (Retold) portions of the corpus are used here. The Solo, Read (synchronous) and Spontaneous recordings are recorded in the first session and the Fast recordings are recorded in the second one (which despite the intention of collectors makes it difficult to separate the effect of inter-session variability from conditions of speaking).

The original sample rate of recordings in Chain corpus was 44.100 kHz. The data was re-sampled to 16 kHz to allow comparability with IVIE data and used in speaker verification experiments. In this research, the solo sentences are divided into two parts: one for training (Solo Subset) and the other for test (Test Subset). Sentences from 30 speakers are used for verification and five sentences from another six persons are used for training the global model (Global Subset). Four fast recordings from 30 speakers (the same speakers that appear in the test subset) are included in the 'Fast Subset' (three sentences and one read passage) for the analysis of the effect of fast speaking on the verification results. The Retold Subset' contains one recording for each of the 30 speakers who re-tell the Cinderella story in their own narrative style. Similarly, the 'Read Subset' contains one recording from 30 speakers reading a paragraph of the Cinderella story. Figure 5-3 illustrates and summarises this information.

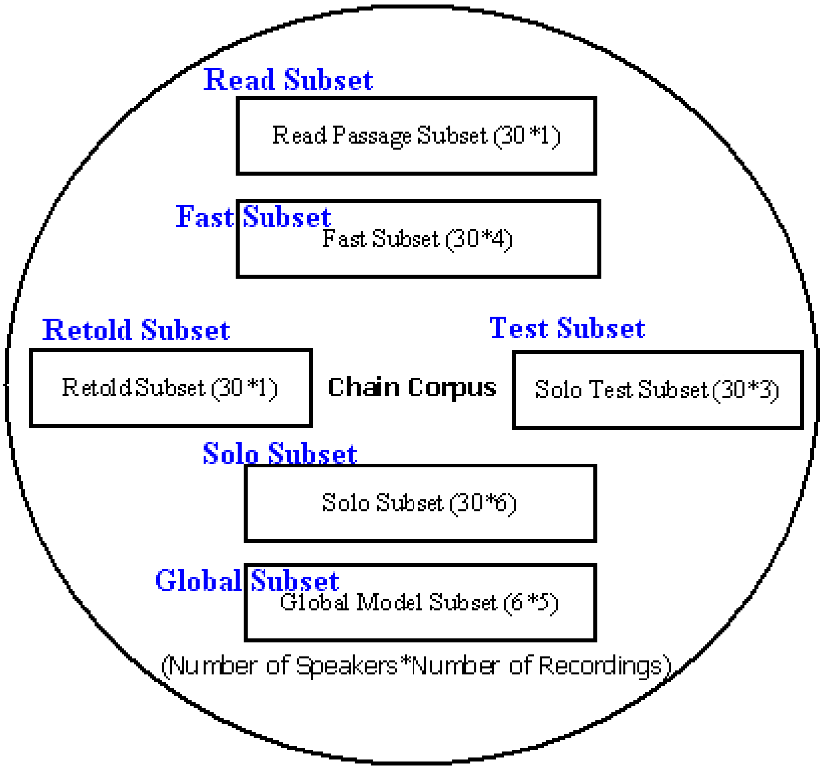


Figure 5-3 Classification of Chain Corpus Speakers and Sentences for Speaker Verification

5.5 Components of the Prototype Speaker Verification System

A person's voice holds constant stochastic characteristics over a long period of time regardless of the contents of speech, which makes it suitable for biometric identification. The possibility of prompting for a new phrase in each verification attempt gives the voice an unparalleled advantage over the other biometrics. The most promising design for a voice verification system, therefore, consists of two separate blocks: One for verification of speech contents and the other for matching characteristics of speech against the speaker's statistical model as depicted in Figure 5-4. In the voice verification system implemented in this thesis, the speaker verification decision is based on scoring a piece of speech by 'user' and 'global/world' models, both trained on enrollment data. A decision threshold is set to separate the user output from the global output. In contrast, content verification (when applied) is based on the decision process involving hidden Markov models as described later in this chapter. A silence removal function removes pauses in speech, which, as discussed in the next chapters, can affect the robustness of verification. A verification attempt is successful if both speaker and content verification steps are successful.

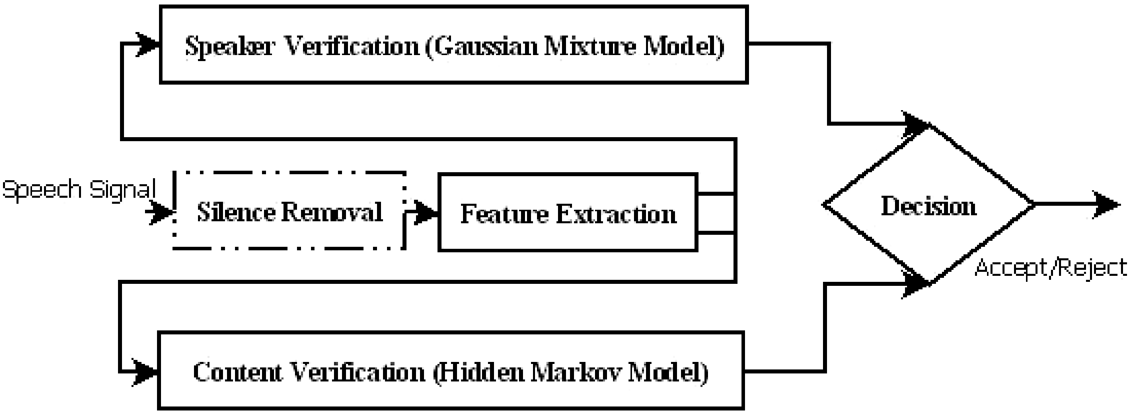

Speaker Verification (Gaussian Mixture Model)
Speech Signal
Silence Removal
Feature Extraction
Decision
Accept/Reject
Content Verification (Hidden Markov Model)

Figure 5-4 Elements of the Voice Verification System

5.6 Variations of Voice Verification Systems and Adjustable Parameters

The number of choices for the structural decisions and parameters of a voice verification system is vast. A thorough, yet pragmatic, investigation of all possible variations is likely to yield a list similar to the following:

- Variations in pre-processing filters such as the use of band-pass filters
- Variations in the feature extraction techniques: Cepstral, Linear Prediction Code (LPC), Chaotic Features, Wavelet (Packet decomposition, Continuous or Discrete), Perceptual Linear Predicative (PLP) and Formant Tracking. A comparison of some of these features for text-independent speaker recognition is made in Kinnunen's thesis (2004). Apart from the feature type, the number of features used (for example, the number of spectral coefficients) and inclusion or exclusion of rate of change ('delta coefficients') is another source of variation.
- Subsequent to feature extraction, any of the feature normalisation techniques such as Cepstral Mean Normalization (CMN) or Mean Variance Normalisation (MVN) could be used (see chapter seven).
- The modeling and scoring techniques may include the use of Gaussian mixture models, or of artificial neural networks, or of vector quantization or dynamic time warping. If one uses, as in this thesis, Gaussian mixture modeling, the most common statistical approach for text-independent recognition (Reynolds, 1995), the number of Gaussians in the mixture and

choice of the initial values of Gaussians is another source of variation. Similarly, for content verification, if one uses hidden Markov models the same choices for Gaussian mixtures arise and, in addition, the number of states and transition details is open to choice.

- The inclusion or exclusion of a voice activity detection or silence removal module which will be discussed later in this chapter is the origin of many differences in speaker verification systems.
- Any of the score normalisation techniques may be used for handset and channel compensation such as Z-Norm, T-Norm and HT-Norm (Auckenthaler et al, 2000).

It is noteworthy that there are no theoretical grounds for justifying which of the many possible choices listed above, is superior. The arguments in favour of a particular choice are mostly based on empirical results, which may not necessarily replicate using different datasets. While it is not possible to investigate experimentally the results in all the possible dimensions, there are enough experimental results in the sections that follow to justify the options that have been chosen notwithstanding the fact that the choices are not perfect for all datasets or even for the corpus at hand.

5.7 Elements of the Verification System

5.7.1 Content Verification

The greater emphasis in this thesis is on how speaker recognition and content verification is used mainly in spoof detection experiments to verify whether or not counterfeit speech signals can pass both tests. Therefore, the complexity of this module is less than one that a commercial speech recognition system may use. In a real content verification system the hidden Markov models for any sentence may be built by concatenation of triphones described in chapter 3. In my content verification experiments the hidden Markov models are fitted on the entire sentences and the score of each observation given the HMM of sentence (λ_i) is calculated as defined below:

Equation 5-1

$$Score\ (O \mid sentence = i) = log\ P(O \mid \lambda_i) - \frac{1}{N-1}\sum_{j \neq i} log\ P(O \mid \lambda_j).$$

Where models, λ_i is HMM model for sentence number *i*, *O* is the sequence of observations (in other words a sequence of feature vectors) and *N* is the number of prompt sentences in use for content verification. This score is compared with an adjustable threshold for deciding whether or not the observation is close enough to the prompted sentence. The calculation of $P(O \mid \lambda_i)$ is based on the equations presented in chapter 3 for hidden Markov models.

For training sentence-level HMMs, the sentences are split into overlapping segments, each corresponding to one state in the Markov model. The frames available in each segment are used for K-means clustering and the results of K-means algorithm is used to determine the mean, variance and the weights of Gaussian mixture for each state. The values for initial state transition parameters are chosen randomly. Diagonal covariance matrices are used and the HMM models are all left to right.

5.7.2 Speaker Verification

In speaker verification experiments the score of each observation sequence, given the speaker model (Π_i) is calculated as follows:

Equation 5-2

$$Score(O \mid speaker = i) = \log P(O \mid \Pi_i) - \log P(O \mid \Pi^M)$$

Where Π^M is the world model or global model (also known as the background model)-is a GMM model trained on a large subset of training data. Model Π_i is the Gaussian mixture model trained for speaker number *i*. *O* is the sequence of observations, or in other words a sequence of feature vectors

If this score exceeds the adjustable threshold, the sentence with feature vector sequence *O* is verified as belonging to a user with speaker model Π_i.

Training of Gaussians is based on Expectation Maximization (EM) algorithm. All the Gaussians are diagonal. For setting initial values of the mixture of Gaussians, a modified version of the K-means clustering algorithm is devised, which has the following steps:

1. Normalise training points to make Euclidean distance meaningful

2. Find best clusters according to the minimum squared Euclidean distance criterion in normalised space (MATLAB *kmeans* function is used for this task). Five initial sets of start points are chosen at random and the best of these five after a maximum of 100 iterations of the K-means algorithm is selected. During the algorithm execution, if one cluster becomes empty

then the K-means function creates a new cluster consisting of the one point that is furthest from its centroid.

3. Find the mean, weight and covariance of the original cluster points in the ‘non-normalised’ space for use as initial values of the Gaussian mixture model. The covariance matrices are diagonal and the values on the main diagonal are variances and the other elements of the matrix are zero. The weights of the Gaussians are calculated based on the ratio of the number of the points in their corresponding clusters to the total number of training points for the speaker.

5.8 Searching for Optimum Number of Clusters and Gaussians

One of the immediate questions after choosing Gaussian mixture modeling for verification is ‘How many Gaussian components should be present in the mixture?’. Since the initial values for training Gaussians come from clustering, the challenge is in finding the suitable number of clusters in an individual’s speech data.

Regardless of the problem under investigation, deciding the optimum number of clusters in multi-feature data is a subjective and disputable task. There are many approaches in the literature for determining the number and quality of clusters, some of which are unreliable (Webb, 2002) and all focused on a particular application area for cluster analysis.

Furthermore, in speech analysis the ’How many clusters?’ question takes two distinct forms, each with separate implications:

1. The first form is ‘How many distinct clusters are naturally present in an individual’s feature space?’ and targets one person in isolation, without considering whether other individuals have similarly structured feature spaces. The question can be answered using the tools of cluster analysis, but the issue of whether other speakers have similar feature spaces depends on connecting clusters to generalised phonological categories, itself a major research topic in speech science~~.~~

2. The second form is ‘What is the optimum number of clusters which minimises speaker verification error rate?’, and refers to clusters as a basis for Gaussian mixture modeling, and using distance between a new utterance and a speaker’s previously chosen cluster centers as a similarity measure for verification decisions.

A first attempt to answer the first question was made by Kinnunen et al. (2001). They tried to determine whether speech data is clustered in a space with three cepstral features only. They concluded on the basis of F-tests that there are no separable clusters in such a limited feature space, and that this is an indication that more features should be used.

Cluster analysis techniques are various and diversified (Webb, 2002). According to a comprehensive comparison carried out by Milligan and Cooper (1985) of 30 methods of hierarchical classification, optimization of a metric proposed by Calinski and Harabasz (1974) outperforms the other criteria. The optimization method has since been used in speech science and other contexts, for example cluster analysis of handwritten characters' (Vuori and Laaksonen, 2002). It is therefore used in this thesis, alongside another for the sake of contrast – the stopping rule criterion of Hartigan (1975), which has, for example, also been used for word sense discrimination (Savova, et al., 2006).

Answering the second question, on the other hand, is purely experimental – trying different number of Gaussians and recording error rates. There is little agreement on optimal numbers in the literature. For example, Reynolds has shown that the mixture components around 40-50 are sufficient for separation of 51 speakers at low error rates (1995). Experiments carried out by Tinnuen (2004) on two datasets using vector quantization methods, indicated that the error rates decline as cluster numbers increase, until a limit of 64 is reached, after which error rates increase. Results obtained by Yu et al (1995) on text-independent speaker recognition by hidden Markov models, indicate insensitivity to state-transition parameters and optimality using Gaussian mixtures with 32 components. In contrast, in some verification experiments (Nordström et al. 1998) up to 256 components were used and in others (e.g. Auckenthaler et al, 2000) as many as 1024 components were thought to be optimal.

In the following sections, some new cluster analysis and decision error-rate studies are carried out in an attempt to clarify both questions.

5.9 Cluster Analysis Criteria and Results

5.9.1 Criteria Used for Cluster Analysis

The widely used Calinski-Harabasz (CH) variance ratio (1974), an indicator of between-cluster to within-cluster variances and analogous to the F-statistic in univariate analysis (Savova et al. 2006) is adopted for cluster analysis here. With *K* clusters the ratio is defined as:

Equation 5-3

$$CH(K) = \frac{\sum_{k=1}^{K} n_k sqd(C_k, M) / (K-1) / N}{\sum_{k=1}^{K}\sum_{j=1}^{n_k} sqd(C_k, O_{kj}) / (N-K)}$$

where *sqd(.,.)* is the square of Euclidean distance, C_k is the center of k_{th} cluster, *M* is the mean of all samples, *K* is the number of clusters, *N* is the total number of samples, and n_k represents the number of samples in cluster *k*. O_{kj} is the j_{th} sample in cluster *k*.

The goal in CH analysis is to find the value of *K* which maximises this ratio or in other words:

$$\arg\max_{K} CH(K) \quad .$$

By way of contrast, an approach based on optimizing within-cluster sum of squares (WCSS) Hartigan (1975) may be represented as follows:

Equation 5-4

$$WCSS(K) = \sum_{k=1}^{K}\sum_{j=1}^{n_k} sqd(C_k, O_{kj})$$

$$Hartigan(K) = (N - K - 1)(\frac{WCSS(K) - WCSS(K+1)}{WCSS(K+1)})$$

The Hartigan stopping metric is a measure of how much within-cluster variance is reduced by adding an extra cluster. Adding extra clusters is stopped when *Hartigan(K)* falls below an agreed threshold. The result could be considered as the ‘natural’ number of clusters in the data.

Before clustering speech data, we generate 4 datasets of Gaussian random samples to examine our procedures. The results are plotted in Figure 5-5. The circles depict the center of clusters chosen by K-means algorithm, and the number of clusters is chosen between 2 and 9.

The figure demonstrates that K-means algorithm and CH criterion have worked well for obviously separate and distinguishable clusters.

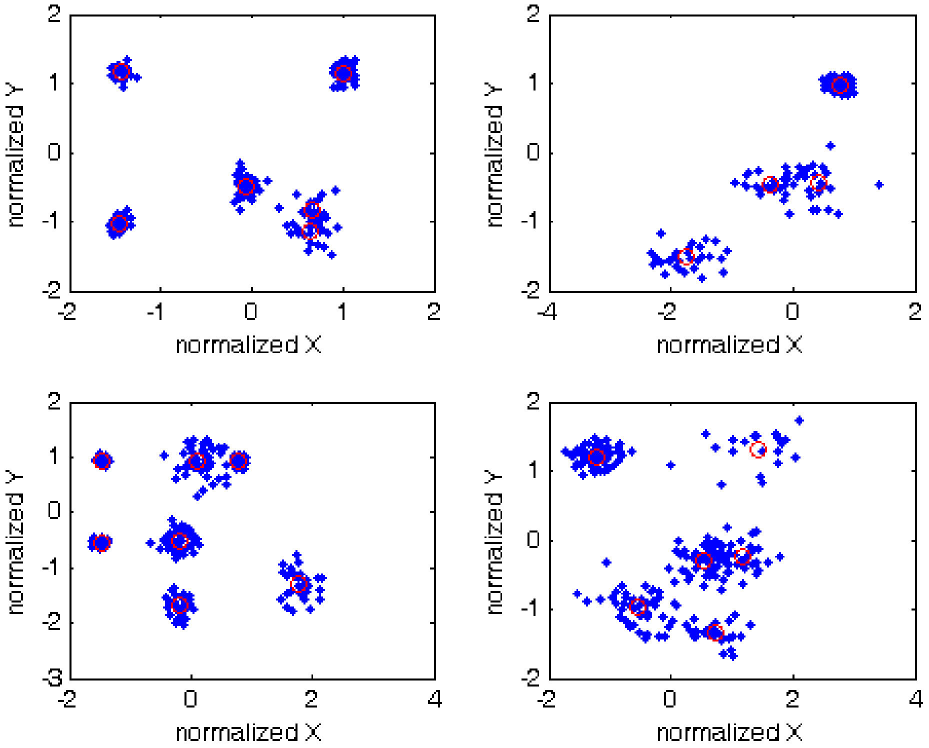


Figure 5-5 Result of K-means clustering and CH analysis on 4 randomly generated datasets

5.9.2 Speech Data Used for Cluster Analysis

For cluster analysis on speech data, 30 speakers from Subset A (Training samples for GMM training), and 8 sentences per each speaker were selected. After Hamming windows were applied to half overlapping frames of 16msec, MFCC features (12 cepstral coefficients) were extracted from sentences. The K-means algorithm was executed on samples of data (normalised to have mean of zero and standard deviation of 1 across all dimensions). This was necessary to ensure distances were equally sensitive to all dimensions of feature space[1].

[1] The distance in this case is called *standardized* Euclidean distance.

To avoid the problem of local minimum, the K-means algorithm was executed 5 times from different initial random samples, and the best clusters (minimum squared Euclidean distance criterion) were chosen. (The built-in MATLAB *K-means* function was used for this task)

5.9.3 Results of Cluster Analysis on IVIE and Chain Corpora

Figure 5-6 shows the average CH ratio for clusters found in the sentences from first 30 speakers of subset-A (IVIE) plotted against the number of clusters (*K*).

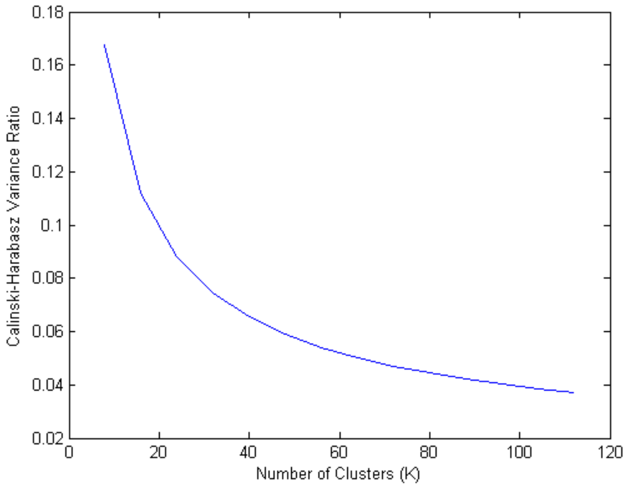


Figure 5-6 K-means clustering and CH analysis on data from 30 speakers, 12 MFCCs, IVIE corpus

The absence of a maximum in the plot suggests that speech samples for each speaker in cepstral space consist of fewer than 8 clusters. Experimentally, this is not true as the classification errors reported later in 5.10 suggest. These perplexing CH analysis results were confirmed by repeating the same procedure on CHAIN corpus (for 12 and also 20 cepstral coefficients). The results for 20 features are plotted in Figure 5-7). The results while according with the findings of Kinnunen et al. (2001) based on other criteria and methods remain perplexing.

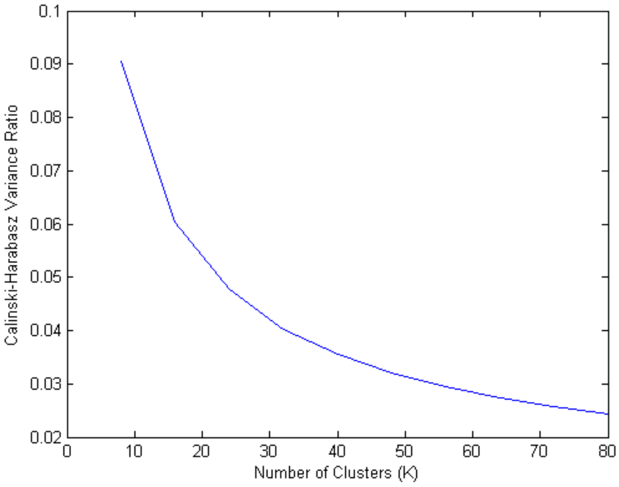


Figure 5-7 K-means clustering and CH analysis on data from 30 speakers, 20 MFCCs, CHAIN corpus

The Hartigan stopping metric, on the other hand, suggests that after about 50 clusters, adding new clusters decrements the WCSS value by about the same fractional amount (see Figures 5-8 and 5-9).

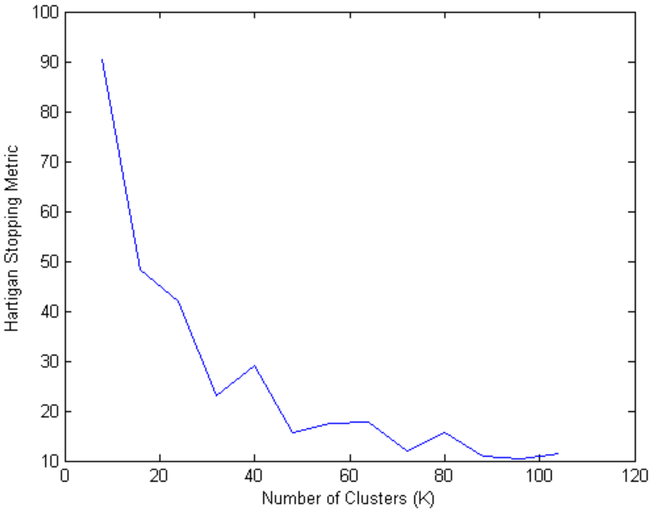


Figure 5-8 Hartigan stopping metric, 12 MFCCs, IVIE corpus

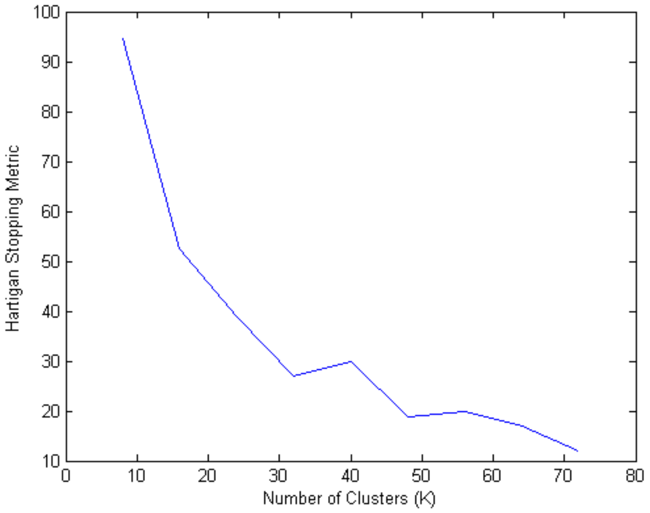


Figure 5-9 Hartigan stopping metric, 12 MFCCs, CHAIN corpus

With the Hartigan 'naturalness' threshold above this amount, the procedure would not stop. There is no fixed way to set the threshold, but these results suggest that it should be set to stop before 50 clusters.

A possible explanation for these puzzling results is indicated in Figure 5-10 showing a large sample of features from one speaker. There is no sign of separable clusters in any dimension, possibly because of the overlap of clusters associated with different phonemes.

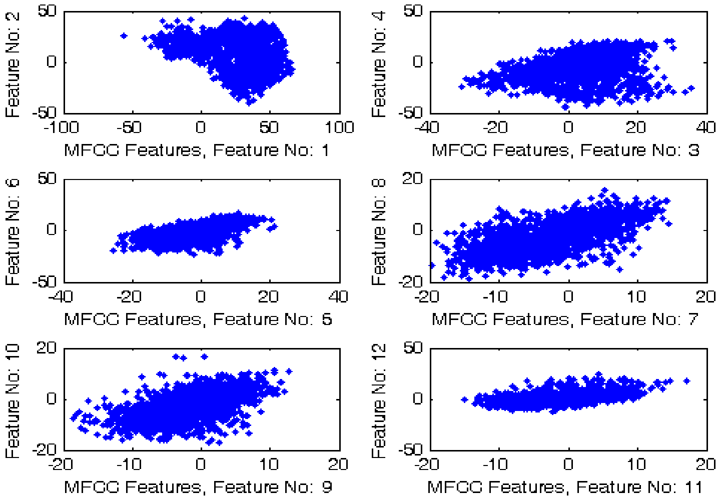


Figure 5-10 Twelve cepstral features extracted for 8 sentences uttered speaker 1

5.9.4 Cluster Analysis based on the Ratio of Correctly Attributed Data

The experimental approach toward deciding the suitable number of clusters in speech data for speaker identification purposes would involve analysis of the percentage of test samples correctly attributed to their corresponding speakers when various numbers of clusters are used. For a feature vector $O_{S,j}$, we define $SP(O_{S,j})$ as the speaker to which this vector is attributed. The attribution is based on the distance to the centres of clusters found for the speakers. First the closest cluster centre is chosen as the cluster which this sample belongs to, and consequently the speaker to which chosen cluster belongs will be $SP(O_{S,j})$.Test samples consisted of 400 samples (MFCC features from 16ms frames) per each of 30 speakers from Subset C. Clustering was performed based on K-means algorithm (with different number of clusters from 8 to 112, by steps of 8). Clustering data was composed of 8 sentences uttered by the same 30 speakers (selected randomly from the 70 speakers available in Subset A). Figure 5-11 shows that the correctly attributed percentage reaches a plateau for about 20 clusters and thereafter fluctuates by about 1 percentage point around the same value for larger number of clusters.

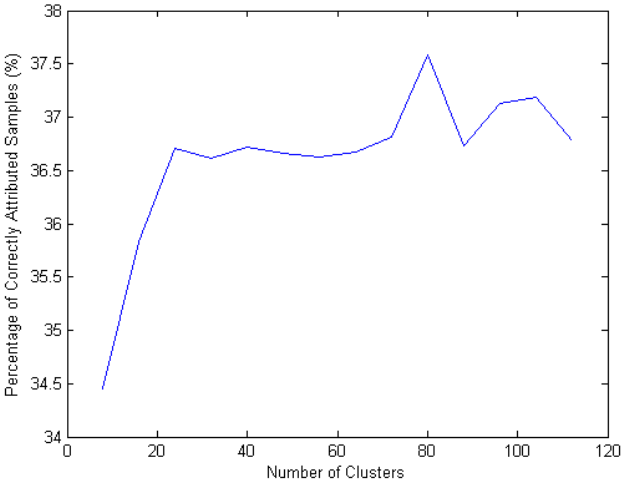


Figure 5-11 Percentage of correctly attributed samples against number of clusters for IVIE corpus, 12 MFCCs, Test Subset: C

Figure 5-12 displays the results of similar experiments on the 'hand trimmed retold' subset. Despite the fact that the percentage of correctly attributed samples is lower, the results show the same qualitative trend for this subset also.

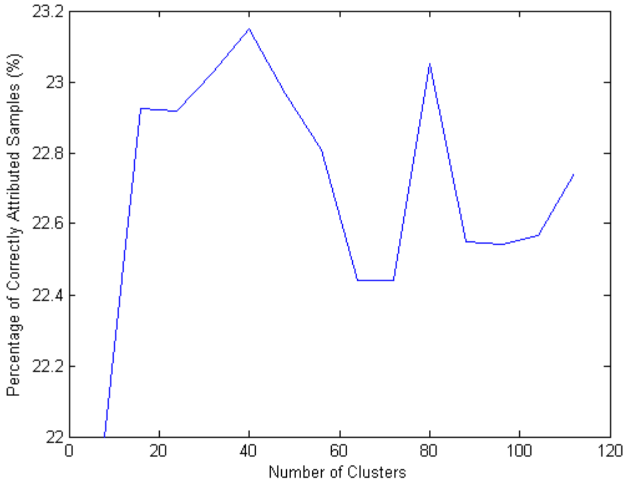


Figure 5-12 Percentage of correctly attributed samples, 12 MFCCs, IVIE corpus, Test Subset: 'Hand Trimmed Retold'

Cross checking the experiments on the CHAIN corpus reveals the same pattern. Figure 5-13 shows the test results for the solo subset of chain corpus when 20 cepstral coefficients were used as features.

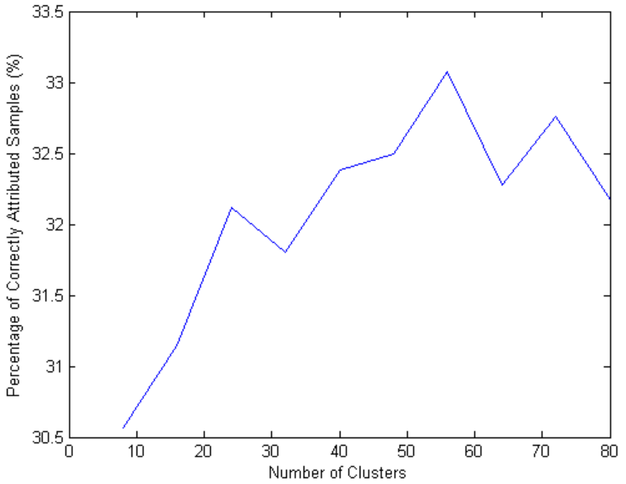


Figure 5-13 Percentage of correctly attributed samples, 20 MFCCs, CHAIN corpus, Test Subset: 'Test Solo'

The reported results here show an upward trend in the percentage of correctly attributed samples followed by a fluctuation which slightly exhibits an increasing moving average until 80 clusters. The fluctuation appears in a small range which is negligible and can be attributed to the bias in the test data. It appears from the plots, that all cluster numbers above 20 have nearly the same discrimination power.

5.10 Deciding Optimum Number of Gaussians and Coefficients Based on Errors

The final step in examining the optimum number of clusters and Gaussian components involves Gaussian mixture modeling (GMM). Equal error rate will be treated as a metric for evaluating suitability of employing various numbers of clusters and consequently Gaussians in speaker models and the global model for verification.

Clusters obtained by K-means algorithm were used as basis for GMM training. More specifically for each user:

1. Twelve (or more) cepstral features were extracted from 8 sentences (uttered by any of 70 speakers in Subset A). Cepstral feature space was normalised along all dimensions (standard deviation=1)
2. K clusters were chosen using K-means algorithm with 5 different starting points to avoid falling into local minima.
3. K independent diagonal Gaussian models were trained over the members of each cluster obtained in step 2 as described in section 5.7.
4. A Gaussian Mixture is trained over the data with K Gaussians. The initial weights were estimated using the proportion of each cluster's samples to the total number of samples.
5. EM algorithm was used to re-estimate the parameters of Gaussians in 20 iterations.

Speaker verification was carried out on Subset B. Three sentences from each of 70 speakers were used for genuine attempts (total 210 genuine attempts). For each speaker 3 imposture attempts using sentences belonging to other speakers in Subset B was made (total 630 imposture attempts).

Seven sets of experiments were designed to shed light on different aspects of parameter selection and model training.

Experiment Set 1: Various Number of Mixture Components

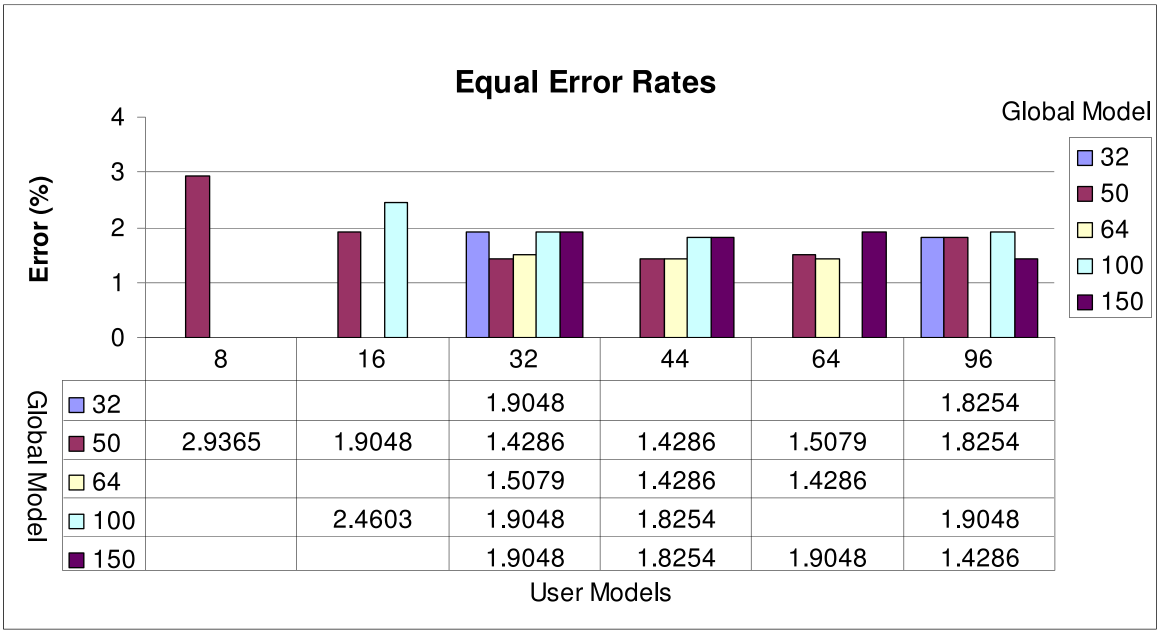
In the first set of experiments, Gaussian mixtures with various numbers of components were trained over the user data (Subset A) and global model data (Subset E)[1]. The number of MFCC coefficients used in this set of experiments was 12. The results show that Gaussian components

[1] The amount of global model training data in all this particular set is one third of Subset E. The global models are trained with one out of three consecutive frames from global model feature vectors which still maintain the generality of data and provide enough training data.

around 50 provide satisfactory results and adding components to the mixture both for user models and global model does not significantly reduce the error rates.

Experiment Set 2: Various Model Sizes and Feature Vector Sizes

In the second set of experiments error rates of verification by various models of different sizes and different numbers of cepstral coefficients are examined. Two types of errors have been used as indicators of performance: Equal error rate (EER) and false rejection rate at secure point where the false acceptance rate is one third of the false rejection rate (FRR@SP). The focus has been on the combination of 32 and 64 Gaussians in the mixtures. It could be observed that increasing the number of coefficients up to 28 improves the verification results (for 32 Gaussians in GMM mixtures) however for 32 cepstrals, since the number of MFCC filters has changed from 29 to 33[1] the error rates indicates degradation in verification performance.



Global Model	8	16	32	44	64	96
□ 32			1.9048			1.8254
■ 50	2.9365	1.9048	1.4286	1.4286	1.5079	1.8254
□ 64			1.5079	1.4286	1.4286	
□ 100		2.4603	1.9048	1.8254		1.9048
■ 150			1.9048	1.8254	1.9048	1.4286

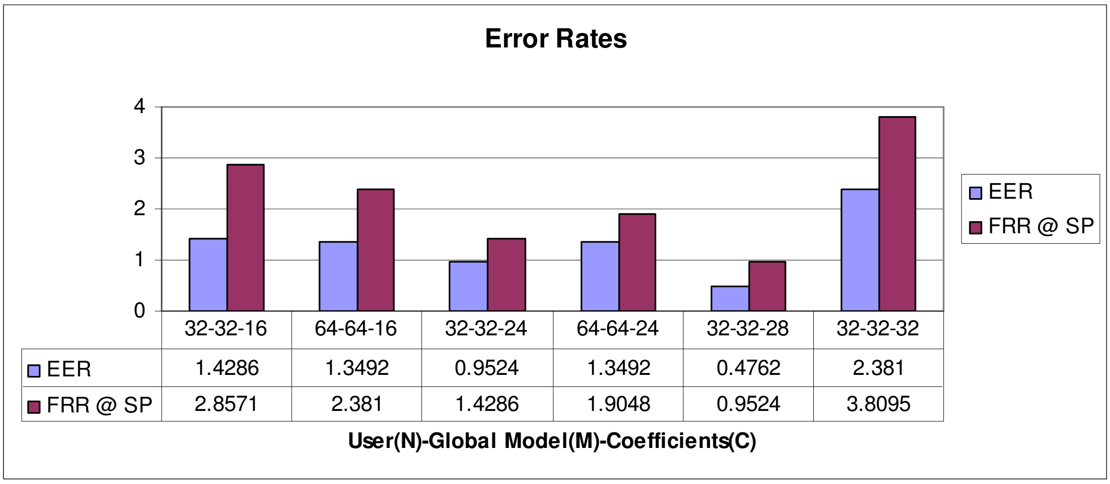
Figure 5-14 Error rates for various model sizes, 12 MFCCs, when one third of data (subset E) was used for global model training

Comparing the error rates in the middle of bar chart in Figure 5-15 exhibits that 32 and 64 mixture components have similar results and even 32 Gaussians in the mixture produces better

[1] The number of cepstral coefficients should be less than the number of filters which discrete cosine transform is applied to their output.

results. Comparison of groups two and four (from left) on the bar chart shows that the same pattern of slight reduction in error rates with an increase in the number of cepstral coefficients exists for 64 Gaussians as well.

In Figure 5-16, the outcome of four experiments carried out on two combinations of features and mixture sizes is depicted. In one group of experiments all the feature vectors in subset E are used for global model training and in the other, similar to Experiment set 1, one third of features are employed. The difference in the error results is insignificant and shows that enough training data has been available in both cases. In one experiment (64-64-24) the results of using one third of the data is even slightly better.



	32-32-16	64-64-16	32-32-24	64-64-24	32-32-28	32-32-32
🟪 EER	1.4286	1.3492	0.9524	1.3492	0.4762	2.381
🟥 FRR @ SP	2.8571	2.381	1.4286	1.9048	0.9524	3.8095

Figure 5-15 Error rates for various model sizes and cepstral coefficients

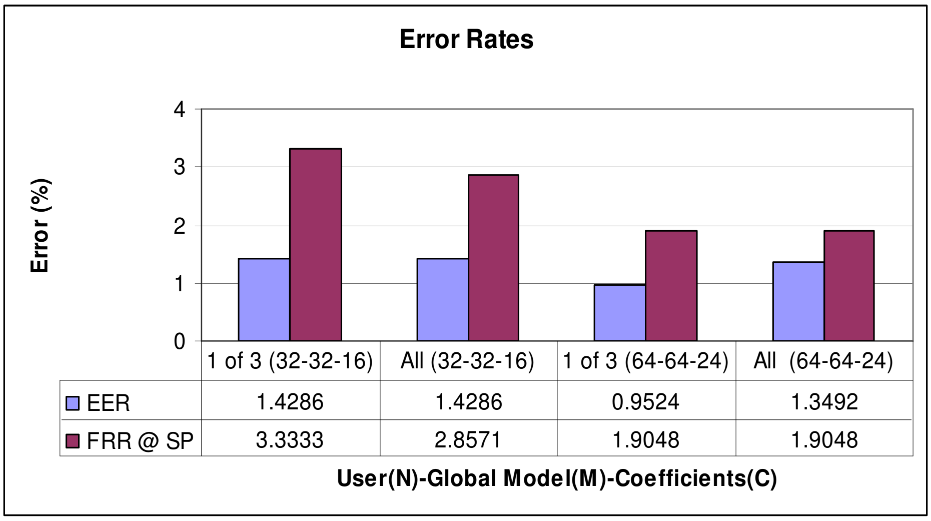
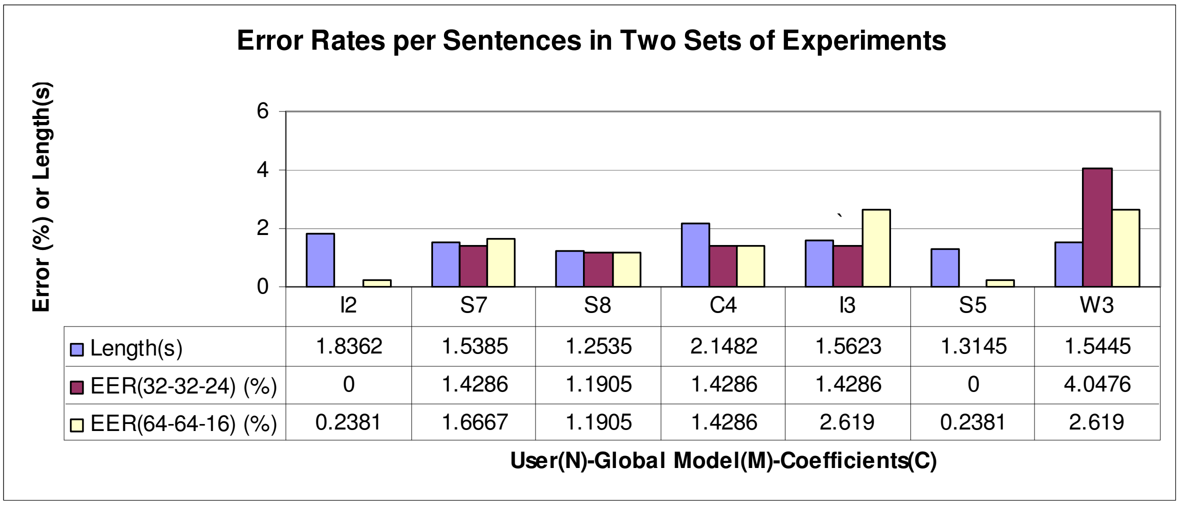


Figure 5-16 Comparison of error rates when a portion or the entire subset E is used for global model training

Experiment Set 3: The Effect of the Types and Lengths of Sentences

In the third set of experiments, the objective is to analyse the errors within each group of sentences. The vertical axis in Figure 5-17 displays equal error rates in percentages for two sets of models as well as the length of sentence in seconds. It is worth recalling that training user models is performed on sentences S1, S2, S3, C1, C2, C3, I1 andW1. The comparison of error rates for S5, S7 and S8 shows that despite their relatively equal lengths and the same type of sentence the errors are lower for S5. This cannot be associated with any other factor except particular differences in the instances of the sentences in the corpus. There is no correlation between the length of the sentences and the error rates within that group of sentences. The high error rate of W3 can neither be explained by the frequency of that type of sentence in training data (comparing to I2 and I3 which have the same frequency) nor by the average length of this sentences. It is obvious from the bar chart that the type of sentence has little impact on the error rates and particular conditions in the recordings accounts for the errors. It is interesting, however that both models have produced similar errors which indicate that model parameters are reliable.

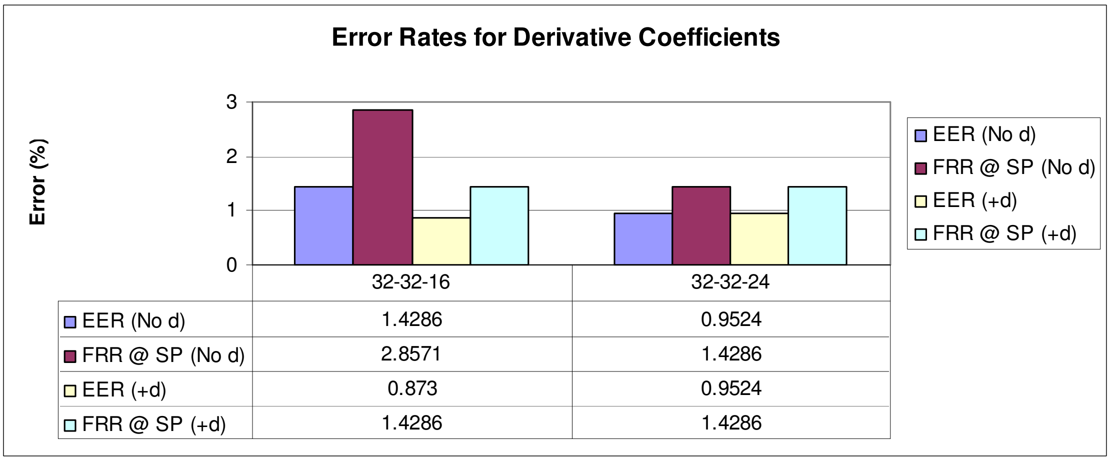


	I2	S7	S8	C4	I3	S5	W3
🟦 Length(s)	1.8362	1.5385	1.2535	2.1482	1.5623	1.3145	1.5445
🟥 EER(32-32-24) (%)	0	1.4286	1.1905	1.4286	1.4286	0	4.0476
🟨 EER(64-64-16) (%)	0.2381	1.6667	1.1905	1.4286	2.619	0.2381	2.619

Figure 5-17 Error rates for various sentences in Subsets B and C

Experiment Set 4: Use of Derivatives

In this fourth set of experiments, first derivatives are used for analysis of the result of inclusion or exclusion of these coefficients. (No d) in Figure 5-18 indicates exclusion of the first derivative and (+ d) denotes its inclusion. Two sets of models with 32 components in the mixture and either 16 or 24 coefficients are used. In one case adding derivatives has improved the results (left group) but it is clear comparing Figure 5-18 with Figure 5-15 that keeping 24 or 28 normal coefficients has been more effective than adding derivatives.

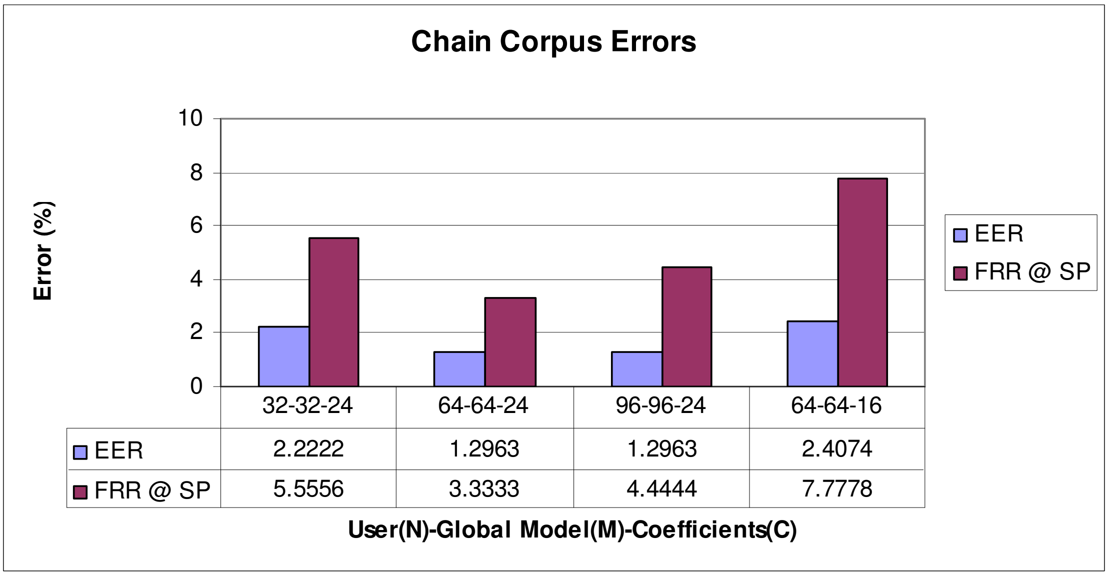


	32-32-16	32-32-24
🟦 EER (No d)	1.4286	0.9524
🟥 FRR @ SP (No d)	2.8571	1.4286
🟨 EER (+d)	0.873	0.9524
□ FRR @ SP (+d)	1.4286	1.4286

Figure 5-18 Impact of inclusion or exclusion of derivatives on error rates

Experiment Set 5: CHAIN Corpus

In this set of experiments, Gaussian mixtures with various numbers of components are trained over the user data and global model data of the CHAIN corpus. The same patterns which were present in the IVIE results could also be observed in Figure 5-20. As a general rule, increasing the number of coefficients improves the verification accuracy. On the other hand, a mixture size of between 32 and 64 produces an acceptable verification performance.



	32-32-24	64-64-24	96-96-24	64-64-16
EER	2.2222	1.2963	1.2963	2.4074
FRR @ SP	5.5556	3.3333	4.4444	7.7778

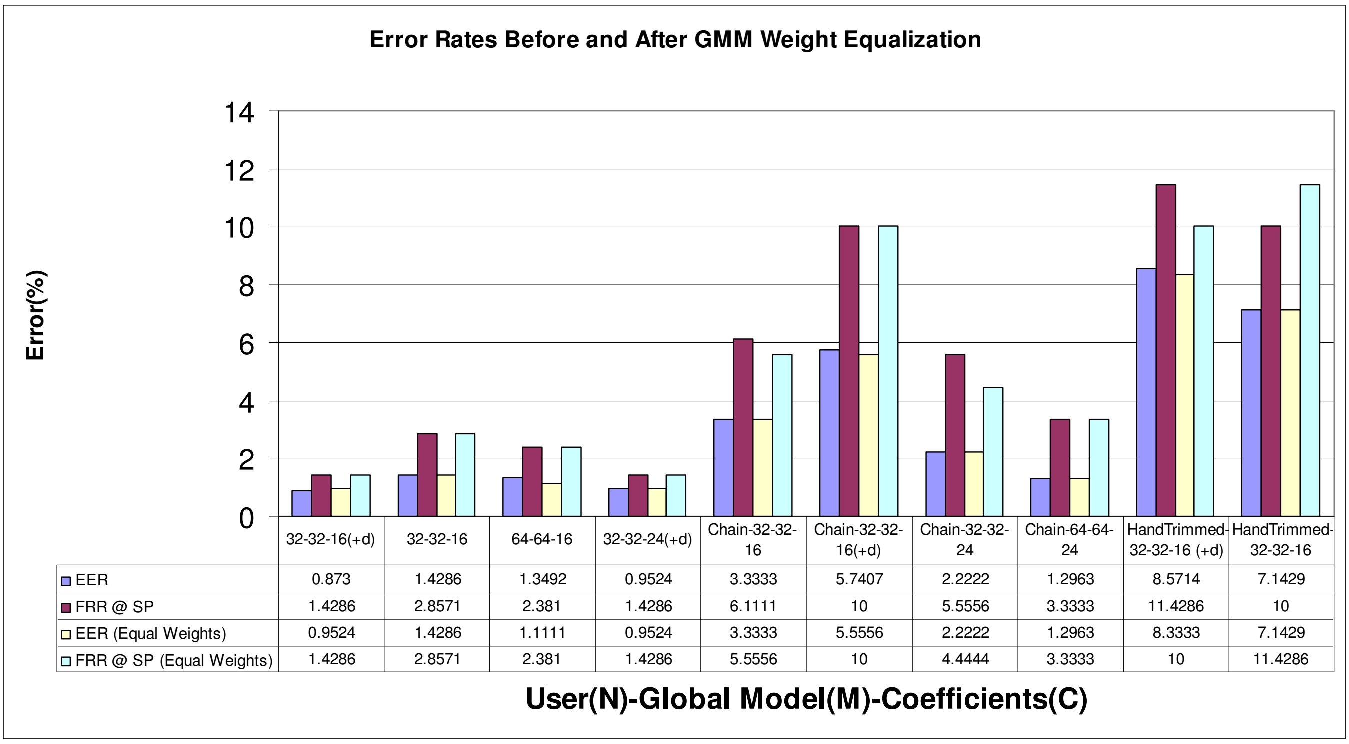
Figure 5-19 Error rates for various combinations of mixture size and cepstral coefficients, CHAIN corpus

Experiment Set 6: The Usefulness of Gaussian Weights in the Mixture

In the final set of experiments a new hypothesis is verified through various experiments. As mentioned before, Yu et al (1995) have shown that for text-independent speaker recognition, hidden Markov models perform as well as Gaussian mixture models and transition probabilities between the states do not play any role in improving verification results. I put forward the hypothesis that even the Gaussian weights in the mixtures have little or no effect in deciding the accuracy of voice verification. It could be explained through the fact that these weights are representative of the frequency of parts of speech in the training data which may be different from that of the test data. Therefore, I expect no significant change in verification outcome as a result of modifying all of the weights in the test experiments with equal values (1/*K* where *K* is the number of Gaussians in the mixture).

Figure 5-21 displays the results of ten sets of experiments in different settings. Neither EERs nor FRRs at the point of high security (FRR=3*FAR) show a meaningful difference between the time equal weights are used and when these weights are normally trained. The standard deviations of weights, when trained, are significant and refute the assumption that the weights automatically approach equal values through training (e.g. standard deviation of 0.0186 and mean of 0.0312 was observed when 32 Gaussians were present in the mixture and standard deviation of 0.0081 for the mean of 0.0156 when 64 Gaussians were used).

The experiments suggest that mixture size between 32 and 64 is a good choice for optimizing verification results in normal conditions while keeping the computations on an acceptable level. As the number of MFCC coefficients reaches 16 and more, the verification errors reach a plateau. The weights of the Gaussians are not important for the purpose of verification. Derivative coefficients are instrumental and use of derivative of cepstral coefficients reduces the error rates.



	32-32-16(+d)	32-32-16	64-64-16	32-32-24(+d)	Chain-32-32-16	Chain-32-32-16(+d)	Chain-32-32-24	Chain-64-64-24	HandTrimmed-32-32-16 (+d)	HandTrimmed-32-32-16
□ EER	0.873	1.4286	1.3492	0.9524	3.3333	5.7407	2.2222	1.2963	8.5714	7.1429
■ FRR @ SP	1.4286	2.8571	2.381	1.4286	6.1111	10	5.5556	3.3333	11.4286	10
□ EER (Equal Weights)	0.9524	1.4286	1.1111	0.9524	3.3333	5.5556	2.2222	1.2963	8.3333	7.1429
□ FRR @ SP (Equal Weights)	1.4286	2.8571	2.381	1.4286	5.5556	10	4.4444	3.3333	10	11.4286

Figure 5-20 Verification results for trained and equal weights in Gaussian mixtures

5.11 Verifying Validity of Clusters and Proposing a Distance Measure

5.11.1 Comparison of Two Recordings and Need for a Distance Measure

One of the common challenges in the field of forensic speaker identification is to decide whether or not two recordings belong to and are uttered by the same person. To be able to automate this process, a similarity measure between two utterances should be defined. Forensic methods described in chapter 3 examine resemblances in parts of speech, such as, formant and duration of phonemes but still lack a universal guide for similarity scoring. On the other hand, text-independent speaker recognition largely relies on statistical data which is usually absent in the forensic cases. For this reason the similarity measures proposed here would be of little use in practical applications. Nevertheless to verify the validity of the clusters produced by the K-means and Gaussian mixture models, within the training data, we need to tackle the same type of problem which is assigning a distance to two recordings based on some or no prior information from the speakers.

A measure for quantifying the distance between two segments of speech (for example sentences) is proposed here and hierarchical clustering methods with their associated dendrograms[1] are employed to illustrate how well the clusters trained on data perform for discriminating speakers.

5.11.2 A Statistical Distance Measure between Two Segments of Speech

Considering two different sentences uttered by the same or different speakers our objective is to define an affinity or distance measure between these two pieces of speech. While in cases where two sentences consist of the same parts of speech, phonetic analysis of the corresponding parts of speech is useful, in automatic text-independent speech analysis (and when the transcription is not available) the distance measure should be built upon overall stochastic characteristics of speech samples.

As with cluster analysis criteria, the distance measures could be defined in numerous ways and the superiority measure of one method over another is its performance. Despite variations two classes of such distance measures could be identified. In the first class lie the metrics which assume previous information from the speakers, e.g. a data model trained on previously collected

[1] “A dendrogram (from Greek dendron "tree", -gramma "drawing") is a tree diagram frequently used to illustrate thearrangement of the clusters produced by a clustering algorithm” (Dendrogram, Wikipedia)

speech data. The second class of distance measures comprises the methods which are just based on the information available in those two pieces of speech with no additional or prior knowledge. The distance measures in the first class are useful in applications which assume knowing the identity of the speaker. They could also be used for verification of the validity of clusters and trained models. The second type of distance measures, have some applications in forensic sciences, where we should establish whether or not two recordings belong to the same person..

I will suggest two distance measures in each category based on GMM probabilities and the sum of standardised Euclidean distance of samples to the cluster centers.

Assuming that *X* and *Y* are two sequences of features from two sentences uttered by speakers *Sx* and *Sy*, M_x and M_y are the models trained on previously collected speech data from speakers *Sx* and *Sy*, and C_{xi} and C_{yi} are *i*-th cluster centers for speakers *Sx* and *Sy,*

Equation 5-5

$$D_{gmm}(X,Y) = -\frac{1}{N_x}\sum_{i=1}^{N_x}\log P(X_i \mid M_y) - \frac{1}{N_y}\sum_{i=1}^{N_y}\log P(Y_i \mid M_x)$$

$$X = \{X_1, X_2, ..., X_{Nx}\}, Y = \{Y_1, Y_2, ..., Y_{Ny}\}$$

This distance measure like cosine distance does not necessarily satisfy the triangle inequality condition. It satisfies non-negativity (assuming that log probabilities are negative) and symmetry conditions.

Similarly another distance measure (D_{Clust}) based on the distance to the centre of clusters could be defined as:

Equation 5-6

$$D_{Clust}(X,Y) = \frac{1}{N_x}\sum_{i=1}^{N_x}\min_j d(X_i, C_{yj}) - \frac{1}{N_y}\sum_{i=1}^{N_y}\min_j d(Y_i, C_{xj})$$

where $\min_j d(X_i, C_{yj})$ denotes the Euclidean distance between X_i and the closest cluster centers of C_y. (C_{xj} and C_{yj} are *j*-th cluster centers for speakers *Sx* and *Sy*)

The feature spaces should be normalised along all dimensions for *X* and *Y* using the same standard deviations to make distances comparable.

It is notable that when there is not any previous information from the speakers (class-II distance measures) the same equations could be used, except that the models and the cluster centers are not pre-defined and are trained based on two speech segments being compared.

5.11.3 Cluster Analysis based on Proposed Distance Measures and Dendrograms

Using the proposed distance measure allows us to assign an affinity value to each pair of training sentences. The four steps involved in cluster analysis of speech data based on the distance metric and plotting dendrograms are:

1. The segments of speech are made by concatenating the sentences. A segment could comprise one or more sentences from one speaker. Two or more segments are built for each speaker.
2. For class-II measure, GMMs are trained for each segment produced in step one (alternatively data clustering can be performed instead of GMM training). Mixture of 32 Gaussians and 16 cepstral coefficients are used in the experiments which are reported in this section (for both classes).
3. The distance between each two segments is calculated based on Equation 5-6 and a distance matrix is built.
4. Dendrograms are plotted based on the distance matrix produced in step 3.

1. Class I Results and Dendrograms

For plotting Dendrograms in class-I category the same GMM models which are trained on speaker data (subset A) are used in Equation 5-5. The main objective here is to verify whether or not the trained model has provided adequate discrimination between sentences from different speakers within training data. As depicted in Figure 5-22 a perfect discrimination is achieved for 48 sentences from 16 speakers. With a cut-off value of 96, 16 clusters will be formed each consisting of the sentences from one speaker only (the scatter plot in Figure 5-23 confirms this). In other words all the sentences from each speaker are placed in the same cluster and are closer to each other than any other sentences based on the proposed distance measure.

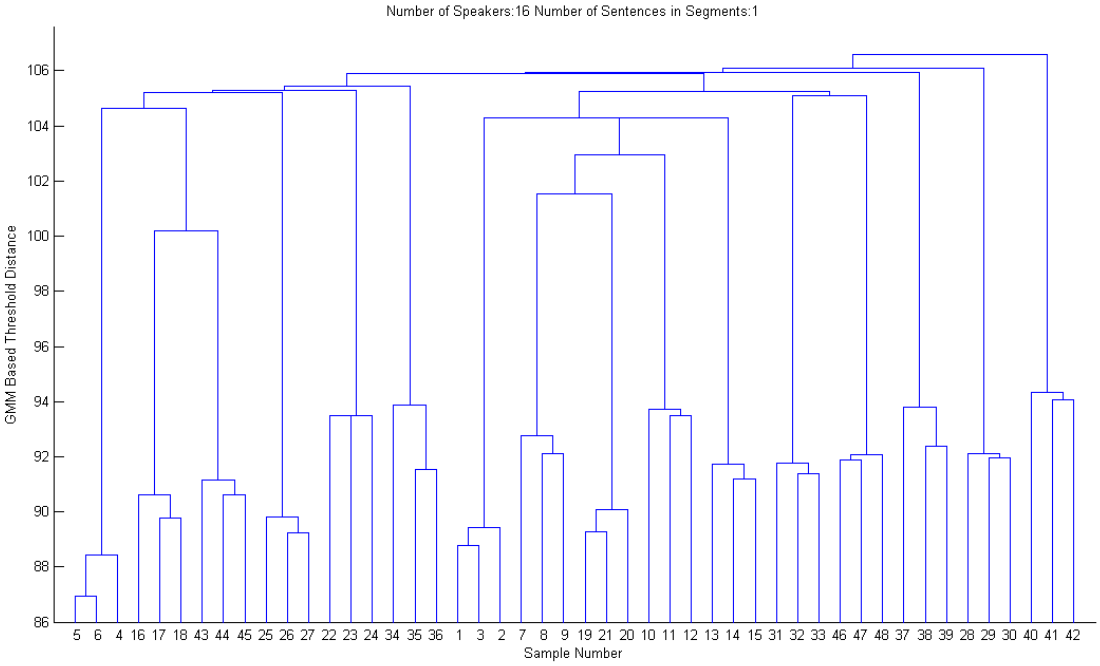


Figure 5-21 Dendrograms for class-I measure, 16 Speakers, 48 Segments each consisting of one sentence, 3 Sentences per speaker

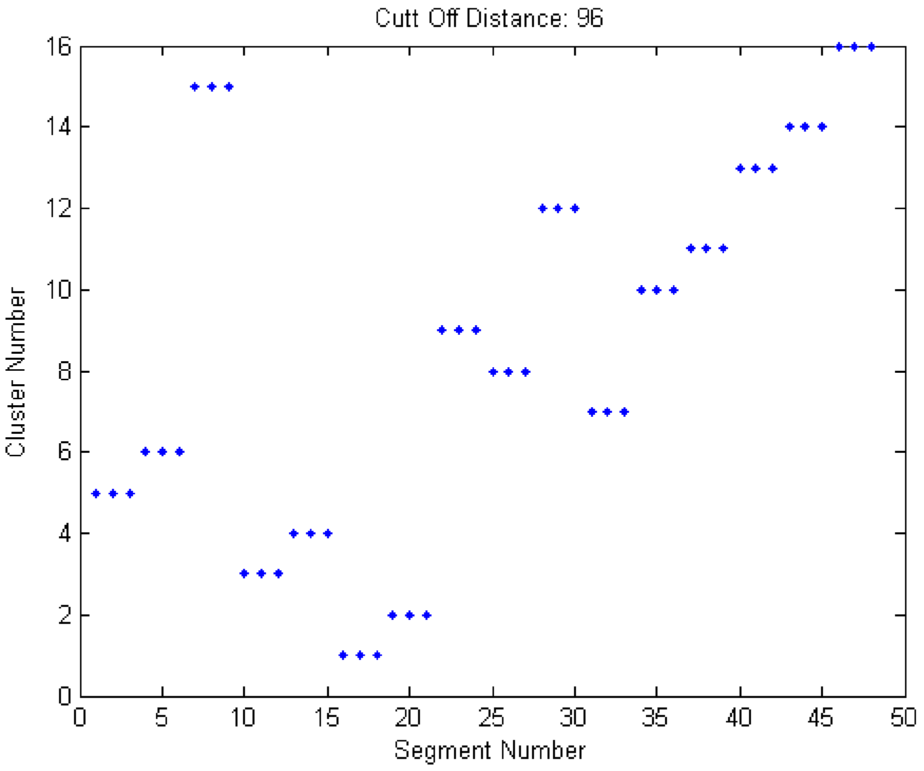


Figure 5-22 Clustering the sentences based on Dendrograms

1. Class II Results and Dendrograms

In contrast to the class I measure of distance, when no prior information about the speakers and the segments being compared is available, the distance measure and dendrograms act as markers of how well two arbitrary sentences can be compared based on their statistical characteristics and regardless of the contents of speech. The length of the sentences should apparently be longer in this case and for a perfect discrimination between segments from different speakers the number of speakers should be fewer.

Figure 5-24, illustrates that for 16 segments, each consisting of 3 sentences, the discrimination between clusters has been perfect and no error has occurred. For shorter segments including 2 sentences and for 24 segments just one error can be observed: (segment 12) in Figure 5-25.

The analysis confirms the clusters' validity in the training data based on the class-I distance measures and shows that the distance measure can discriminate speakers based on statistical measures, when two or more sentences are available from each speaker.

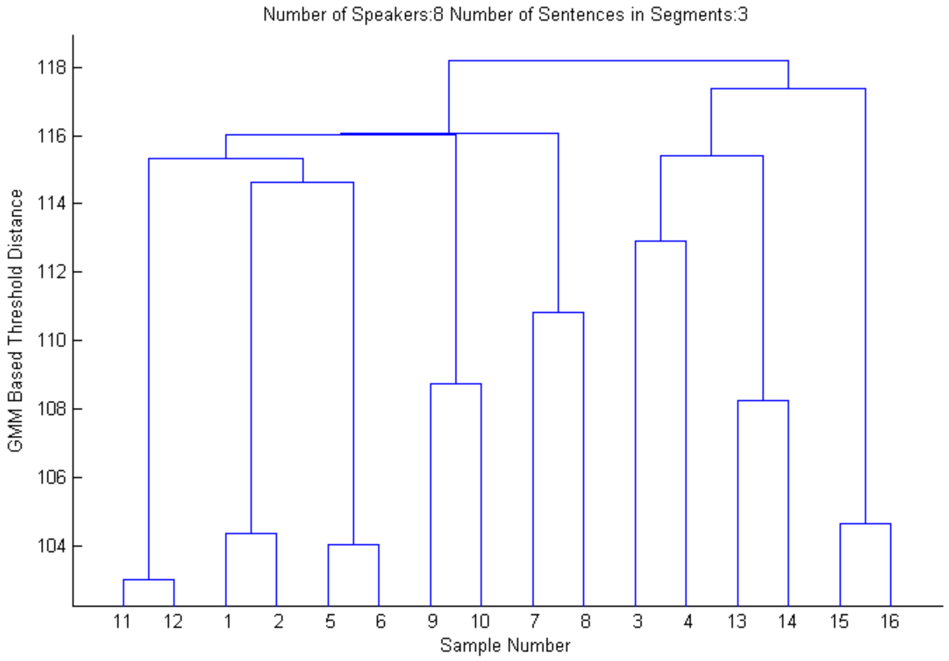


Figure 5-23 Dendrograms for class-II measure, 8 Speakers, 16 Segments each consisting of TWO sentences per speaker

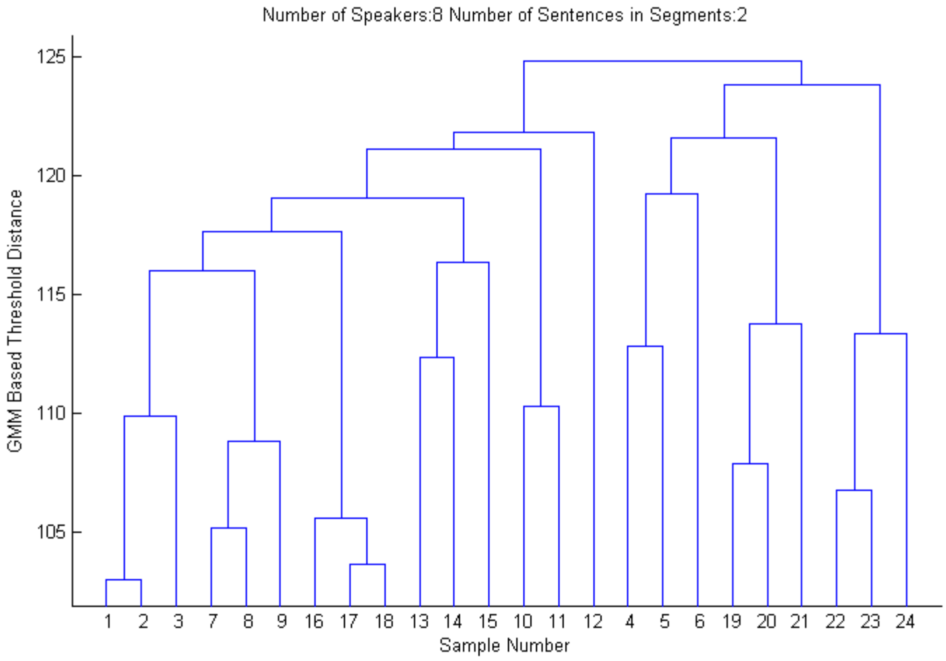


Figure 5-24 Dendrograms for class-II measure, 8 Speakers, 24 Segments each consisting of THREE sentences per speaker

5.12 Highlighting the Need and Describing the Method Adopted for Silence Removal

5.12.1 Rationale for Voice Activity Detection in Speaker Verification Systems

In actual voice verification scenarios where the user is prompted to read out a phrase, the delay, the pauses and the boundaries of the speech are unknown to the system. It is very likely that the speaker does not keep the microphone at a suitable distance, changes the position of the microphone (which results recording at various ranges of amplitude) or cannot follow the instructions given by the system. Under such conditions, the voice data has random pauses in voice activity.

In real-life circumstances, the segment which has to be verified is unknown and the system should decide when to terminate recording and what parts of the recording should be discarded. Therefore, the voice activity detection (VAD) module has a major role in making the interaction with the user robust against possible pauses.

In addition to scenario tests, in the technology tests (test of algorithms) a silence removal module is similarly an indispensable part of the voice verification system. The pauses or non speech parts exist in the start, middle, and in the end of the recording unless they are manually removed from both the training and verification data. Due to the fact that enrollment and verification algorithms do not discriminate between 'silence' and 'speech' frames and since the statistical characteristics of silence or low amplitude frames are decided by the environmental noises, inclusion of these frames in the training and testing data can both distort the trained GMM models and falsify the score assigned by them[1].

5.12.2 Possible Approaches to Removing Non-Speech Segments

One of the earliest attempts for distinguishing between speech and non-speech segments and finding the boundaries of words was made by Rabiner and Sambur (1975) in which they tried to determine the start and end of 'isolated words' for recognition purposes using energy and zero crossing measures. Haigh and Mason (1993), while mentioning works based on the combination of these features (zero-crossing and energy-related measures) adopted a different approach based on the distance in cepstral space. According to Górriz et al. (2005, p. 1), the different approaches for voice activity detection include "those based on energy thresholds, pitch detection, spectrum analysis, zero-crossing rate, periodicity measure, and higher order statistics in the LPC residual domain or combinations of different features". They employed two-dimensional discrete Fourier transform of third order cumulant function with statistical tests (generalised likelihood ratio and the central Chi-squared test) in different noise conditions and signal to noise ratios to separate speech and non-speech segments. In more recent work, Kinnuen et al. used support vector machines in Mel-cepstral space to classify speech and non-speech frames. On the frames of 30 milliseconds they reported equal error rates of around 9 % in classification (2007).

5.12.3 Non-Speech Segment Removal Module Implemented in This Work

In this research the majority of the experiments are carried out on manually trimmed data. There are only two subsets of the IVIE Corpus which have considerable silent segments and pauses in their recordings: 'Read Passages' and 'Retold'. I have manually removed silences and pauses in

[1] Even if we neglect the environmental noise and assume that all silence frames are equal or close in the cepstral space, the score assigned to a piece of speech with large silences *in the verification session* will be a function of how much silence has been present in the data from the *enrollment session* of the speaker being verified. The more silent parts in training data, the more it is likely that a recording containing silence passes the verification threshold which is a security threat in itself.

sentences in the ‘Retold’ subset to create a ‘Hand Trimmed Retold’ subset but there are still experiments discussed in chapter 7 which need automatic removal of silences. A rather simple voice-activity detection module is used in the system, which is based on the combination of zero-crossing and either normalised energy or sum of amplitudes in the frame for silence detection.

For testing the VAD module, fifteen passages were chosen from the *speakers* present in the Subset E (global GMM training subset). The passages were taken from ‘Read Passages’ of the IVIE corpus (since the speakers are different from test speakers this data was not used in any verification experiment). Care was taken to ensure:

1. There was no overlap between the *test speakers* and the speakers from which these passages were chosen
2. There was no overlap between this *data* and any other *training* or verification data.

A total 13,450 frames of speech (160 milliseconds, half overlapped) and 4,949 frames containing silence were labeled manually for testing.

The two measures used for voice activity detection were:

1. Zero crossings (number of times the x-axis was crossed in a frame);
2. Either normalised energy (energy of the frame divided by the highest energy per frame in that utterance) or magnitude measure (sum of the amplitudes in the frame divided by the highest value of this sum among the frames in that particular utterance).

Frames for which a metric based on these two features falls below certain values are treated as non-speech and are removed.

Figure 5-25 displays the points corresponding to speech/non speech frames. Despite the fact that points corresponding to the non speech frames are accumulated near the origin of the graph there is some overlap between two classes. A linear classifier can be used to make a distinction between speech and non-speech frames. Two different lines and sets of thresholds are shown in this figure: (50, 0.1) and (25, 0.05). The classification error rates for these two sets of values are presented in Figure 5-26.

In Figure 5-26 a three dimensional graph of frame level classification errors is plotted against zero crossings and an energy measure when these two measures are used for linear classification. It is assumed that the line that connects two points of (energy measure, 0) and (0, zero crossings) is used to separate these two classes. The points below this line are considered as corresponding to non-speech frames and the points above the line are considered to be related to speech frames.

The two surfaces in the plot correspond to the percentage of speech points below this line (classification error for speech frames) and the percentage of non-speech points above this line (classification error for non-speech frames). Similarly Figure 5-28 shows the same diagram for zero crossings and the magnitude measure.

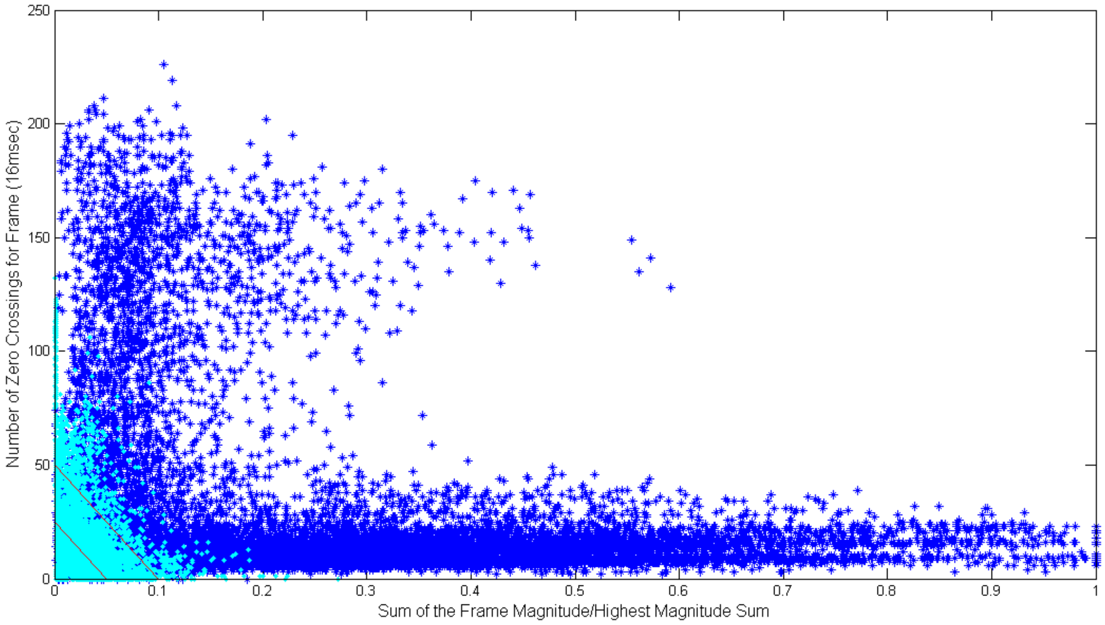


Figure 5-25 Plot of zero crossings against magnitude for frames of two groups of manually extracted speech and non-speech data

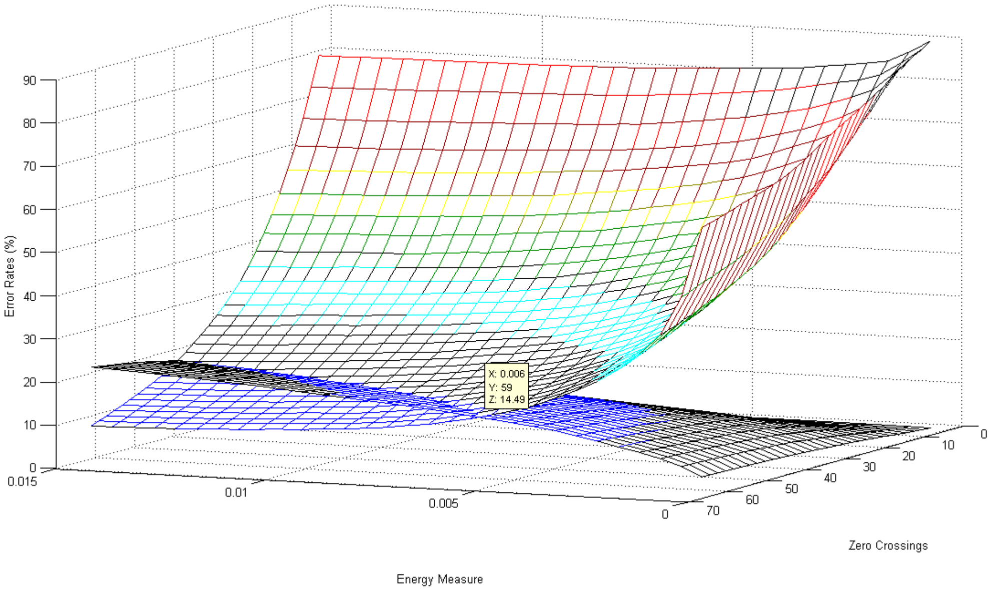


Figure 5-26 Frame level classification error as a function of zero crossings and energy measure

Following the line of interception of these two surfaces shows that minimum frame level classification's equal error rate for the energy measure is 14.63% and that for the magnitude measure is 13.97%. Figure 5-27 shows the curve which relates the zero crossings values to the magnitude measure at the points of equal error rates for the magnitude measure (points of interception). There is no reason however to assume that it is desirable to choose the equal error rate parameters for optimization of verification results.

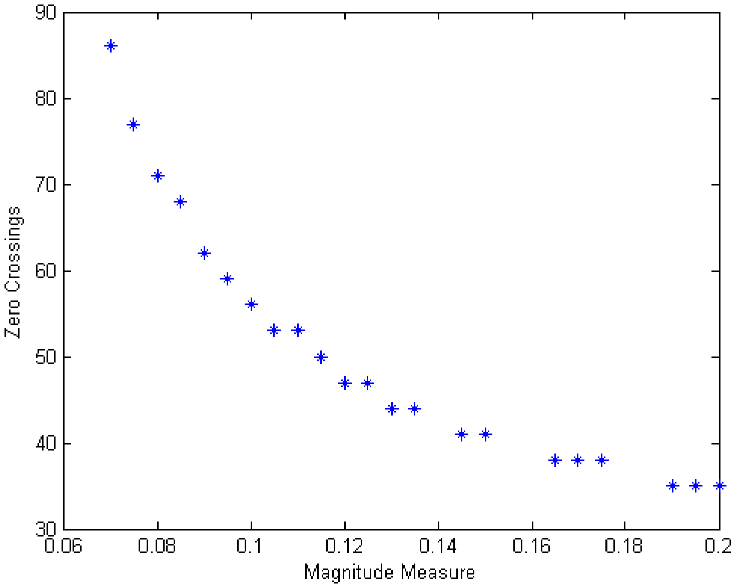


Figure 5-27 Zero crossings and magnitude measures along the interception line

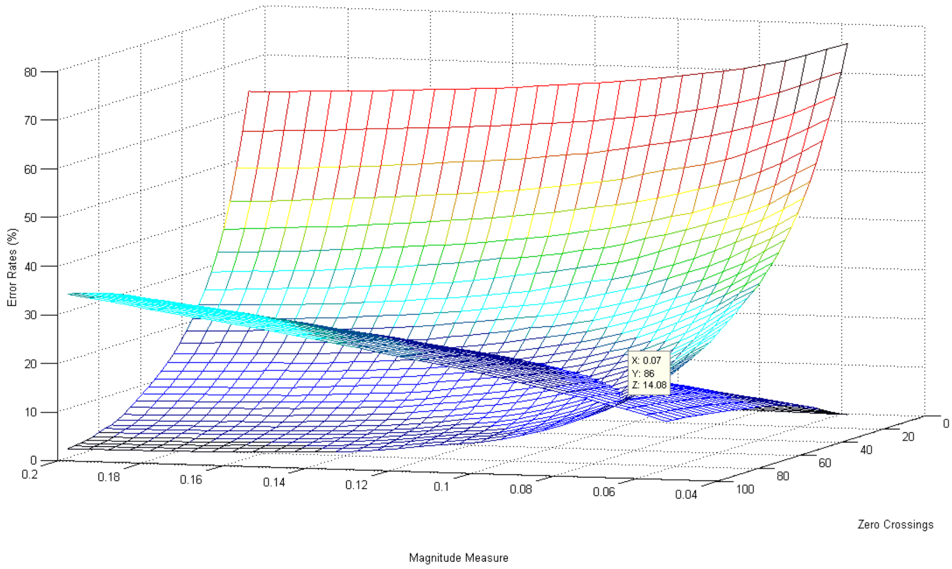
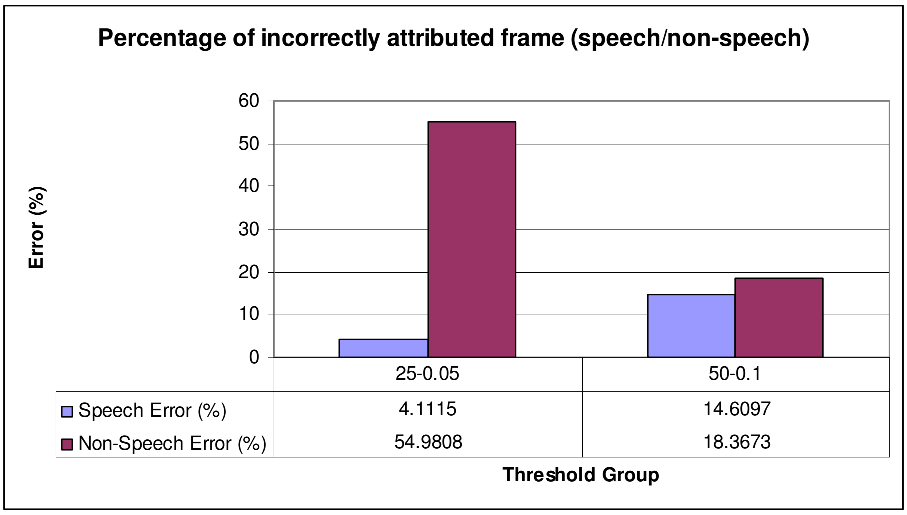


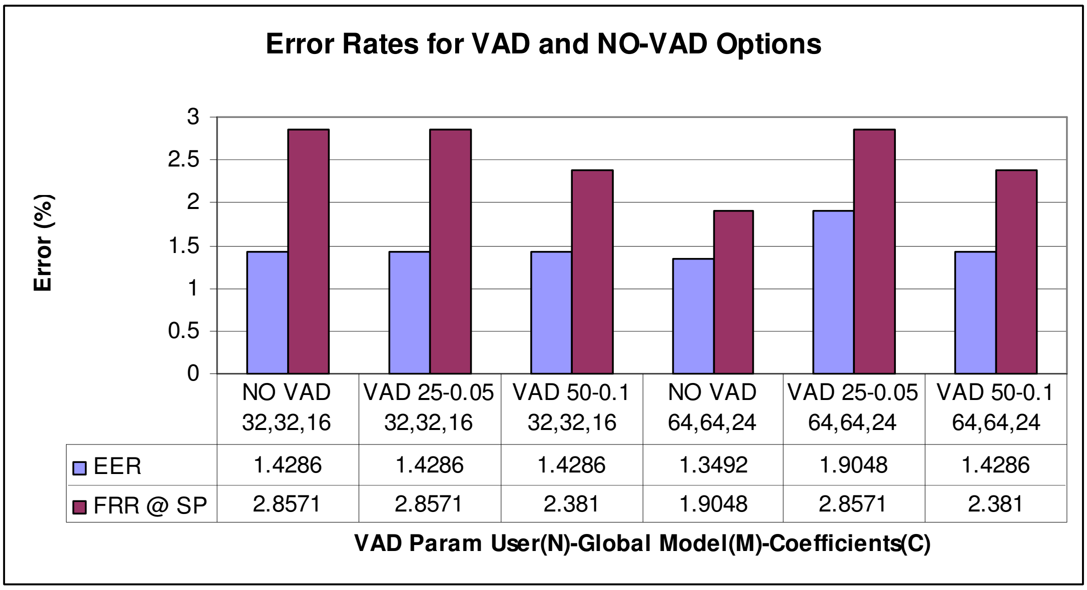
Figure 5-28 Frame level classification error as a function of zero crossings and magnitude measure

Finally Figures 5-29 and 5-30 display frame level classification errors and speaker verification errors for two different sets of thresholds. Several model parameters and coefficient numbers are used in combination with the voice activity detection module. The verification results indicate that the impact of the choice of VAD parameters is minimal on the errors for the isolated sentences. In most cases there is sufficient data left after VAD for training the models during enrollment and in the verification process.



	25-0.05	50-0.1
□ Speech Error (%)	4.1115	14.6097
■ Non-Speech Error (%)	54.9808	18.3673

Figure 5-29 Frame level classification error in detection of speech/non-speech for two sets of thresholds (zero-crossings/magnitude measure)



	NO VAD 32,32,16	VAD 25-0.05 32,32,16	VAD 50-0.1 32,32,16	NO VAD 64,64,24	VAD 25-0.05 64,64,24	VAD 50-0.1 64,64,24
🟦 EER	1.4286	1.4286	1.4286	1.3492	1.9048	1.4286
🟥 FRR @ SP	2.8571	2.8571	2.381	1.9048	2.8571	2.381

Figure 5-30 Verification errors for various voice activity detection and model parameters[1] (zero crossings/magnitude measure)

[1] The first two values are VAD parameters. The next three numbers are size of user models, global model and number of coefficients respectively.

5.13 Summary of Findings and Provisional Conclusion

The aim of this chapter was to provide information about the prototype voice verification system that has been developed for carrying out the rest of the experimental analyses in this research. The structure of the system and the two corpora used in this research were presented in the preceding sections. Choices in the selection of the parameters were discussed and the effect of changes in those parameters was studied through seven sets of experiments.

In addition to choosing suitable values for parameters such as the size of Gaussian mixtures, number of clusters, number of cepstral coefficients, number and inclusion or exclusion of derivative coefficients, the secondary objective was to identify how different choices alter the verification results. In order to shed light on what qualities Gaussian mixtures capture in cepstral space, an attempt was made to calculate the number of distinct clusters present in each speaker's feature space and what these Gaussians are tantamount to. It was demonstrated through visual aids as well as by the cluster analysis criteria that the overlap between speaker's spaces and parts of speech in one person's cepstral space is considerable. The percentages of correctly attributed frame level observations were around 37% and 33% for the IVIE and CHAIN corpora with 30 speakers and the curves reached plateaus for a large number of clusters which indicates that the problem associated with the overlap could not be alleviated by the addition of more clusters.

Additional outcomes of the experiments can be summarised as follows:

1. Increasing the number of clusters and Gaussians in the mixture for each speaker improves the verification results when those numbers are small but the improvement becomes negligible when the mixture size outgrows the range from 32 to 64.

2. It was known before that the transition probabilities of HMMs have no impact on speaker recognition results. It was proposed here that even the mixture weights in GMMs are more reflective of the frequency of parts of speech in the sentence. The experiments showed that substituting these weights with equal numbers does not cause any deterioration in the verification results.

3. The percentage of correctly attributed samples shows that there is a significant overlap between 'speaker spaces'. In the normalised cepstral space the probability that a single frame, assigned to the nearest cluster centre, is associated with its true speaker is approximately around 35% for 30 speakers regardless of how many clusters are used.

4. Increasing the feature vector size by adding derivative of features are helpful but not as helpful as adding more normal features.

5. The verification results did not seem to be affected by the type of sentence.

After trying to build a robust speaker verification system and rationalizing some of the choices made, the vulnerability analysis of such a verification system in face of security threats and adverse conditions will be carried out. In the next chapters two types of spoof attacks are proposed, one in each category of threats (conversion and synthesis). Based on the recommendations made in chapter 4 it will be examined whether (commonly used) speaker recognition systems (based on the techniques proposed here) are vulnerable to such types of attacks or can withstand (at least) these counterfeit techniques which are put forward and implemented.

Chapter 6 : Security Analysis of Voice Verification Using the Proposed Framework

6.1 Goal of This Chapter

This chapter provides a practical exercise of the recommendations made in chapter four by analysis of how the commonly used GMM/HMM speaker recognition systems (and in particular the one detailed in chapter 5) would withstand two types of attacks one in each category introduced previously: speech synthesis and voice conversion. The fact that we can not evaluate 'all' voice verification systems against 'any' type of attack may seem somehow disappointing but for practical purposes we are limited to testing specific algorithms separately and for different systems. A security evaluation system is not complete unless it consists of a range of such attacks including the two spoofing techniques introduced here. Therefore, the experiments conducted here enhance our knowledge about these two new types of attacks. In addition, development of these types of attacks allows exercising the procedure proposed before with the details deemed to be significant for interpretation of the results.

Theoretically, the results of the experiments carried out here can just reveal whether or not the *voice verification system under investigation* endures *these two types of attacks.* Nevertheless, from a practical perspective the following experiments indicate to what extent a *typical* voice verification system could be secure and how much the similarity of the verification and spoofing modules can change the results.

The speech signal produced by the spoofing modules will be used as raw material for analyses of chapter 8 where we determine how far (if at all) the objective of automatic recognition of synthetic or manipulated speech is realistic.

More specifically this chapter will seek to answer the following questions:

1. To what extent does the success of spoof depend on the similarity of the conversion algorithm / parameters and the algorithms /parameters used for verification?
2. Are there any connections between reliability and security?
3. How much speech data from a target speaker do intruders need to enable them to fool the verification system?

4. How much knowledge about the verification system and its parameters is required for launching a successful attack?

The focus is mainly on the speaker recognition module. As mentioned in chapter five, a real content verification module would consist of triphone HMM models for each phoneme (which is trained separately) and concatenated to form any desired sentence. In this work, the HMM models are trained for the entire sentences and the score of a particular sequence of observation is derived through calculating the probability of observing *that* sentence given *all* the HMM models.

The rest of this chapter is organised as follows,. The next section describes the theory of voice conversion algorithms and this is followed by a discussion and analysis of the experimental results of spoof attacks using the developed conversion method. Following that in section 6.4, the HMM based speech synthesis is introduced and in 6.5 the security evaluation results are discussed. Section 6.6 contains a discussion about the implications of these experiments for other areas where voice can be used as a security biometric. Finally, in section 6.7, some tentative conclusions are reached and the aims of the following chapter are described.

6.2 A Voice Conversion Algorithm Based on Hill Climbing

6.2.1 Goal of Voice Conversion

For a general conversion algorithm which seeks to maximise or increase the score of a piece of speech with regard to a speaker model three steps can be identified:

1. The speech signal is transformed to a feature space e.g. cepstral space or frequency domain;
2. The features are altered to resemble the target features or models (e.g. those which belong to a target speaker);
3. A speech signal is built from the results of step two which produces the same sequence of features

Since the relation between signals and features are normally many to one, there are usually infinite signals which produce the same feature set and therefore there is huge room for variation and innovation in step 3.

The method used here for voice conversion is based on making successive changes in the features extracted from the frames of the source speaker’s voice in order to increase the probability of the features given the target speaker's Gaussian mixture models while preserving the nature of the uttered phrase (contents of speech). It uses a gradient ascent algorithm in combination with spoof GMM models to transfer features towards local maxima in the target speaker's cepstral space starting from true cepstral coefficients extracted from the source data.

If *O* ($O = \{o_1, ..., o_T\}$) is the sequence of observations of the length *T* and each observation is a feature vector extracted from a frame of speech, the probability of the observations given the HMM model (used for verification of speech contents) and a path (*Q*) on the model will be:

Equation 6-1

$$P(O \mid \lambda, Q) = \prod_{t=1}^{T} a_{q_{t-1}q_t} \sum_{i=1}^{M} K_i N(o_t, \mu_{iq}, \delta_{iq})$$

where *a* is the transition probability and a_{0i} is the prior probability of state *i* (probability of starting from state *i*) and q_0=0. *Q* is the path or sequence of states on HMM starting from q_1 and ending in q_T. μ_{iq}, δ_{iq} are the mean and covariance of the Gaussian *i* of state *q*. The probability of observations given the model, $P(O \mid \lambda)$ is the sum of this probability for all possible sequences of states(*Q*) where λ symbolises the HMM model for the phrase.

The probability of the observations given the GMM can be written as:

Equation 6-2

$$P(O \mid \Pi) = \prod_{t=1}^{T} \sum_{i=1}^{M} K_i N(o_t, \mu_i, \delta_i)$$

where Π is the Gaussian mixture model with *M* components. *N* is the normal distribution and μ_i, δ_i are the mean and covariance of the *i*-th Gaussian in the mixture.

For step 2, we seek to maximise either the GMM probability or a combination of GMM and HMM probabilities to meet speech and speaker recognition criteria.

6.2.2 Finding the Optimum Sequence of Observations in Feature Space

In a typical conversion algorithm the features of speech are converted and transformed to resemble those of a target speaker. In the GMM based voice conversion, new features can be found that are based on the models the intruders can train or obtain from the speaker. The idea behind the algorithm suggested and developed here is that of finding and moving towards the local maxima of the distribution function of the features belonging to a target speaker in cepstral space. The distribution functions are the speaker's GMM models.

It may seem a trivial task to find the local maxima of a GMM model (also known as the modes of a GMM) but no direct method currently exists for locating these local maxima even in the simple cases. Carreira-Perpiñán has given a partial proof that the number of modes cannot be more that the number of Gaussians and has developed two mode searching algorithms; one based on gradient ascent and the other based on a fixed point iterative scheme for finding these modes (2000). It is notable, however, that for our purpose the objective is not to find these local maxima because substituting the source vectors with them changes the contents of speech drastically and makes the sentence unable to satisfy the content verification criteria (HMM equation).

To solve this problem we choose to increase this probability by relocating each observation vector (o_t) in the direction which increases the probability given the target speaker's GMM (hill climbing by gradient ascent). To find this direction we use the gradient of Gaussians at the given point. The Gradient of a Gaussian distribution at point *Ot* could be worked out as follows (Carreira-Perpiñán, Chapter 8, 2001):

Equation 6-3

$$\nabla N(o_t, \mu_i, \delta_i) = \delta_i^{-1}.(\mu_i - o_t)$$

where the left side of the equation is the gradient, μ_i, δ_i are the mean and covariance of the *i*-th Gaussian in the mixture and o_t is the observation at time *t* ($\nabla g(x, y, z)$ is a vector of partial derivatives of $g(x, y, z)$ i.e. $(dg / dx, dg / gy, dg / dz)$).

Combining equations 6-2 and 6-3 yields that:

Equation 6-4

$$\nabla P(o_t \mid \Pi) = \sum_{i=1}^{M} K_i \delta_i^{-1}.(\mu_i - o_t)$$

This specifies the direction in which we should relocate the observation at time t:

Equation 6-5

$$o_t' = o_t + k.\nabla P(o_t \mid \Pi)$$

The process of maximizing the probability of observation vector (o_t) given the GMM model, is iterational and the above adjustment should be applied to the observation vector several times, until an ending criterion is satisfied.

Through the experiments it was observed that finding a proper and robust value for *k* is easier and the process will be faster if we choose to maximise log-probability instead of just the probability. We can work out the gradient in log domain as follows (Carreira-Perpiñán, Chapter 8, 2001):

Equation 6-6

$$\nabla \ln P(o_t \mid \Pi) = \frac{1}{P(o_t \mid \Pi)} \sum_{i=1}^{M} K_i \delta_i^{-1} .(\mu_i - o_t)$$

It is worth mentioning that if instead of going towards the local maxima we had chosen to find the modes or global maxima given the GMM model, the nature of speech would have changed drastically and the probability of the observations, given the HMM model, would have decreased considerably.

There is always a trade-off between the two goals of increasing probability based on GMMs and keeping speech contents represented by HMM unchanged. One may choose to maximise the following term instead:

Equation 6-7

$$Score(O \mid \lambda, \Pi) = f(P(O \mid \lambda), P(O \mid \Pi))$$

where *f* is an arbitrary function of probability of the observation given the HMM model and the path and the probability of observation given the speaker's model. Function *f* may be chosen to weight any of these two terms over the other, for example, if preserving the content of speech is more important, *f* could be chosen so that the weight of the HMM term is higher. There are, however, obstacles to maximizing $P(O \mid \lambda)$. It is not possible to obtain analytically the observation sequence which maximises $P(O \mid \lambda)$ in a closed form (Masukoy (2002)). In order to synthesise a sequence of observations based on a hidden Markov model, Masukoy chose to maximise $P(O \mid \lambda, Q)$ where Q is a path on the HMM[1].

While similar methods could be applied for HMM based adjustments (which due to the trade-off worsens the scores given the GMM) in this work the emphasis has been on increasing

[1] With the goal of HMM based speech synthesis which his PHD thesis revolves around rather than conversion

$P(O \mid \Pi)$ through a number of iterations, therefore, adjustments based on HMM-score have not been carried out.

6.2.3 Generating Speech Signal from Features

As mentioned before, the third step of artificial speech generation involves reproduction of speech samples from the sequence of features obtained through hill climbing. The problem could be described as finding signal *S (*consisting of the sequence of frames $X' = \{x_1', ...x_T'\}$) for a sequence of feature vectors $O' = \{o_1', ..., o_T'\}$ (which are obtained through gradient ascent) where each vector is of the size *C* and corresponds to a frame in *S* with the length of *2N* samples. *C* is generally smaller than *2N* and in our method is based on the dimension of spoofing GMM models (for example, if spoofing GMMs are trained on vectors with 16 coefficients *C* equals 16 while for 16msec frames of speech length of f_i is *2N*=256).

Figure 6-1 displays the steps involved in the calculation of Mel-cepstral coefficients over a speech frame of *2N* samples which was described in chapter 3. The process of extracting these coefficients from voice samples is awkward (i.e. some information is lost through the process) and the reverse path is not unique.

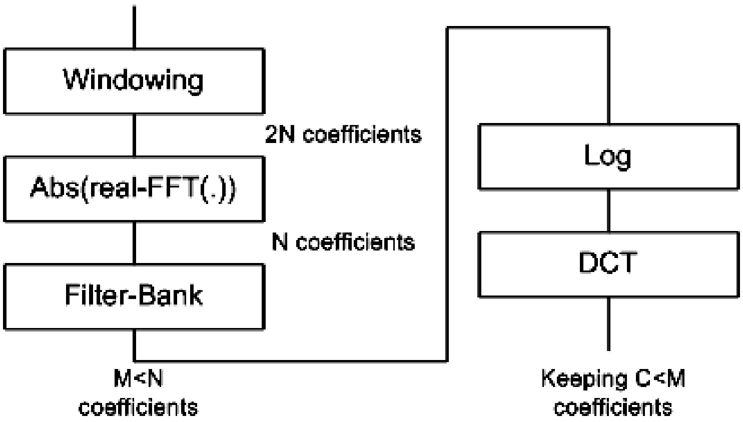


Figure 6-1 Process of extracting Mel-cepstral coefficient from a frame of *2N* samples

Several methods have been suggested for regeneration of speech from MFCC coefficients. An algorithm based on Mel-log spectral approximation (MLSA) filters which uses MFCC coefficients as well as pitch (fundamental frequency) is widely used for HMM based speech synthesis (Zen and Toda, 2005) and decoding speech from sequence of MFCC features (e.g by Tokuda, et al. 1998 and Chazan, et al, 2000). Another approach is based on attempting to reverse the extraction process and reconstruct speech through sinusoidal or source filter models of speech (Milner and Shao, 2006).

The inverse process proposed here has some overlap with the Milner and Shao methods in reversing the extraction procedure, but uses some information from the source speech to approximate the envelope of the spectrum for each frame.

The detailed description of the process is as follows (for each frame)[1]:

1. Calculate *M* MFCC coefficients from source frame and substitute first *C* coefficients in the DCT vector of the source speech's frame with the target coefficients that are found through hill climbing.

Assume that o'_t is the modified feature vector from the source frame vector o_t which is extracted from the source frame x_t (o'_t and o_t are *C* by1 and x_t is *2N* by 1). We aim for finding x'_t which has feature vector o'_t knowing o_t and x_t.

One difference between our problem and the common synthesis problems is that we have access to the real data available from the source speaker's frames of speech (x_t).

In a few works Milner and Shao have used zero padding for the MFCC vector to reach the dimensionality of the filterbank and then have applied an inverse DCT followed by an exponential operation. In contrast, in this work, the rest of the MFCC coefficients from the source frame are used for obtaining the dimensionality of filter-banks which are supposed to preserve some fine structures in the source signal. Before extracting features from source, a signal hamming window is applied to the frame.

The outcome of this step is that we will have v_t which is *M* by 1 and is obtained by adding *(M-C)* elements from source frame DCT coefficients to o'_t.

2. Inverse Discrete Cosine Transform (DCT) from the *M*-element vector is taken (*IDCT(* v_t *)*).

3. To reverse the effect of taking *log(.)* function *exp(.)* of the M-element vector is calculated similar to Milner and Shao approach (w_t *=exp(IDCT(* v_t *)))*.

4. To obtain the spectral coefficients from the results of step 3 a new method is used which is based on finding the inverse of filterbank matrix.

[1] The reader who is uninterested in the details of proposed algorithm could skip the rest of this section and continue reading from 6.3.

If the filter-bank matrix was square the inverse matrix could be used for reversing the filtering step. Alternatively for rectangular matrices the Moore-Penrose inverse matrix can be obtained. For any matrix A *(*m by n) there exists a unique Moore-Penrose inverse, denoted by A^+ (n by m) which satisfies the four conditions (Schmidt,2008): $A.A^+.A=A$, $A^+.A.A^+=A^+$, $(A.A^+)^*= A.A^+$, $(A^+.A)^*= A^+.A$ where * operator denotes conjugate (Hermitian) transpose (obtained by transposing matrix and taking complex conjugate)[1]. MATLAB *pinverse* function is used for calculating the inverse of filterbank matrix once for all frames.

The result of multiplying N by M matrix A^+ which is the pseudo-inverse of the filter-bank matrix by M by 1 vector obtained in step 3 (w_t) is an N by 1 vector which corresponds to the magnitude of the first half of the discrete Fourier transform of the frame (Since the speech signal is real, the second half of the Fourier transform is the mirror of first half): $f_t = A^+.w_t$

5. Reconstruction of signal from spectral information

The result of step 4 will be N coefficients representing amplitude of the Fourier transform of desired signal. There are many signals which produce the same spectrum. The following methods are among techniques which could be used to produce $2N$ samples with the desired frequency amplitude:

Frequency response of a random white noise is assumed to be flat. A random white noise frame can be used to stimulate the spectral envelope into time domain frame of speech.

The same procedure can be done by a frame containing Dirac delta (which has flat frequency response)

Suppose that f_t is the desired spectral amplitude vector obtained in step 4 and g_t is the Fourier transform of source signal x_t. In the third method which was used in the experiments a filter with the frequency response of f_t / g_t was applied to the original source frame x_t. The result will have the desired amplitude of Fourier transform and carries some of the original signal's characteristics (such as phase and the first Fourier element which represents the amplitude of the signal) in addition to the spectral characteristics.

While all the above techniques produce the desired spectral amplitude the preliminary experiments with these methods show that the third method produces slightly better results in terms of false acceptance rate on a small portion of data therefore this technique was adopted.

[1] The matrix is named after Moore and Penrose who described the matrix in Moore (1920) and Penrose (1955).

6.3 Results of Security Evaluation Experiments for Voice Conversion

6.3.1 Design of Experiments

The algorithm described above, was implemented as an independent module in MATLAB which also could be accessed from the prototype application in Java. Separate GMM models for each speaker were trained based on data in Subset C as described in chapter 5. The spoofing interface allows any evaluation system that adheres to the rules of evaluation to call this module. The spoofing module receives the source file, target file, and the user-ID, that is, the user for whom the conversion is intended to be performed. The source file is any of the files in the test subset (Subset B). The target file is the result of conversion and the conversion module stores the converted signal in the designated location. The evaluation module sends all the source files directly to the verification system to work out and plot the false rejection rate curve. To evaluate the false acceptance rate for different thresholds, the evaluation system sends the source file to the conversion module and sends the target file (converted voice) to the verification system. It is notable that the false acceptance rate obtained in this case is only the false acceptance rate of the converted signal and the normal false acceptance rate (as a result of the inaccuracy of the verification system) still exists but is not included in the error rates.

Within the conversion module, the hill climbing algorithm was applied to each of the half overlapping frames of speech (25ms frames, with 12.5ms overlap). The length of frames was deliberately different from that used for voice verification and content verification (16ms) to rule out any claim that the success of spoofing had been due to temporal manipulation of identical frames. The resultant (converted) frames were multiplied by the hamming window and the shared halves of each of two consecutive frames were summed to produce the final sequence of samples.

As described in chapter 5 Subset C was used for training spoofing GMMs for voice conversion. It was assumed for the spoof experiments that the conversion module did not have access to the same data which training models were built upon therefore subset C had no overlap with subset A (used for training users' GMM models) and B (test sentences). If the conversion performed in this fashion proved unsuccessful we can redo the experiments and use the shared data or some of the speakers' model parameters[1].

6.3.2 Security Evaluation Details for Conversion Experiments

The summary of evaluation details for conversion spoof is presented in Table 6-1.

Table 6-1 Evaluation Details for Conversion Experiments

Evaluation Details	Description
Category of spoof attack	Voice conversion: It is assumed that the intruder can alter his/her voice (source voice) to sound like the target voice of an arbitrary speaker using a portable device or a computer.
ToD	ToD4: The average length of each of 6 sentences used for training conversion GMM models is 1.7s which makes the total trimmed data available from each speaker to intruders 10.2s. Neither 'model parameters' nor the 'same data used for training verification models' is available to intruders.
Vulnerability	Reported in 6.3.4.
Rate of Detection by Human	Signal is detectable by human.
Scenario Implications	
Unsupervised, Remote / In Person	The intruder can speak out any requested sentence (or phrase) and alter his/her voice to resemble the target speaker's voice using a portable device or a computer. Unsupervised verification whether in person or remote is vulnerable to this type to spoof attack.
Supervised	Detection is possible in all supervised cases (remote / in Person).
Automatic Detection of Attack	See Chapter 8.
Footprints of the spoof attack	See Chapter 8.
Signal Analysis, Verification Scores, Speech Anomalies	See Chapter 8.
Usage traces	A very small delay in response may be observed but it depends on the device of operation.
Rate of spoof detection	See Chapter 8
Time to detect (When will a detection alarm be issued?)	See chapter 8 for detection of speech conversion and chapter 9 for the system design.

[1] In fact the experiments will show that there is no need to have access to training data or model parameters to conduct successful spoof attacks.

6.3.3 Variety of Experiments

The four sets of experiments reported in this chapter are chosen from several others with slightly different parameters and conditions to shed light on different problems related to the security of voice verification system developed so far. Three baseline systems were utilised for this goal. For speaker verification, the first system uses 16 cepstral coefficients and 32 Gaussians. The second system makes use of 24 coefficients and 64 Gaussians. The last system uses 16 coefficients with 16 derivative coefficients and 32 Gaussians. All the systems have a content verification module making use of HMMs. Each HMM is trained for one sentence on data available from all speakers in Subset F and HMMs consist of 60 Gaussians in 30 states (each state consists of 2 Gaussians). In the HMM experiments 16 cepstral coefficients and 16 first derivatives were used. In the verification system and for calculation of the MFCC features 29 triangular filters were used in the filterbank. Table 6-2 presents the models parameters.

Table 6-2 Parameters of Speaker and Content Verification Models

Model	Gaussians	Filters	MFCCs	Window/Frames	Used For
GMM-A	32	29	16	Hamming/16msec	Speaker Verification
GMM-B	64	29	24	Hamming/16msec	Speaker Verification
GMM-C	32	29	32(16+d)	Hamming/16msec	Speaker Verification
HMM-A	60 (30*2)	29	32(16+d)	Hamming/16msec	Content Verification

Voice conversion GMM models were trained in two different setups. In the first set, all GMMs have 36 Gaussians and only use 12 MFCC coefficients (GMM-X). The models in the second set use 20 coefficients (GMM-Y). Table 6-3 presents the parameters of the conversion models.

Table 6-3 Parameters of Voice Conversion Models

Model	Gaussians	Filters	MFCCs	Window/Frames	Used For
GMM-X	36	33	12	Hamming/25msec	Conversion
GMM-Y	38	33	20	Hamming/25msec	Conversion

Among all the parameters there are only two that are common to both the verification and conversion modules: the shape of the MFCC filters (triangular) and the shape of the time domain windows[1]. There is not a high degree of similarity between them because the number of filters and the length of frames are different. Table 6-3 presents the combination of verification and conversion models in four sets of experiments reported in this chapter.

Table 6-4 Set-up of Experiments and Models Used for Verification and Conversion

Experiment	Verification Models	Conversion Models
Ex-1	GMM-A + HMM-A	GMM-X
Ex-2	GMM-B + HMM-A	GMM-Y
Ex-3	GMM-B + HMM-A	GMM-X
Ex-4	GMM-C + HMM-A	GMM-Y

We aim to find out whether or not: differences between conversion and verification systems in terms of parameters, higher complexity of the verification system (higher number of Gaussians, higher number of coefficients) or use of derivatives can increase security of verification system.

6.3.4 Result of Baseline Experiments

In baseline experiments:

- *Content verification models* were evaluated independently and EER and high security point thresholds were acquired (for HMMs)
- In all the evaluation experiments the point of high security was chosen as the operation point for HMM models (An inclination towards security).
- Speaker verification thresholds for EER and the high security point were calculated when both content and speaker verification are in effect with HMM threshold set to the threshold adopted in step 2 (high security).
- FAR and FRR curves for the system were plotted for different speaker verification thresholds (GMM threshold).

Figure 6-2 displays the FAR and FRR curves for the content verification module. By choosing the HMM threshold at the point of high security we incline toward security rather than user convenience[1].

[1] Some experiments showed that rectangular windowing before feature extraction for training voice conversion models does not produce good results. Since the lengths of frames were different for conversion and verification the observation could not be attributed to the degree of compatibility of these two systems. Instead it may be due to spectral artifacts of rectangular windows and its incompatibility with the windowing techniques used in conversion algorithm for concatenation of converted frames.

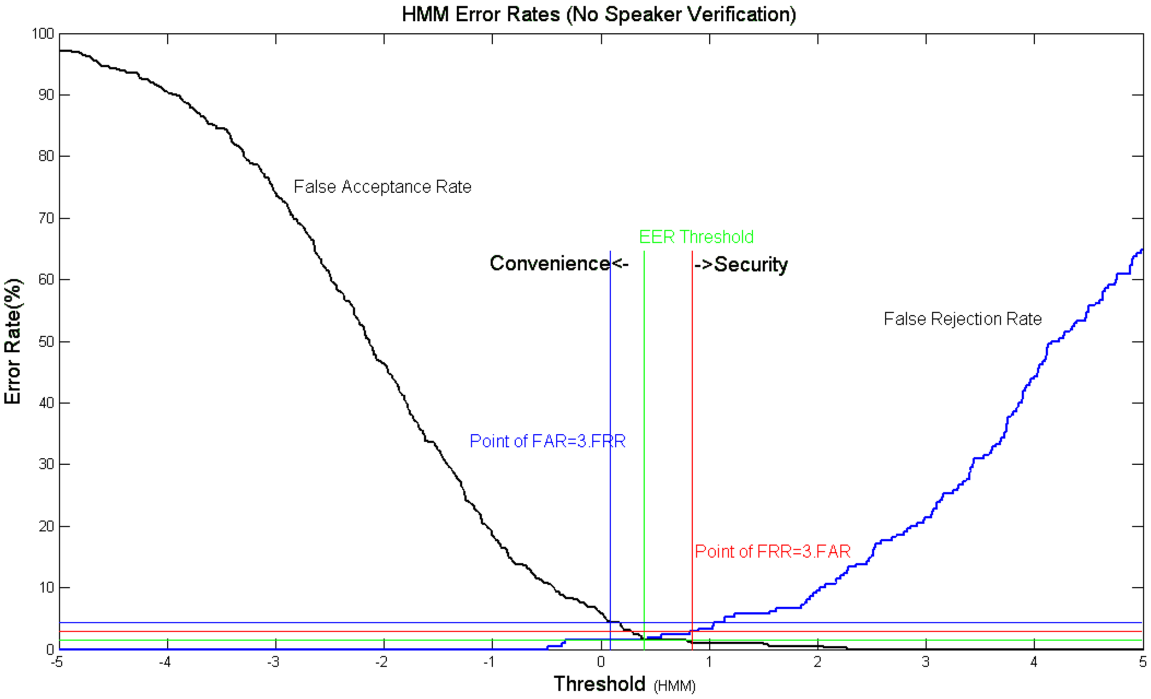


Figure 6-2 Baseline Content Verification Errors for HMM-A

From Figure 6-2 and corresponding matrices, the threshold for point of high security is chosen for HMM-A (EER is 1.43%, FRR at point of high security is 2.86% and the threshold at this point is 0.84). The next step involves fixing the HMM threshold to the adopted threshold and completing the verification and speaker verification together with various GMM thresholds.

The evaluation system should record the GMM threshold for the EER point and the point of high security as reference points. In Figure 6-3 the results of this step are plotted.

[1] Convenience used here could also be explained as lack of inconvenience caused by false rejection.

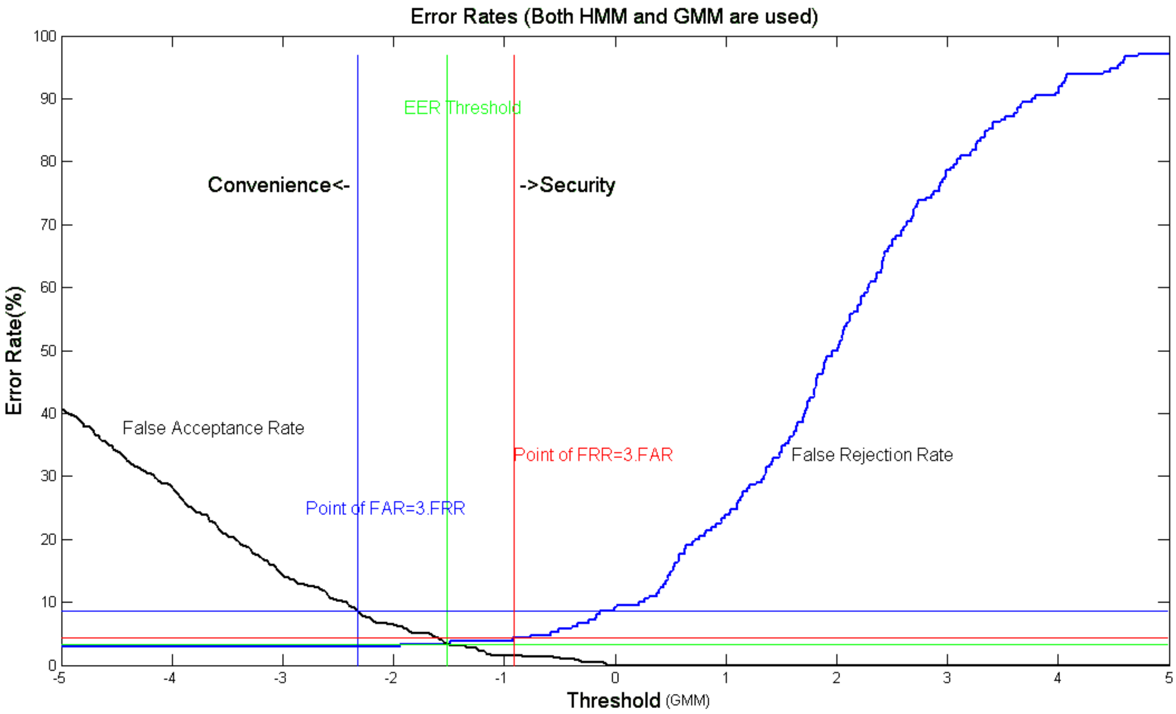
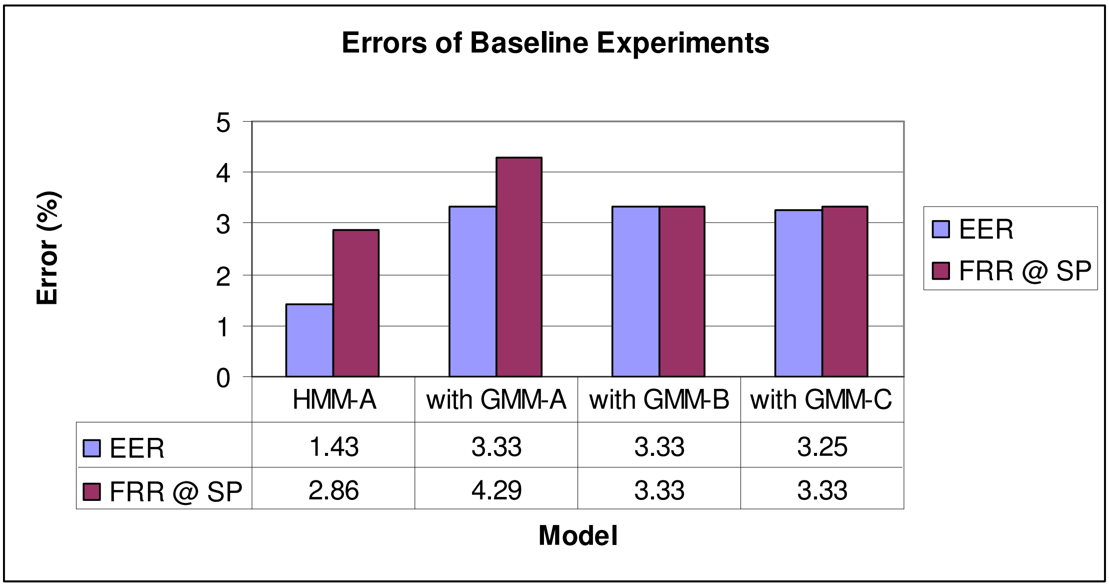


Figure 6-3 Baseline Speaker and Content Verification Errors for GMM-A and HMM-A

In the next figure (6-4) the equal error rate and false rejection rate at the point of high security is reported for all three baseline systems as well as for the content verification module.



	HMM-A	with GMM-A	with GMM-B	with GMM-C
🟦 EER	1.43	3.33	3.33	3.25
🟥 FRR @ SP	2.86	4.29	3.33	3.33

Figure 6-4 Verification Errors for All Baseline Systems with Various Models (The second column for example indicates the verification errors for the system making use of HMM-A and GMM-A)

6.3.4 Result of Conversion Experiments

1. Result of Experiment 1 (Ex-1)

The models used in the first set of experiments were specified in Table 6-3. The conversion and verification model parameters are completely different. Although the difference in frame length eliminates the doubt about locality of changes made through conversion due to the fact that window shapes and filterbank shapes are in common we can declare that '(Some) Knowledge about verification system's parameters is available to intruders' otherwise the choice would be that 'General knowledge of the algorithms employed for verification is available to intruders'.

The error rates for the first set of experiments are plotted in Figure 6-5. The baseline EER and secure point thresholds are recorded for each set of baseline systems and are specified in this figure. The baseline EER line indicates how much the FAR curve has been shifted to the right. The FAR continues to rise steadily when the threshold is decreased which implies that the rejection of converted voices is done by the GMM model (as well as the HMM model). In contrast, if the FAR reached a plateau on the left hand side (as we will see in the synthesis experiments) it would show that changing GMM threshold does not affect the error rates, and the spoofed signal is rejected by the HMMs regardless of the GMM threshold.

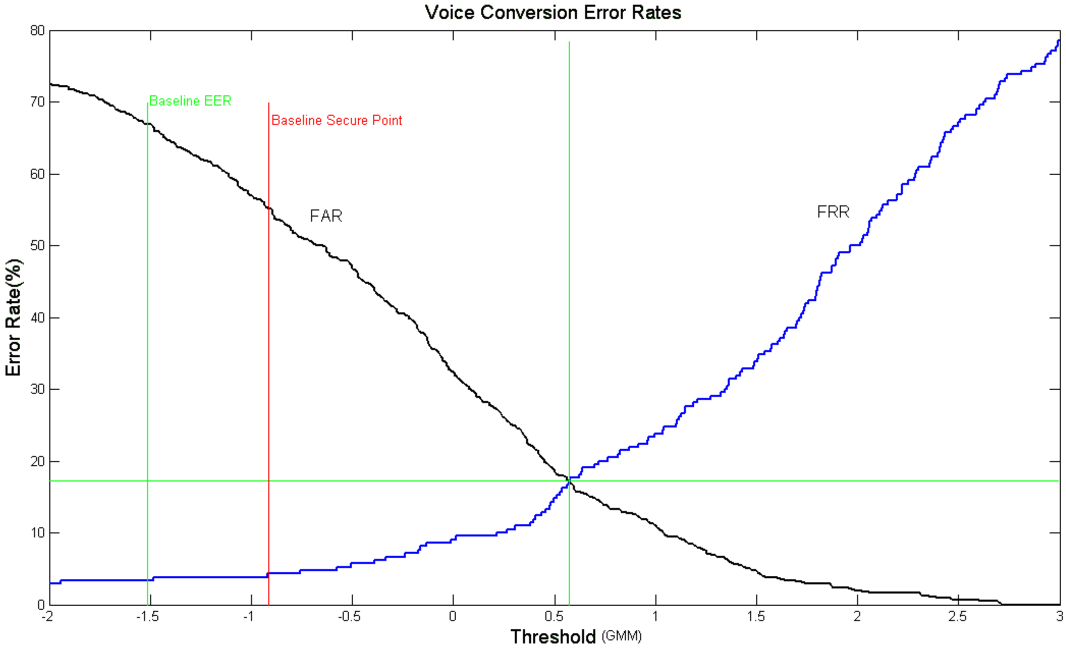


Figure 6-5 Voice Conversion Errors for Experiment One (Ex-1)

2. Result of Experiment 2 (Ex-2) and Experiment 3 (Ex-3)

Since both sets of experiments (two and three) are performed on the same verification system, we will be able to plot the error rates for both evaluation experiments in the same diagram. The error rates for these two sets of experiments are plotted in Figure 6-6. The baseline system in these experiments consists of more complicated models with a higher number of coefficients and Gaussians in the mixture (Table 6-1). The hypothesis that such a system withstands conversion spoof is apparently refuted by both experiments. At the EER and previous secure point, the conversion model with higher complexity has worked better but at the baseline EER threshold, both conversion models have produced the same results. Notice that in similar vein to the previous experiment, FAR continues to grow for both conversion models especially for lower thresholds (0 to baseline EER threshold).

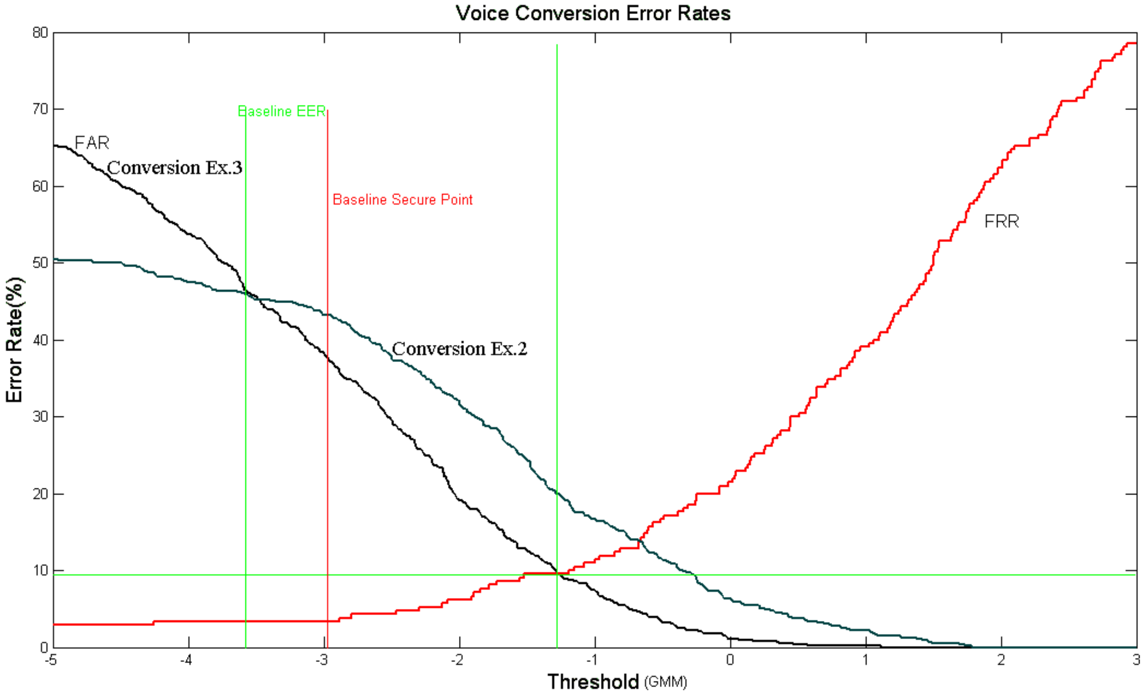


Figure 6-6 Voice Conversion Errors for Experiment Two (Ex-2) and Three (Ex-3)

3. Result of Experiment 4 (Ex-4)

In the last set of experiments we verify the hypothesis that use of derivative coefficients can guarantee the security of the voice verification system. Figure 6-7 shows that this hypothesis is also rejected.

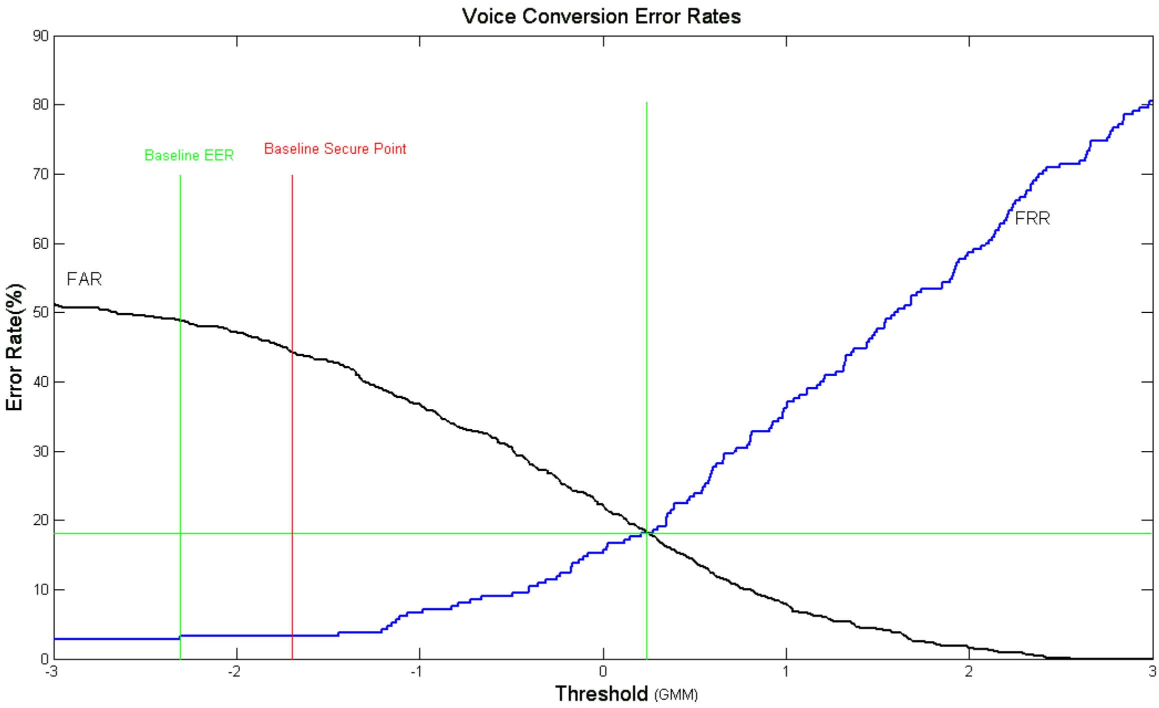
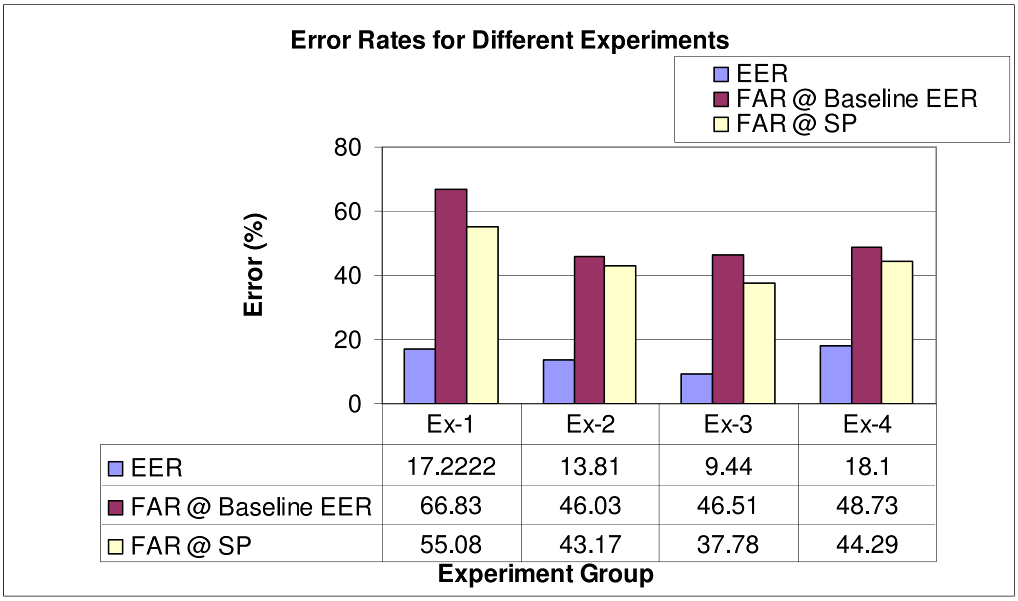


Figure 6-7 Voice Conversion Errors for Experiment Four (Ex-4)

The error rates in Figure 6-7 are significantly high even though our conversion method does not use derivative coefficients and the verification system does.

Figure 6-8 summarises the results of all four sets of experiments by specifying the EER, FAR at the baseline EER point and the FAR at the baseline secure point.



	Ex-1	Ex-2	Ex-3	Ex-4
▨ EER	17.2222	13.81	9.44	18.1
■ FAR @ Baseline EER	66.83	46.03	46.51	48.73
□ FAR @ SP	55.08	43.17	37.78	44.29

Figure 6-8 Summary of Results for Conversion Experiments 1 to 4

6.4 HMM Based Speech Synthesis

6.4.1 Goal of Speech Synthesis

While many techniques have been proposed for HMM-based speech synthesis (some of which were mentioned in the previous chapters) the goals of those techniques are different from the one pursued here. Speech synthesis has many applications in which a need for transferring data, e.g. text, to speech exists. Examples include services for people with disabilities such as speech impairment (in which the system reads the text for the user) or visual impairment (in which the system reads the text to the user). In such applications the quality of generated speech is a deciding factor in the choice of the algorithm. On the contrary in spoofing attacks the goal is to circumvent the defense mechanisms devised by the verification system.

The technique proposed here for speech synthesis targets both GMMs for speaker recognition and HMMs for speech recognition by rearrangement of genuine parts of speech available from a source speaker. It is assumed that a few sentences from the target speaker are available to the impostors. The sentences are different from the sentence prompted by the verification system. Using the HMM of the prompted sentence and re-arrangement of the speech units, a new phrase can be built that meets the score threshold for both the HMM of the sentence and the GMM of the target speaker.

It is noteworthy that this algorithm and the results obtained here could be a good representation of all concatenative synthesis algorithms which use the genuine parts of speech, e.g. phones or words from a source speaker to build a new sentence through re-arrangement and concatenation.

6.4.2 Speech Synthesis Algorithm

In the developed speech synthesis algorithm, suitable parts of a target speaker's voice are concatenated so that the generated sequence of concatenated parts sounds as the desired sentence (requested by the verification system). The method could be considered as one of the algorithms under the category of concatenative speech synthesis in which the sequence of parts and the selection is based on hidden Markov models of the desired sentence.

If O ($O = \{o_1, ..., o_T\}$) is the sequence of observations of length T and each observation is a feature vector extracted from a frame of speech ($X = \{x_1, ..., x_T\}$ is the sequence of frames in time domain), Equation 6-2 shows that any rearrangement of observations o_t in any sequence does not change the probability of observations given the GMM model. It gives the idea that by relocation of segments of speech available in recordings obtained from a speaker, we will be able to still get an intact verification score. The remaining challenge will be bypassing the content verification test.

Recalling the probability of the observations given the HMM for a path (*Q*) on the model (Equation 6-1):

$$P(O \mid \lambda, Q) = \prod_{t=1}^{T} a_{q_{t-1}q_t} \sum_{i=1}^{M} K_i N(o_t, \mu_{iq}, \delta_{iq})$$

The probability of the most probable path on the model, or Viterbi path (found through Viterbi algorithm), is commonly used as a representative for the HMM probability. Viterbi path on the HMM is used here for estimation of the average duration of emissions made by each state. After estimating the proper length of emission made by each state, one segment of speech from the target speaker's available data with this length is chosen by going through all the data available from the source speaker and sliding a window with desired length (for each state). The variable length segments are then concatenated in order after proper windowing to build the final sentence (Figure 6-9).

The following steps fully describe the algorithm used for speech synthesis:

1. Hidden Markov Models for desired sentences are trained over the data available in Subset D. It is assumed that intruders can make a generic HMM of any sentence that is requested by the verification system (These HMMs are generic and are not trained for any specific speaker which maintains the generality of the results). An expectation maximization algorithm is employed for this task.

2. Using the same training data for step 1 (from the speakers neither in test nor in training sets) the average duration of each state is estimated using the Viterbi algorithm. Since the HMMs are left to right, the duration, in terms of number of frames, could be estimated as:

$$D(s = i) = 1 / M \sum_{k=1}^{M} Num\ (q_{k,j} == i)$$

where *M* is the number of sentences used for estimation, $D(s=i)$ is the estimated length of state *i*, $q_{k,j}$ is the state *j* on Viterbi path Q_k corresponding to sentence *k*. In other words, we just estimate the length of each state by averaging the state lengths of the Viterbi path of all the available sentences.

3. Then for each state, we choose the segment with this average duration in the target speaker's available speech which maximises the probability of the segment given the state:

$$ind\ (s=i) = \arg\max_{k} P(o_k \ldots o_{k+D(s=i)-1} \mid s_i) = \arg\max_{k} \prod_{t=0}^{D(s=i)-1} \sum_{j=1}^{M} K_{ji} N(o_{t+k}, \mu_{ji}, \delta_{ji})$$

where *ind(s=i)* is the index of the start of the selected segment in the observation sequence. The duration of this segment in terms of frames is evidently $D(s=i)$. μ_{ji}, δ_{ji} are the mean and covariance of the Gaussian *j* of state *i*. K_{ji} is the mixture weight of Gaussian *j* in state *i*. In simple words a segment with length of $D(s=i)$ is selected in all the data available from the target speaker for each state of the HMM.

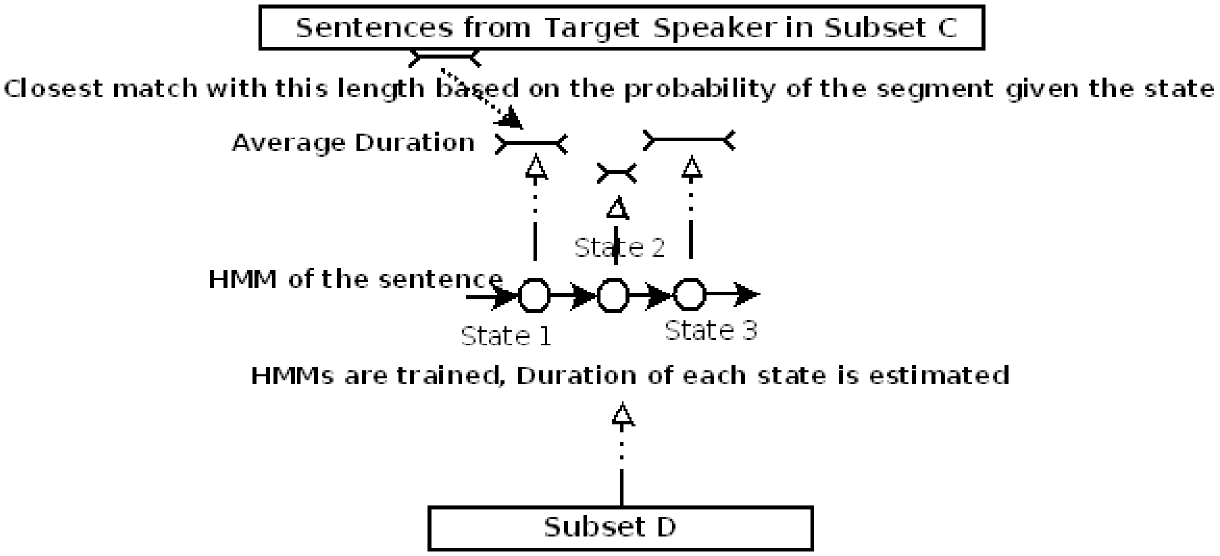

Sentences from Target Speaker in Subset C
Closest match with this length based on the probability of the segment given the state
Average Duration
State 2
HMM of the sentence
State 1
State 3
HMMs are trained, Duration of each state is estimated
Subset D

Figure 6-9 Illustration of the Synthesis Algorithm

4. Using the index of the most suitable sequence of vectors, we find the segment of speech which corresponds to this sequence of feature vectors ($o_{ind(s=i)} \ldots o_{ind(s=i)+D(s=i)-1}$ is the sequence of vectors corresponding to the sequence of frames $x_{ind(s=i)} \ldots x_{ind(s=i)+D(s=i)-1}$ of raw speech samples).

5. Finally we concatenate the segments selected for each state (in sequence) after applying hamming window to build the pass phrase prompted by the system.

Since the synthesised sentence is built upon raw samples of the speaker and the GMM is order-agnostic, the synthesised sentence passes the GMM test. Also because it follows the emission properties of the states it passes the HMM test as well.

6.5 Results of Security Evaluation Experiments for Speech Synthesis

6.5.1 Design of Experiments

The spoofing module receives the source file, target file, and the user-ID of the user for whom the conversion is intended to be performed. The target file is the result of synthesis and the spoofing module stores the synthesised speech in the designated location. The evaluation module sends all the source files directly to the verification system to plot the false rejection rate curve. The false acceptance rate curves are plotted through sending synthesised sentences to the verification system.

As described in chapter 5, Subset D was used for training spoofing HMMs for speech synthesis. It was assumed that the synthesis module did not have access to the same data which training models were built upon. The synthesis module also has access to subset C (which was also used for conversion experiments) for finding the most suitable parts of the speech corresponding to the states.

6.5.2 Security Evaluation Details for Synthesis Experiments

Similar to the conversion experiments the evaluation datasheet for the synthesis experiments should be completed appropriately. Table 6-5 presents the evaluation details.

6.5.3 Variety of Experiments

Five sets of experiments are reported in this chapter which are chosen from several others carried out on the course of research. Specification of the baseline system is given in Table 6-4. For speaker verification the system uses 16 cepstral coefficients and 32 Gaussians. In the HMM experiments, 16 cepstral coefficients and 16 first derivatives were used. In the verification system and for the MFCC features, 29 triangular filters were used in the filterbank.

Table 6-5 Evaluation Details for Synthesis Experiments

Evaluation Details	Description
Category of spoof attack	Synthesis: It is assumed that the intruder can synthesise speech in real time using a portable device or a computer which is connected to an input device e.g. a keyboard through which he/she can specify the desired sentence.
ToD	ToD4: Total trimmed data available from each speaker in Subset C is 10.2402s. Neither 'model parameters' nor the 'same data used for training verification models' is available to intruders. The average length of sentences in Subset D (for training HMMs) is 1.6 and 60 recordings are available from 20 speakers.
Vulnerability	Reported in 6.5.4.
Rate of Detection by Human	Signal is detectable by human.
Scenario Implications	
Unsupervised, Remote / In Person	The intruder can input any requested sentence (or phrase) to the device and produce the desired speech. Unsupervised verification whether in person or remote is vulnerable to this type of spoof attack.
Supervised	Detection is possible in all supervised cases (remote / in Person).
Automatic Detection of Attack	See Chapter 8.
Footprints of the spoof attack	See Chapter 8.
Signal Analysis, Verification Scores, Speech Anomalies	See Chapter 8.
Usage traces	A quite considerable delay in response may be observed as a result of inputting the prompted phrase through a device.
Rate of spoof detection	See Chapter 8
Time to detect (When will a detection alarm be issued?)	See chapter 8 for detection of speech conversion and chapter 9 for the system design.

Table 6-6 Parameters of Speaker and Content Verification Models

Model	Gaussians	Filters	MFCCs	Window/Frames	Used For
GMM-A	32	29	16	Hamming/16msec	Speaker Verification
HMM-A	60 (30*2)	29	32(16+d)	Hamming/16msec	Content Verification

Voice conversion HMM models were trained in five different setups which are specified in Table 6-7. For HMM-Z with Hanning window in time domain the shape of MFCC filters is Hamming like (as opposed to rectangular in the verification system and in all other experiments so far).

Table 6-7 Parameters of Synthesis Models

Model	Gaussians	Filters	MFCCs	Window/Frames	Used For
HMM-X	3x15	33	18	Hamming/25msec	Synthesis
HMM-Y	3x60	33	14	Hamming/13.75msec	Synthesis
HMM-Z	3x60	33	14	Hanning/13.75msec	Synthesis
HMM-Q	3x80	33	14	Hamming/13.75msec	Synthesis
HMM-R	2x45	33	14	Hamming/13.75msec	Synthesis

Table 6-8 presents combination of verification and synthesis models in five sets of speech synthesis experiments reported in this chapter.

Table 6-8 Set-up of Experiments and Models Used for Verification and Conversion

Experiment	Verification Models	Synthesis Models
Ex-1	GMM-A + HMM-A	HMM-X
Ex-2	GMM-A + HMM-A	HMM-Y
Ex-3	GMM-A + HMM-A	HMM-Z
Ex-4	GMM-A + HMM-A	HMM-Q
Ex-5	GMM-A + HMM-A	HMM-R

6.5.4 Result of Speech Synthesis Experiments

Since the verification system has not changed in all five experiments, we will be able to compare the success of all five groups of spoofing attacks in one diagram. Figure 6-10 displays the error rates for experiments one to five.

The fairly straight line on the left side of false acceptance rate curves indicates that reduction of the GMM threshold does not allow any new artificial speech produced by the synthesis module to be accepted by the system. In other words, the content verification and not the speaker verification module rejects the synthesised speech. There is a duality between synthesis and conversion problem. The content verification module in the synthesis experiments plays the role of the speaker verification module in the conversion experiments. The challenge in conversion attempts is to satisfy the speaker verification requirements and in the synthesis experiments the main objective is to produce a phrase which sounds like the required one since the speaker verification requirements are apparently satisfied.

Not surprisingly, the success of spoofing with HMM-X is limited. This could be ascribed to the longer frames and lower number of states (15) compared to the verification system. The signal synthesised by this method is coarse and does not meet the content verification requirements. Despite that, the voice verification error is over 20% around the operation point of the system.

In experiment two (Ex-2) hidden Markov models with higher number of states (60) and numbers of Gaussians per state (3) were used. Error rates exceed 40% for this experiment. It is notable that the similarity of the verification system and the synthesis system is not a deciding factor. To examine this hypothesis in Ex-3 the same set-up is used with two slight changes, the shape of the windows used in the time domain was chosen differently from that of the verification system (Hanning windows) and instead of rectangular filter-banks in the frequency domain for extracting MFCC features, Hamming like filters (smoother filters) were employed. These changes do not significantly affect the results implying that a fine synthesis method, although employing different parameters and algorithms, can be a risk to the verification system.

In experiment four, the number of states has been increased to reach 80. Some degradation in the results was observed (Ex-4) which could be because of the lack of sufficient data for training the model parameters or because of the difference between the synthesis models and verification models.

To clarify how much the closeness of the models can benefit the spoofing results, in Ex-5, hidden Markov models with 45 states and 2 Gaussians per state were used which despite some difference with the verification models, had the highest similarity among all the experiments. The error rates at the operating point of the system goes beyond 55%.

Figure 6-11 displays the error rates for all five experiments in a bar chart.

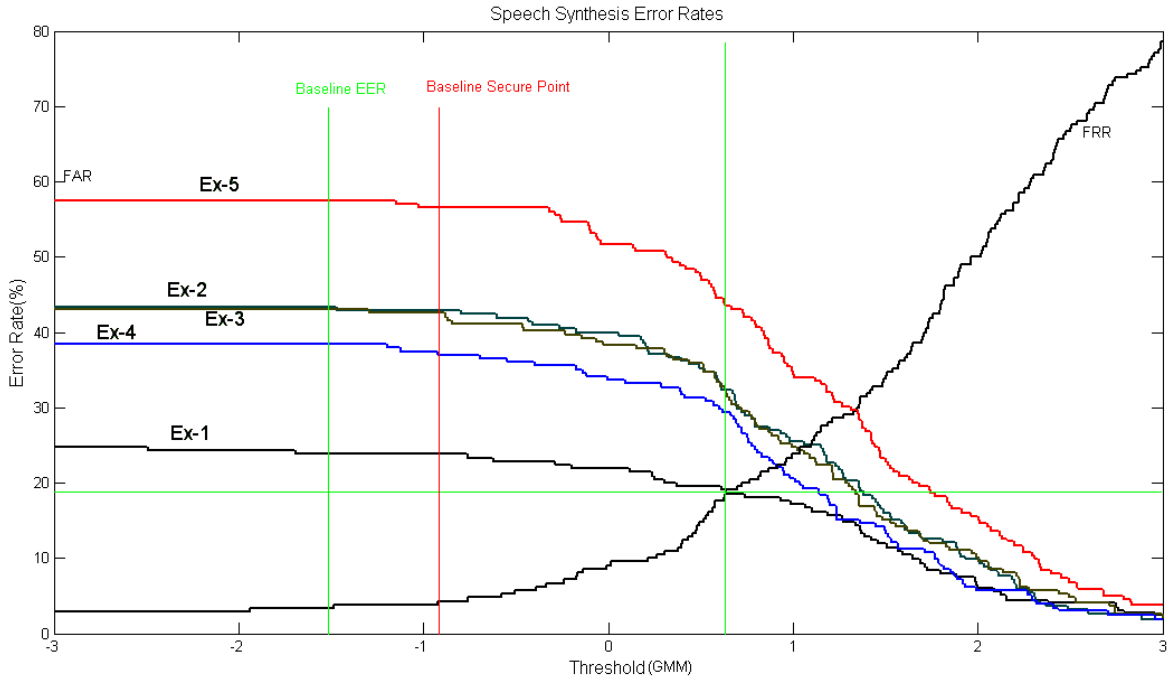
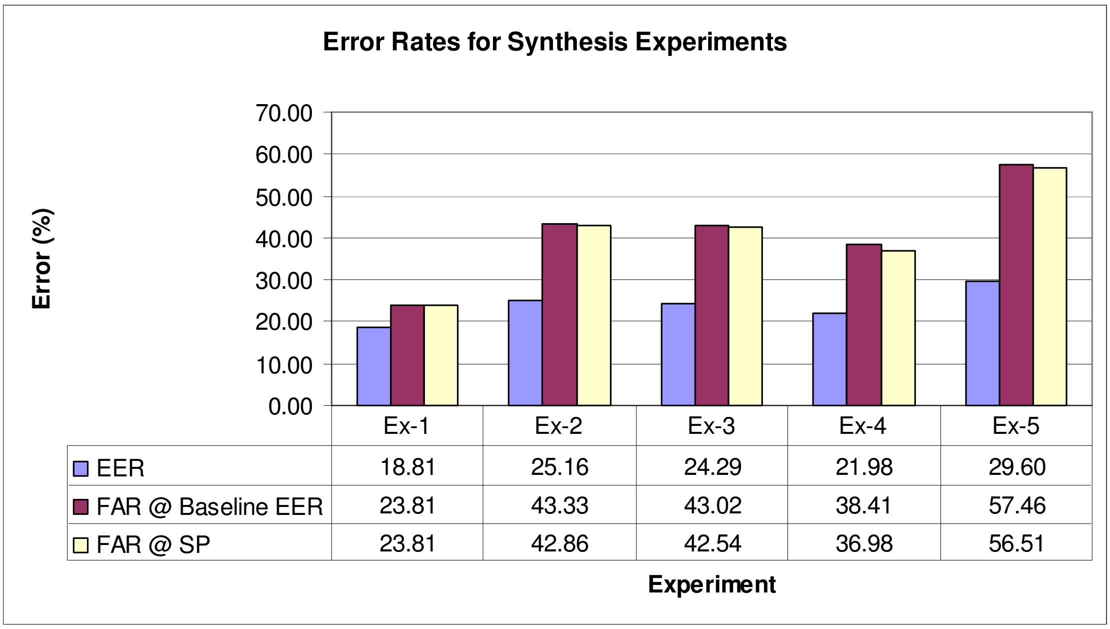


Figure 6-10 Error of voice verification system when under five sets of speech synthesis attacks



	Ex-1	Ex-2	Ex-3	Ex-4	Ex-5
□ EER	18.81	25.16	24.29	21.98	29.60
■ FAR @ Baseline EER	23.81	43.33	43.02	38.41	57.46
□ FAR @ SP	23.81	42.86	42.54	36.98	56.51

Figure 6-11 Summary of verification errors for five sets of speech synthesis attacks

6.6 A Discussion over Implications of Spoofing for Forensic Applications

The spoofing techniques presented here, and the framework explained so far, are most relevant to automatic-especially text-independent-speaker verification systems which contain a content verification module and are based on acoustic features. There are three major categories of systems which may be prone to similar types of attack.

The first types of system are text-dependent verification systems which are subject to speech replay attack, and can hardly be recommended in any practical situation.

The second types of system are the automatic systems which are based on the ideas borrowed from the field of forensic speaker identification. Such systems rely on acoustic linguistic features, for example, second and third formants in the vowels which are more robust in average quality recordings (Rose, 2006). A concatenative speech synthesis algorithm can certainly bypass the tests offered by such systems since the building blocks of the synthesised speech are raw data from the true speaker. For these systems, simpler methods such as formant based speech synthesis can also be source of significant threat. It could be added that description of a fully automatic forensic speaker verification system is not present in the literature and few forensic verification systems e.g. the one described by Botti et al. (2004) use the same acoustic features.

The final category of systems comprises those which are based on human input and both auditory and acoustic analysis of speech signal. A recommendation for such systems is rejecting all ‘low quality’ perceived speech recordings.

Ironically, the reliability of all types of systems, as well as automatic text-independent verification systems analysed so far, seems to be contingent on the detection of manipulated speech whether through ‘human interception’ or ‘automatic methods’. Even assuming that an operator or expert in forensic applications can detect the converted voice through its low quality or the synthesised voice for lack of smoothness or interaction (which is a correct assumption for the spoofing techniques presented in this chapter and also was shown in the previous analyses presented in chapter 3 and 4 based on mean opinion scores reported on all artificial methods so far) in practical systems with large number of users such as financial systems the hope for human supervision is slim or non-existent. For this reason, finding altered and suspicious voices through automatic techniques which are examined in chapter 8 and combining automated and human supervision (discussed in synthesis chapter, chapter 9) turn into crucial objectives in establishing the security of voice verification systems.

6.7 Early Discussion and Conclusion

Two types of spoof attacks in each category of conversion and synthesis were proposed, implemented and elaborated on in this chapter. Several experiments were performed to evaluate hypotheses about vulnerability of speaker verification systems against spoof attacks. These included the following hypotheses, namely:

- that more complex and simpler models may work equally in normal conditions but the former is stronger against spoof attacks;
- that the accuracy of a voice verification system translates into its security against spoofing algorithms
- that knowledge of the details of adopted algorithms, model parameters and voice data used for training models is essential to the success of spoofing attempt
- that using derivative coefficients may be helpful for rejection of spoofed signals

The results indicated that none of the above hypotheses could be accepted[1].

It was observed that there is a parallel between content verification for concatenative speech synthesis and speaker verification for voice conversion, since these two blocks pose the real challenge and line of defence against spoofing technique. The starting point in conversion is a piece of speech which passes the content verification test, and we intend to alter it so that it bypasses the speaker verification test as well. In comparison, in concatenative speech synthesis, the building blocks of speech, which may be phones, words or statistically chosen parts of speech, have the desired characteristics and the challenge is to join them properly to circumvent the content verification tests.

A huge variety in the algorithms used for voice verification exists. In the algorithm used here, normalization was applied through the world / global model as described in the earlier chapter. Variations may include cohort normalization or use of different features.

[1] Although the similarity of model parameters between the verification and spoofing systems imposes higher risks on the verification system, this similarity is not essential and the false acceptance rates are alarming even when the similarity is negligible. Neither having the same training data nor template/model parameters is required in the case of voice biometrics for spoofing which is in contrast with static biometric identifiers such as fingerprints. More accurate, more complex systems may still be prone to spoofing, and use of derivative coefficients does not guarantee that converted speech will be rejected.

By scrutinizing the scores obtained from altered speech we will notice that the score attributed to those instances by speaker and world model are lower. Auditory tests also show a lower quality of speech for converted voice. A question which links this chapter to the next two chapters is the degree of similarity between these scores and those which are obtained from recordings made in poor conditions. In other words, we aim at comparing these scores with the ones obtained by normal and poor quality samples which could be collected on the phone, by mobile and through noisy channels. The results reveal whether we will be able to discard altered voice just because of the quality of voice or by doing so that we will lose a lot of genuine cases.

The experiments in this chapter showed that the detection of manipulated speech is essential to establishing the security of voice verification systems. The verification module should be equipped with layers for detection of altered voices. The practical applications demand automatic scoring of voices based on the likeliness of being manipulated since constant and ubiquitous human supervision is not a realistic option for large systems. Therefore, chapter 8 is devoted to the analysis of the success of initiatives taken for detection and elimination of altered and spoofed voices for two categories of speech synthesis and conversion. The outcome of this investigation plays a crucial role in deciding the future of voice verification. From the assumption that detection algorithms are likely to fail, it follows that remote unsupervised voice verification will not qualify as the sole security measure for remote and unconstrained use unless a degree of risk is tolerable.

A comprehensive discussion about the suitability of using voice as a security biometric as well as recommendations for the design of systems will be offered in the synthesis chapter (Chapter 9). For example, based on the insight gained from the concatenation methods and detection ideas, it is evident that the natural continuity of a prompted signal is a critical factor in discriminating between counterfeit and authentic signals. Therefore, a system which prompts for a sequence of isolated digits, or words is more prone to attacks. Armed with the results from all these analyses we can draw conclusions about the suitability and security of using voice as a biometric in various applications and can identify the most appropriate set-up for the deployment of voice verification modules in a way that ensures that the necessary requirements for the security and performance of such systems can be met.

Chapter 7 : Assessing the Reliability of Voice Verification in Adverse Conditions

7.1 Goal of This Chapter

The accuracy of a typical voice verification system reported in chapter 5 in normal conditions is generally high. It would be tempting to utilise a voice verification module in any authentication system especially in scenarios where capturing less than two seconds of speech allowed us to reliably verify the identity of a person hundreds of miles away. Further examinations revealed that a great source of threat to the reliability of voice based authentication originated from spoofing attacks. As mentioned briefly in chapter 2, we have been able to classify different factors affecting the reliability of voice verification in the following classes: intra-speaker variability, impersonation and spoofing, channel characteristics and noises.

In this chapter, as was the case with spoofing, the adverse conditions are classified and a few less well researched factors are studied through experimental methods. These factors include fast speaking, distance at which the recordings are made and coding by a prevalent standard, Adaptive Multi-Rate (AMR), at different bit rates which is used in mobile transmission.

To study the prospect of mitigating the effect of adverse conditions a few compensation methods are tried in each case. In the first study, I have examined the possibility of improving the features and especially the filter-banks used for cepstral analysis by finding the most discriminative spectral areas for speaker recognition. Two similar studies have independently been made on this subject producing slightly different results and fairly different conclusions. We will also examine whether multi-feature algorithms and combining scores from features can reduce the effect of adverse conditions.

The results obtained in this chapter are used in chapter 8 where discarding the low quality or low score recordings is probed as a strategy for fighting spoof attacks.

In section 7.2 the influential parameters which are sources of variability in the speech are introduced and classified. The purpose of 7.3 is to lay the foundation for further use of multiple-features which focus on various spectral areas. A new metric for the evaluation of feature discrimination power is introduced and a comprehensive study on speaker specific frequency components is carried out which extends two previous studies with the same aim. In section 7.4 AMR coding is used along with voice verification and a solution to reduction of the effect of coding is suggested. A thorough research on intra-speaker variations based on previous research as well as new experiments on speaking styles and rate of speaking is reported in section 7.5. This section ends with a discussion of the practical implications of reported studies and possible solution to the problem of style of speaking. The effect of channel distortion and distance to the microphone is theoretically and experimentally studied in section 7.6. In addition to the thorough study of previously suggested solutions, the novelty of this section is providing a new re-recorded speech corpus, the use of multiple features for focusing on various spectral areas and the analysis of various score fusion techniques which causes substantial improvement in the verification results. In common with Section 7.5, sections 7.6 and 7.7 aim is to ascribe the error to specific parameters. In 7.7 several types of noise are studied along with the solutions to noise reduction and improving the verification reliability in noisy conditions including sub-band filtering and multi-feature-fusion. We will be able to draw conclusions on the reliability of voice verification in adverse conditions, assess the suitability of each proposed solution so far, and discuss why each solution works and whether or not it provides a consistent improvement in all conditions.

7.2 Classification of Adverse Conditions and Reliability Evaluation

The important factors in degrading the reliability of voice verification could be classified in the same manner that spoof attacks were classified. The same set-up for the evaluation of security could be used for the reliability analysis in adverse conditions. The message put across for uniform and comprehensive analysis of the system against all types of attacks holds here for adverse conditions. A reliability analysis is not complete until all the automated tests as well as tests on datasets collected in bad conditions are completed. While a certain algorithm may work well under one set of conditions it may not necessarily produce the same results in another. This point will be expanded on throughout this chapter and will be discussed again at the end of chapter.

Examples of the wide range of factors that can affect the reliability of voice verification is shown in table 7.1

Table 7-1 Classification of Factors Affecting Reliability of Voice Verification

Class of Variability	Subclass or Factor	Comments
Variability in prompted phrase	Type of phrase/sentence	Type of phrase for example string of digits, question, inversion, etc.
Intra-speaker	Language factors	Speaking second language, different accents or dialects
	Effect of colds/ vocal tract inflammation	Or other diseases affecting vocal tract
	Inter-session variability	Including long term effect of aging
	Eating and drinking	Immediate before or during recording
	Emotions	Anger, sadness, happiness
	Rate of speaking	Fast/Slow speaking
	Style of Speaking	Interactive, prompted, free reading, spontaneous, etc.
	Clarity of Speech	Clear or casual speech
	Speaking Effort	Vocal effort in furtive and whispered speech or when trying to project speech over noise (Lombard effect) and distance.
Channel Distortion	Microphone	Portable, Electrolyte, Carbon
	Telephone	Receiver, Analog lines
	Mobile	Transmission, coding and compression
	Medium	Including air, effect of distance
Noises	Speech like noises	Such as 'office noise' or 'cocktail party' noise, also known as speech interference
	Background and non-speech noises	White noises, color noises, short duration occasional noises (sound of horn, car passing, etc.),
	Collection device noises	Noises added due to interference with collection device (mainly electrical) e.g. mains noise

It is contended that in the same set-up explained in chapter 4, an evaluation framework should possess separate datasets for each of the above categories and regardless of how the scores are calculated by the system or the expert, the reliability of decisions have to be evaluated.

For a number of categories, the reliability tests can take the form of security tests, in which, the clean speech is given to a simulation module and after manipulation (for example adding noise) the altered speech is given to the verification module. For the others, new datasets should be collected and the evaluation module should send the sentences from those databases instead of regular datasets to the verification system.

7.3 Prospect for Improving Reliability of Verification Based on Use of More Speaker-Discriminative Spectral Features

Finding spectral components with speaker-discriminative information can be useful in the design of feature extraction methods that focus on those specific areas rather than entire spectrum. The Mel-filter-banks are optimised for speech recognition and may or may not be optimum for speaker recognition.

There are two pieces of work which have independently examined the possibility of this optimization. In the first one, Kinnunen (2004) used two datasets (Helsinki and part of TIMIT, DR7) with the sample rate of 11025 Hz, and the F-Ratio method to search for such spectral areas. His results showed that apart from the low frequencies, a speaker discerning spectral area exists between 2 and 3 kHz. Total bandwidth in his work, however, due to Nyquist's Theory was limited to around 5 kHz. He recommended the use of linear frequency warping (Kinnunen, 2004b, p. 121)[1] and rejected the idea that Mel banks are optimal for speaker recognition in all cases. In a more recent study, and without any reference to Kinnnen's research, Lu and Dang (2007) employed theoretical discussions based on the modeling of the vocal tract and formants, as well as statistical approaches based on the F-Ratio and Mutual Information to show that speaker specific spectral areas are in the ranges of 4-5 kHz, 7-8 kHz and low frequencies. The idea was further developed by the design of cepstral features with higher number of filter-banks in those areas. The results reported on the NTT-VR speaker recognition database for

[1] Kinnunen states that "We conclude that there is no globally optimal frequency warping method, but it must be tailored for each corpus. Although the results show that the mel-scale is better than linear scale in some cases, the author prefers to use a linear-frequency filterbank. In this way, controlling of the important frequency bands is more easy and the implementation is also more simple".

identification by GMM models showed some improvement over the traditional MFCC and uniformly distributed filter-banks.

Appendix 7.1 presents two available measures for gauging discrimination power along with a third new discrimination power marker. It is shown that the F-Ratio itself may be a misleading marker of discrimination power when the data is not concentrated around one point. The new metric which could be named discrimination power is suggested with the same functionality as mutual information but with some advantages which could be used, not only, for this problem but also for similar ones. Appendix 7.1 also reports the results of spectral analysis on several subsets of CHAIN and IVIE. Based on the results, 8 filter-banks are designed focusing on various spectral areas.

The verification results do not necessarily follow the expectation from filter-banks based on discrimination powers. The best filterbank is the one which focuses on the low-frequency area (Filterbank-1). Comparing the error rates of this filter-bank with those reported in chapter 5, reveals that this filter-bank works as well as Mel-bank. Verification errors for adjacent banks are consistent with the outcome of discrimination analysis when 24 coefficients are used for verification. For 16 coefficients however there are some incompatibilities. The study suggests that even though the discriminative power of spectral components is not uniform across the entire frequency domain, the gain obtained by focusing on various spectral areas is too small for verification purposes. Nevertheless these features are used in the rest of this chapter for multi-feature analysis to reduce the effect of channel distortion and noise contamination.

7.4 Effect of Coding and Compression

7.4.1 Description of the Problem

Due to bandwidth limitations, the raw speech data is usually coded and compressed at a source point and decoded and decompressed at the destination. The encoding/decoding algorithms unavoidably distort the frequency components of the speech signal and affect the reliability of voice verification systems based on acoustic parameters. From many available coding algorithms, the effect of Adaptive Multi-Rate (AMR) coding is analysed in this chapter since it is widely used in mobile transmission and is deployed in mobile handsets.

AMR is adopted by the Third Generation Partnership Project (3GPP) as the mandatory codec for third generation (3G) cellular systems[1]. The underlying coding algorithm of AMR is algebraic code-excited linear prediction (ACELP) which is a patented algorithm and was proposed by Abdoul et al. (1987). The algorithm is based on combining the results of a short term predictor (based on linear prediction of coefficients of an all pole model) and a long term predictor of the signal (using a long term predictor filter obtaining information about the pitch and long term periodicity of signal) and choice of an excitation function based on a fixed code-book and with some modifications that can be used for other signals except for speech (Carotti, 2007). AMR is one of the multi-rate coding algorithms which allows link adaptation (by the use of several pre-defined schemes for various target bit-rates). The target bit-rate is defined by a mode. Eight possible modes are defined in AMR for output bit-rates of 4.75 to 12.2 kbps. In addition to the main coding algorithm, AMR supports voice activity detection (VAD) and discontinuous transmission (DTX) which enables muting or using low bit-rates in the absence of speech components. AMR frames allow the sending of comfort noise (CN) parameters which can be used for the generation of noise-like sounds at the destination during silent periods (RFC4867, 2007). The sampling frequency adopted in AMR is 8 kHz and eight output bit-rates are 12.2, 10.2, 7.95, 7.40, 6.70, 5.90, 5.15 and 4.75 kbps, respectively (Ekudden et al.1999).

The distortion caused by encoding/decoding and the speech acquisition device together account for the higher verification errors. In the next section the amount of degradation due to coding at different bit rates is analysed experimentally. For AMR coding, the implementation made by the VoiceAge Corporation[2] is used.

7.4.2 Experimental Results

For AMR the tests Subset B, containing test recordings from IVIE corpus was used as before. After gaining access for research purposes, the VoiceAge Corporation encoder/decoder was employed for first, encoding the raw speech into AMR format and then decoding the coded files into raw speech. Since the sampling rate adopted in AMR is 8 kHz for the sake of comparability, the user models and global model were trained after re-sampling the speech recording in Subset

[1] RFC4867 - RTP Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codecs, 2007

[2] VoiceAge Corporation is the License Agent for the AMR patent pool which joined Ericsson and Nokia, the other major patent contributors to create the patent pool in 2004.

A and Subset E to 8 KHz. The models were trained on 24 cepstral coefficients and the mixtures were composed of 64 Gaussians.

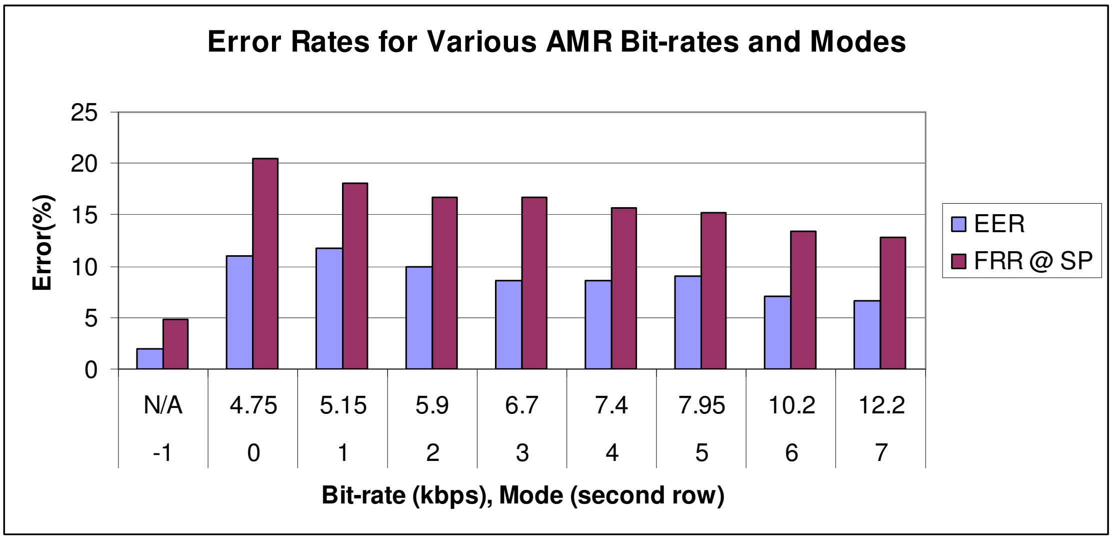


Figure 7-1 Verification Error Rate for Various AMR Bit-rates

The process was repeated for all the possible modes (various bit-rates) and for no encoding/decoding (titled mode -1) here. The error rates are reported in Figure 7-2.

Error rates (both EER and FRR at secure point) suggest that limiting the bandwidth to 8 kHz has had little impact on the accuracy of voice verification and the EER for the mode -1 (no coding) is 1.98% compared to 1.35% for previously obtained results with 16 kHz samples with the same number of coefficients and mixture size (chapter 5). On the contrary the error caused by encoding/decoding is significant and rises fairly constantly with lowering of the bit-rate. The EER error starts from 6.67% for the 12.2 kbps output and goes as high as 11.75% for mode 1 (mode 0 EER is slightly lower:11.03% but its FRR at secure point is higher).

7.4.3 Discussion and Possible Solutions

The experimental results showed that the verification errors due to coding may be significant especially for low bit-rates. A solution proposed here is training and keeping variety of speaker models for different bit-rates. Since the AMR stream and files contain mode information, based on the mode of coding, the proper models could be used for verification of that segment of speech codec under the known mode. This idea is tested in a separate experiment here.

Three types of experiment were carried out for this purpose. In the first experiment, mode two coded speech samples (5.9 kbps) were verified by regular models. As expected, the equal error rate for this experiment was 10%. In the second set of experiments mode-two-coded speech samples were verified by the user and the global models all trained on Subset A and E when both subsets (all training data) were encoded/decoded with AMR coders. The error rate obtained for this set of experiments is drastically lower (EER of 2.78%) which was close to the third set of experiments in which raw speech (8 kHz sample rate) was verified by regular models. Figure 7-2 displays the FAR and FRR curves for all three experiments. It could be noted that AMR coding without the adjustment of models causes both FAR/FRR curves to shift to the left. The curves after using specifically trained models for the target mode are close to those for raw speech (at 8 kHz) with the same EER threshold and just slightly higher EER (2.78% compared to 1.98%).

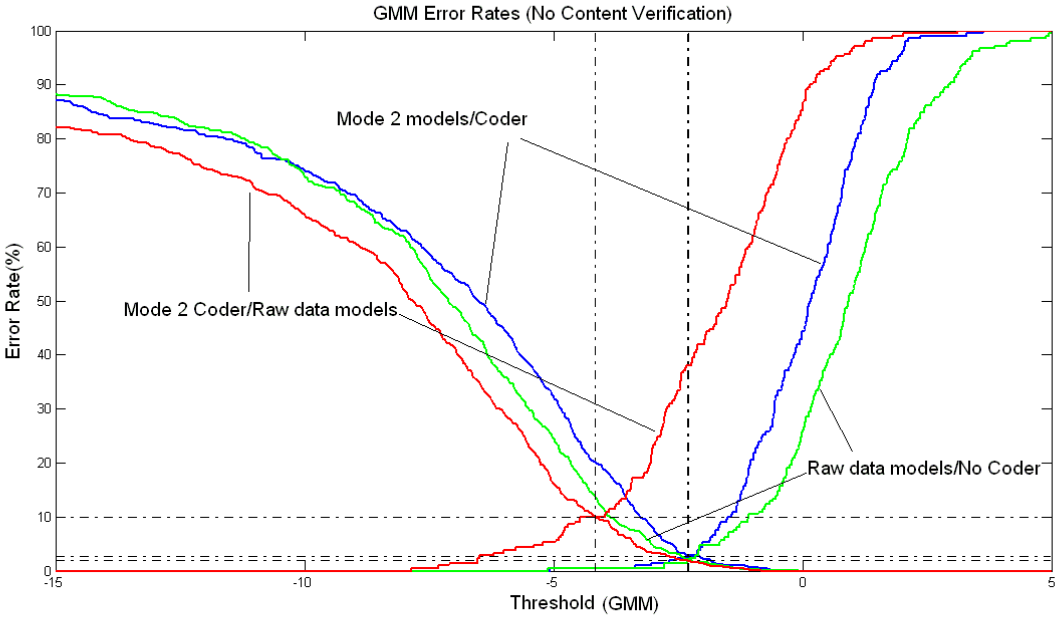


Figure 7-2 Error Rates for Three Types of Experiment: Mode Two Coded Speech verified by Regular models, Mode Two Coded Speech verified by Mode Two Trained Models and Raw Speech Verified by Regular Models

The experiments suggest that coded speech data (including the mode information) holds enough information for reliable voice verification if mode specific models are employed.

7.5 Effect of Style of Speaking and Intra-speaker Variability

7.5.1 Description of the Problem

A large portion of false decisions in speaker verification is made due to intra-speaker variability. Each speaker may adopt different styles of speaking, express different emotions in speech, change his rate of speaking, adjust the effort made to make the speech and speak with different levels of clarity at different times.

Five specific questions are proposed here the answers to have a serious impact on the strategies adopted in voice verification systems to maximise the reliability of decisions:

1. To what extent are speech and speaker verification affected by those variations?

The necessity of counter actions is decided by seriousness of the effect.

2. How can we recognise changes in the style of speaking whether through using the same acoustic properties to those used for verification or through other phonological and acoustic characteristics of the captured voice?

Unless we can identify the style of speaking we are not able to apply the appropriate measures to mitigate the effect.

3. How do these variations affect the acoustic properties of the voice, which are captured by the voice verification system?

The answer to this question would help up modify the speaker features and models based on the type of variation to minimise the errors.

4. Can we find a number of uncorrelated features with these speaking styles that can reliably work for any type of phrase and style of speaking?

5. How much enrichment of the training database with different styles is needed to enable he verification system to deal with the various styles employed by the speaker.

7.5.2 Previous Work on Evaluation of Effect of Intra-speaker Variations

In one of the earliest works on investigating the effect of style of speaking on the speech features (mainly for the purpose of contribution to speech recognition) Eskénazi (1992) analysed a number of features[1] in *careful*, *casual* and *read* speech. Eskénazi concluded that speakers use

[1] Intensity, F0 and its range, number of pauses, speaking rate, phonological changes, F1/F2 shift and number of stop releases.

different strategies to achieve the same perceived results. In a similar research Holm found out that F0 mean was significantly greater for spontaneous speech compared to read speech (2003). F0 also was larger for spontaneous speech as well as parameters of intensity range. No significant effect on F1 and F2 formant was observed. The effects were inconsistent in different subsets and Holm inferred that the speakers use different strategies when changing between read and spontaneous speech. Also 'different speakers use different parameters and different speakers use the same parameters in different ways when switching between reading and speaking'.

Quantitative analysis of style of speaking on *speech recognition* has also been also carried out in the past. Weintraub et al. (1996) experiments with different speaking styles, showed that speech recognition error (word level) is considerably higher in spontaneous conversation compared to read conversation and read dictation (read in a careful dictation style). Sturm et al. (2000) reported a dramatic impact of speech style on speech recognition performance when the styles are different: reading long sentences versus reading or saying words or short commands. They concluded that short phonetically rich utterances, even if they are read, seem to be the best possible source of data to train acoustic models.

Some studies in the literature have targeted speaking style problems in the context of speech generation[1]. Also while there are no style-robust features available there have been some efforts to include prosodic and long term information in speaker models[2].

One of the possible solutions to reducing the effect of intra-speaker variations is widening the range of styles available from each speaker in the training database. Karlsson et al.(2000) tried to 'elicit' voluntary speech by directing the users to speak in a number of different modes including normal, fast, slow, weak, strong and denasalised (pinched nose). For each mode the users were asked to read aloud 6 sequences of 6 digits. While their results do not show a substantial improvement for structured training (using various styles) they concluded that the neutrally

[1] Yamagishi et al. (2005) proposed two kind of approach for modeling speaking style in speech synthesis. By use of style-dependent models and style-mixed models they managed to produce speech in four styles of polite, rough/impolite, joyful, and sad. The context for their 42 phonemes (from one male and one female japanese speaker) included many parameters such as number of phonemes (more precisely morae, long phonemes have two morae), position of breath group, position of current accentual phrase in the current breath group and style.

[2] For example Carey et al. (1996) used pitch as prosodic features for speaker recognition and combined it with spectral features to achieve a better recognition rate on NIST-1995 database. Adami et al. (2003) achieved a better speaker recognition performance by using fundamental frequency and energy trajectories as long term information however they declared that: "sparsity of these prosodic features requires a considerable amount of training data. In our experiments, we observed that the F0/energy slope system requires more than 1 conversation half to outperform the baseline system."

trained models give a ‘varying performance’ for the different elicited speaking-styles, while the structured training causes ‘similar performance’ for the different speaking-styles. They also point out that it is important that in a system with structured training, the world (or cohort) model is created with structured training. Scherer et al. (2000) tried to induce different emotional states through a computer aided tool which induced 4 involuntary states of stress, irritation/anger, cognitive stress, anxiety and two voluntary states of positive/negative emotions and acting. While the difference between results of verification by their neutral and emotionally trained models was not significant, Their approach to elicitation of emotions was of considerable importance.

In one of the recent studies[1] Shriberg et al. investigated the effect of style and effort of speaking on verification results (2008). Based on 8 combinations of effort made by the speaker (furtive or low effort, normal effort and high effort with projection over a distance, rather than over noise as in the Lombard effect) and the style of speaking (interview, conversation, reading and oration) in the SRI-FRTIV (Five-way Recorded Toastmaster Intrinsic Variation) corpus they obtained very interesting result. For example, the EER from training on one condition and testing on the other was similar to that when the data sets are reversed i.e. the degree of mismatch rather than inherent properties contributed to error rates. Baseline EERs for furtive matched conditions were lowest for read speech and highest for conversations. The largest effect on EER came from vocal effort manipulation. The EERs varied between 0 to around 18%. Becker (2007) carried out some experiments on speaker identification for read vs. spontaneous and free vs. Lombard speech[2]. The lowest identification rates were obtained when double mismatch conditions existed.

Amano et al. (2009) tried to classify clear (e.g. when spoken in noisy environment or hyper-articulated for any reason) and conversational speech by means of a large number of features. After statistical feature reduction the number of features required for the classification was as low as 9 features for 76.14% accuracy. The final features included pause count and duration, mean and range of F0 and some metrics based on F1-F2 in some vowels.

Krause and Braida (2004) analysed global, phonological, and phonetic properties of conversational and clear speech produced at normal speaking rates. Their results show that talkers may employ different strategies to achieve clear speech at normal rates.

[1] not available at the time of the first literature review of this work and accessed after first round of related experiments.

[2] when the speaker talks louder in a noisy environment

The experiments performed on this subject are reported in the next sections with a brief overview of their purpose and design.

7.5.3 Style-Related Experiments

Appendix G presents the details of experiments carried out on speaking style and its effect on verification errors. Here only the most important points are elicited and offered:

- For the IVIE corpus the errors varied between 7 to 10% for different models, lengths and VAD parameters.
- For the Chain corpus the errors for retold passages were comparable to those of IVIE.
- Fast sentences and fast read passages (Chain) had similar errors which were significantly high (EER of over 15%).
- Inclusion of the same type of recordings (retold passages) in the training data caused reduction of errors to 2.86% (EER) from 10% (IVIE Corpus).
- Use of VAD and hand-trimming data did not lead to a significant error reduction.
- Trying to detect the speaking style by the use of global models was unsuccessful and 67.14% of cases were wrongly identified.
- Various feature-sets focusing on different parts of spectrum were unable to reduce error rates.

7.5.4 Discussion and Possible Solutions to Speaking Style Problem

Both previous research, and experimental results reported in this chapter, showed that each speaker can adopt different speaking styles and these styles have separate acoustic characteristics. The error inflicted as a result of the introduction of a non-existent style in training data during verification session is significant. For similar durations, the error may go up from 0% to 8-10% for matched vs. mismatched conditions.

Detecting the speaking style for verification purposes is not easy. Speakers use their own approaches in applying speaking style. For investigative purposes (i.e. securing criminal convictions or eliminating individuals from enquiries) the speaking style should accompany acoustic and phonological features, since those features largely vary based on the style.

Despite the fact that style-detection is a difficult and error-prone task, by elicitation and adding more speaking styles to the training data, the verification error can be considerably reduced.

It is advisable to model and treat data from a new speaking style as a new source of signal rather than a transformation of the previously collected signal. The results reported in the rest of this chapter show that by the use of multiple features and linear compensation methods, the channel effect can be reversed to a great extent. Those methods are not applicable to difference in speaking style.

Finally, it was shown that almost all the spectrum is affected by change in the style.

The best approach to dealing with the variability in speaking styles, based on the above observation and discussions, would be to narrow the range of possible speaking styles in prompted phrases by choice of better test phrases (matched with training phrases) and eliciting and including speech in different styles from the speaker.

7.6 Channel Distortion and Effect of Distance and Re-recording

7.6.1 Description of the Problem

In this section, the effect of channel distortion through a re-recording experiment is analysed. As shown in Appendix I, despite the complexity of setup in the re-recording experiments at various distances, the unwanted effect of speech signal could still be modeled as channel/noise effect.

The experiments reported in this sub-section simply aim at clarifying the following:

- Determining the effect of distance when no compensation method is used. Evaluating the impact of the distance of the recording both on spectrum and verification results
- Demonstrating the efficiency of use of multi-feature fusion in face of channel distortion
- Showing how far the re-play itself worsens the verification results (with some bearing on the possibility of speech generation and replay)

A channel (with a linear or nonlinear transform function) distorts the source signal and causes a mismatch between the spectral characteristics of source signal (real speech) and recorded signal. Since this effect is not foreseen in the user models and global (background/world) models, the score obtained after distortion is unpredictable and affects each of those models as well as impostor models differently. Part I of Appendix I offers a discussion on a theoretical basis of noise and channel distortion and the re-recording experience model. In part II (of Appendix I) available solutions to reduce the effect of channel distortion in the area of speech and speaker recognition are presented.

To study the effect of channel and success of possible solutions discussed later in this chapter, a new corpus is created by re-recording our test subset of the IVIE corpus at different distances and using two different microphones.

The remainder of this chapter is organised as follows: the next section (with a reference to Appendix 1) provides a theoretical discussion on channel and distance effect, after that previous research on channel effect is summarised and the success rate of each technique, and application in which they could be employed are expanded on.

7.6.2 Available Solutions to Reduce the Effect of Channel Distortion

A detailed account of previous research conducted aiming at reduction of the adverse effect of channels and noises can be found in appendix I. The main ideas are summarised in Table 7-2. The table serves as a guide for choosing the best approach under different conditions. It specifies what are the requirements of and what can be gained as a result of applying different solutions.

7.6.3 Experimental Results

A full description of re-recording experiments (including methods and results) has appeared in Appendix J. The results will be summarised here:

- Use of cepstral normalization and discarding the information about mean and variance of data through mean variance normalization (MVN) inflicts a cost of around 2% increase on EERs for clean speech
- The effect of attenuation by distance both on spectral components of signal and verification results is outstanding
- It is the combined effect of channel (including the non-linearity) and distance that decides the distortion, and therefore the distance effect may not be equal for two different channels
- MVN has been extremely helpful
- Uniform filters have outperformed Mel-filters for channel distorted signal
- Applying a spectral filter based on a linear approximation of the transfer function in the frequency domain for small distances can drastically improve the results. For higher distances, the amplitude of signal falls below certain thresholds and the non-linearity of the microphones and domination of noises result in high error rates.

- The comparison of rules of fusion shows that the fifth rule proposed in this study has a robust performance in all conditions[1] (for channel and noise).
- While EERs are indicators of relative performance of features and techniques, the results show that in mismatched conditions setting the threshold is a real challenge. The same results also show that re-play of a manipulated speech may not be as effective as needed by impostors. In other words, the impostors in some cases will be unable to deceive the system if they try to generate a signal, play it and capture it by a microphone due to the re-play / re-recording distortion. In fact it is a valuable gift to the impostors to allow them to send the waveforms in digital format, instead of recording it at a device[2]. On the other hand, attempts to reverse the channel effect could open doors to acceptance of spoofed signals.

7.6.4 Early Discussion on Channel Effect and Suggested Solutions

Channel effect in general, and the effect of replaying and recording at different distances was examined in this chapter. Previous solutions for reducing channel effects and their assumptions were analysed. In addition to those techniques, multiple feature extraction and score fusion were suggested for the cases in which the original signal is distorted. Many fusion rules were examined. It was observed that score fusion, based on the scored obtained from multiple features focusing on different parts of spectrum significantly improves the verification results. The improvement was substantial considering the improvements reported by other methods and compared to MVN alone.

Linearity assumptions proved disputable considering the standard deviation of attenuations (which was fairly large compared to the mean values in some frequencies) however a linear amplification function could still improve verification results suggesting that having an amplification function for each microphone may be helpful if the handset type is known.

The experiments showed that the combination of distance and microphone can inflict a substantial verification error. For example, for B microphone and assuming that normal MFCC features are used, the error rates for 30cm, 60cm, 90cm and 150cm is 8.57%, 13.65%, 16.11%

[1] The experiments on simulation of a linear channel distortion which was not detailed here also shows that while for 8 features (with MVN) the errors were 4.13% on average and the best EER was 3.81% rule 5 achieved an EER of 1.98% demonstrating a promising performance both for simulation of channel effect and real experiments with different microphones and at various distances.

[2] They will have the chance to send a perfectly crafted manipulated speech signal instead of going through the re-recording process.

and 21.98% respectively showing the effect of attenuation by distance. As the signal attenuates in the air or any media, the influence of noises and distortion by microphone becomes more notable. The low energy signals showed lower relative improvement compared to the best linear compensation.

Table 7-2 Summary of Available Noise/Channel Effect Reduction Solutions

Method	Applies To	Requirements	Improvement/Notes	Reference
Feature Mapping	Channel	Channel specific models should be trained	Little improvement	Reynold (2003)
Feature Transformation	Channel	Data from the channels	Some improvement	Colibro et al. (2006)
Speaker Discriminant Transform	Channel	Data from the channels	Some improvement	Pelecanos et al. (2006)
Speaker Independent Model Transformation	Channel	Channel specific models	Up to 20% over CMN	Teunen et al. (2000)
Cepstral Mean Normalization	Channel (and Noise)	No Data or Requirements	Significant improvement in case on channel distortion	Liu et al. (1993) and Ortega et al. (1999)
Code-word Dependent Cepstral Normalization	Channel (and Noise)	CDCN does not require stereo or simultaneous recording (unlike MFCDCN and FCDCN)	Reported for speech recognition	Liu et al. (1993)
Gaussanization	Channel (and Noise)	CGM parameters should be trained on channel data	Some improvement over CMN	Xiang et al. (2002)
Multi-Feature Score Fusion	Channel (and Noise)	No Data or Requirements	Significant improvement	This work
Normal Distribution Warping	Channel (and Noise)	No Data or Requirements	Some improvement over CMN	Pelecanos and Sridharan (2001)
SNR Dependent Cepstral Normalization	Channel (and Noise)	Stereo Data, Environment Specific Training	Reported for speech recognition	Liu et al. (1993)
CASA Based Speech Separation	Inference	No Data or Requirements	Significant improvement with GFCC and T-F mask	Brown and Wang (2005), Bregman (1990), Han et al. (2006), (Shao et al., 2007)
Independent Component Analysis	Inference	Stereo Data, Linear Mixture, Number of sources less than or equal to sensors	Significant improvement with the assumption of linear mixture	Trivedi et al. (2005) and this work

Method	Applies To	Requirements	Improvement/Notes	Reference
Kalman Filtering and Expectation Proagation	Inference	Single Channel	Different Application	Walsh et al. (2007)
MLE Based Speech Separation	Inference	Stereo Data	Some improvement	Koutras et al. (2000,2001)
GFCC Features	Inference/Noise	Single Channel	Significant improvement with T-F binary mask	Shao and Wang (2007), (Srinivasan and Wang, 2007)
Discarding Low Energy Segments	Noise	No Data or Requirements	Significant improvement especially in case of narrow-band noise	This work
Histogram Normalization and Wiener Filtering	Noise	No Data or Requirements	Little or no improvement on speech recognition	De-Wet et al. (2005)
Improvements in MFCC	Noise	No Data or Requirements	Little improvements for SNR of 10-20 dB	Ravindran et al. (2006)
Median Filtering	Noise	No Data or Requirements	Little/No improvement	Wu and Cao (2005) and this work
Replacement of 'log' in MFCC	Noise	No Data or Requirements	Good results for SNR under 20	Wu and Cao (2005)
Subband Filtering and Score Fusion	Noise (Especially Narrow-band)	No Data or Requirements	Significant improvement in case of narrow-band noise	Damper and Higgins (2003),Chen et al. (2004) and this work
Modeling Log-Amp as Max Function	Noise/Inference	Noise Thresholds' Estimates (First Ref.)	Reported for speech recognition (some improvement)	Varga and Moore (1990) , Deoras and Hasegawa-Johnson (2004) and Ghahramani and Jordan (1996)

7.7 Types and Effect of Various Noises

7.7.1 Description of the Problem

Noise reduction is a very broad topic in speech and image enhancement. Depending on the type of the noise and purpose of enhancement (e.g. better perceived quality, sharpness, smoothness, or higher recognition results) the suitable techniques vary. The goals of the discussions about noise in this chapter (and related appendices) is to reach a classification of the issues pertaining to noise contamination in voice verification applications, classification of the types of noise and producing a list of a series of noise related evaluation scenarios. Further aims include presenting and categorizing the solutions suggested for noise reduction, evaluating the effect of several types of noises on our prototype verification system for comparison purposes, and finally, trying a number of improvement strategies in each case which clarifies the extent of problem and scale of the success of compensation strategies.

Inline with these objectives, the next section (7.7.2) introduces the types of noises for which a complete evaluation test should be carried out on any target voice verification system (elaborated on in Appendix I, part 3). In Appendix I a comprehensive study of a large number of available solutions was presented in two classes of noise and inference. The solutions were classified with their references, assumptions or requirements and reported success rate in Table 7-2.

Since the details of experiments on noise are of lesser importance compared to their implications those details are presented in Appendix K (methods and results of experiment). Here in 7.7.3 the experimental results are summarised and their bearing on the directions of this study assessed.

7.7.2 Types of Noise

Noises with various characteristics can be present in the recording environment. They may be stationary noises (stochastic characteristics of the noise remain constant over the time) or non-stationary (e.g. made in unpredictable points of time: a horn or drilling sound). Type of noise which have to be included in the evaluation scheme are summarised in two categories of speech-like noise (inference) and non-speech noises (stationary and non stationary) in table 7-3. It is imperative that the evaluation framework includes various subsets with these characteristics for test purposes (or manipulation blocks similar to spoof blocks for transforming clean speech to noisy speech with the target noise profile). The distinction between different types of speech-like noises can be best understood in light of previous works (summarised in Table 7-2) and especially experimental results reported in Appendix K.

Table 7-3 Types of Noises to be Included in Evaluation Tests

Noise Type	Subcategory / Description
Gaussian Noise (Stationary)	White noises (equal energy across entire spectrum) at several SNRs.
	Color noises (e.g. Pink, Brown) at several SNRs
	Low-frequency noises
	Narrow-band noises
Non-Stationary Noises	Spike like noises such as gun-shot noises
	Special case noises in specific environments (e.g passing of the car)
	Noises with changing profiles (switching between different types of stationary noise at different powers)
Speech Like Noises/Inference	One dominant voice over speech from another speaker on one channel (linearly mixed or linearly filtered and mixed)
	One dominant voice over speech from several other speakers on one channel (linearly mixed or linearly filtered and mixed)
	One dominant voice over speech from one/several other speaker(s) on stereo/multiple channels, linear mixture
	One dominant voice over speech from one/several other speaker(s) on stereo/multiple channels, with different channel characteristics
	Identification of one (non-dominant) speaker in the speech from several speakers

7.7.3 Summary of Noise Experiments

Appendix K offers details of experiments on a variety of noises. The summary of results is presented here:

- Even for high SNRs (20dB) the EERs are significant: 12.78%, 27.78% and 13.25% for narrow-band, white noise and low-pass filtered noises respectively (for uniform features).

- MVN is generally helpful.
- Mel features work as well or better than uniform filter-banks. This is in contrast with the outcome of re-recording experiments. This has an important implication which will be discussed in this chapter's conclusion
- Use of several temporal averaging techniques such as median filters had no or little effect.
- Rule 5 of fusion which was proposed in this study, also showed a consistent performance in noise experiments.
- Subband features and filtering had little or no positive effect in the majority of experiments, despite the fact that the rule 5 again showed the promising way for fusion of scores.
- Masking effects (discussed in the appendix I) explain the noise contamination effect in the case of cepstral features. Eliminating the low energy frames is a promising approach which has little or no impact on clean speech verification and could be used invariantly in normal conditions.
- Adding similar types of noise to training data was the only solution tried in the appendix that offered a significant improvement for white noise. This move, however, worsened the verification results for some of the other noise profiles. This observation lays ground for a discussion offered in the conclusion part of this chapter.
- The inference error rates are low compared to other types of noise and eliminating low energy frames improves the verification results, especially for high signal to inference ratios.
- Independent component analysis (in the way used in the appendix which needs stereo recording) works very well when two signals are linearly combined but does not work as well when the signals go through different channels. This justifies the distinctions made in Table 7-3 for types of inference that have to be analysed separately.

7.8 Broad Summary and Next Chapter

This chapter examined the effect of adverse conditions on voice verification through GMM modeling and acoustic features. Beginning with the classification of adverse conditions, an important step towards producing a comprehensive framework for evaluation of voice verification systems, it continued with analysing the effect of five reasons for mismatch in voice conditions: coding and compression, intra-speaker variations, distance and channel effect and finally noises.

The primary goal of this chapter was classification of causes of mismatch with the aim of shaping a complete evaluation framework. Table 7-1 paved the way to this goal. It could be argued that any verification system that claims reliability should be tested with subsets comprising signals in the categories presented in the table. This is true regardless of how the verification process is carried out: by experts, by semi-automatic or automatic methods.

A secondary goal, which was determining how a typical voice verification system based on GMM models and acoustic features works in those conditions, was pursued.

Devising features focusing on various parts of the spectrum for speaker verification did not yield an outstanding improvement that could have an impact on the final conclusions. The features however were used for further multi-algorithmic fusion.

The effect of AMR coding was examined and a suggestion for reducing this effect was proposed. It was shown that coded speech data (including the mode information) along with mode specific models could be employed to reverse the effect of coding. The suggestion could be used in forensic applications, as well, for example in calculation and comparison of formants.

With respect to intra-speaker variability and style of speaking five questions were proposed and answered in the light of experiments and previous research (7.5.4). Speakers produce speech with inherently different characteristics which could not be compensated for using basic transformations. It is necessary to have enough data with different styles of speaking in the training corpora. This has a bearing on forensic speaker verification, implying that, unless the forensic method could prove reliability in a variety of styles (adoptable by speaker) it could not be used in crucial applications. In addition, in order to show for example that two spectrograms are not (or are) close enough, we should determine whether they are uttered in the same speaking style. Voice verification results were extremely poor when styles were different, causing an EER of 7 to 10% even after hand-trimming the data.

The effect of channel, both based on the previous studies in the literature and experiments was analysed and a new multi-algorithmic fusion technique was proposed and tested in the channel simulation experiments and for re-recorded signals. The algorithm offered a substantial improvement in the majority of the cases. While this chapter did not aim at re-implementation of algorithms suggested so far, in Table 7-2, a summary of available noise/channel compensation methods with the requirements and offered improvement was presented. The combination of experiments and relative improvements promised in the previous works gives an estimate of how well a system may work employing those methods. For summary of channel and re-recording experiments see 7.6.3 and 7.6.4.

Finally, types of noises which have to be accommodated in a comprehensive evaluation framework were presented in Table 7-3. The experiments demonstrated that the EERs are significant even for low SNRs. More importantly, mismatch conditions provide a shift in the score of the genuine and the impostor population which calls for setting new thresholds. Adjusting this threshold (based on the signal information and estimation of conditions) has a bearing on security problems. This will be elaborated on in the next chapter.

One last point has to be made before the conclusion of this chapter. In noise experiments, Mel features worked as well or better than uniform filter-banks. This was in contrast with re-recording experiments. Many of the algorithms suggested in the literature might offer ‘partial improvement’. Proof of practicability of those methods is contingent on demonstrating that it is possible to detect the conditions in which they are applicable (for example, for noises). This is where the necessity of conducting comprehensive evaluation tests (based on various test subsets) on the same system and in a fixed set-up for all experiments becomes more realizable. By the same token the reader can recall that adding similar type of noise to training data offered a significant improvement for white noise while verification results for other noise profiles were deteriorated.

Assuming that the thresholds could be adjusted (for the new EER of the system for the signal at hand) based on the signal characteristics or by normalization techniques, at least one solution in each condition was proposed to bring the error rates below 10%.

Chapter 8 : Prospect of Spoof Detection and Comparison of Security and Reliability Issues

8.1 Goal and Organisation of This Chapter

In chapter four, the need for spoof detection modules in voice verification systems was highlighted and an outline of such modules and associated performance tests was presented. The experimental results reported in chapter six demonstrate that the magnitude of false decision rates that a typical voice verification system could incur is large enough to justify the development of algorithms for spoof detection. Chapter seven on the other hand, revealed another aspect of the vulnerability of verification based on voice biometrics, which was poor performance under a variety of conditions, even if steps are taken to improve this as far as possible.

The main goal of this chapter is to determine whether or not it is possible to detect counterfeiting in the speech signals and how the detection process affects the overall performance of the system.

The delicate balance between dealing with adverse conditions on the one hand and being able to detect spoof attacks on the other, is confounded when setting an acceptance threshold for the entire system so that it performs satisfactorily in different situations. The threshold setting option when used inappropriately can be exploited by perpetrators to attack the system successfully.

On the other hand the decision to relax any of the system's rules under noisy conditions based on the characteristics of received signal has the consequence of allowing the impostors to simulate those conditions and break into the system more easily. The preceding types of vulnerability necessitate balancing the detection of spoofed signals against massively rejecting the noisy or channel affected recordings.

Three families of spoof detection methods are explored in this chapter for detection of synthesised and converted voices generated by the algorithms described in chapter 6. To extend the research a small dataset was collected from the speech generated by five commercial speech synthesis systems. The aim of the experiments on this dataset was to determine the limitations to spoof detection when elaborate and state of the art algorithms for synthesis are employed. This limitation, however, changes over time as spoofing and spoof detection techniques improve.

The rest of this chapter is structured as follows. In sections 8.2 and 8.3, after a short description of how a spoof detection block could be described and characterised, possible approaches towards spoof detection in the context of voice verification are presented and classified. Following that, previous research on the detection of synthetic speech is reviewed in 8.4[1]. In 8.5 an outline of the purpose and nature of the experiments carried out on this topic is presented. In 8.6, the scores obtained under adverse conditions and by spoofed signals are explored and possible threshold related decisions are discussed. In 8.7 and 8.9 consistency analysis and discontinuity detection are analysed as two techniques for the identification of counterfeiting in speech signals. The chapter ends with an initial assessment on where automatic voice verification stands on the security scale in view of all preceding discussions.

8.2 Description of a Spoof Detection Module/Block

Chapter 4 demonstrated how spoof detection blocks can be developed independent of the system and act as shields for the core system.

In this chapter we will see the mathematical description of the spoof detection module and its effect on the system's performance.

For a signal *S* consisting of a sequence of samples, a spoof detection block, and the verification system, could return a single number representing the probability of the signal *S* being spoofed rather than being natural. This single number can be compared with a threshold to inform a decision about the signal (whether to reject the signal regardless of its score given by the verification system or to combine two scores). In this way the spoof detection scores could be interpreted as a new source of information which its score/decision should be combined with the verification score. Therefore all known fusion techniques can be employed.

In this chapter we assume that although the spoof detection blocks can communicate with the verification system to receive the verification scores, the relation between these two is based on decision fusion.

[1] The reason that previous research is not discussed at the start of the chapter before description of the detection blocks is that the legacy of research presented here is mostly directed towards other goals and there is little research in the literature on spoof detection in the way formulated here, therefore it is necessary to lay foundation for the discussions first.

The spoof detection block rejects the hypothesis of being natural based on evidence or observation (here the signal) if the likelihood surpasses a threshold:

Equation 8-1

$$\frac{P(H_1 \mid S)}{P(H_2 \mid S)} = \frac{P(S \mid H_1)}{P(S \mid H_2)} \cdot \frac{P(H_1)}{P(H_2)} = LR_{SP} \cdot \frac{P(H_1)}{P(H_2)}$$

where H_1 is the hypothesis that *S* is counterfeit and H_2 is the hypothesis of it being natural (regardless of whether it is from the intended speaker or not). Making any assumption about prior probability of H_1 and H_2 is impossible and the system should aim at estimating the probabilities given the hypotheses. It is helpful to think of the spoof detection module as a function which assigns a number to each signal representing the likelihood ratio of being counterfeit to being natural:

Equation 8-2

$$LR_{SP}(S) = f(S)$$

Also let's assume that the score given by the verification system to signal *S* is x: $x = g_i(S)$ where *i* is the index of claimed speaker.

The spoof detection block, based on extraction of a number of features from *S,* assigns a number *f(S)* to the signal which if it exceeds a threshold, the signal is rejected as spoofed. The point that is going to be made here is that this threshold cannot be globally set independent of *x,* and if it is chosen independently then, the false positive and false negative error of spoof detection blocks vary for different values of *x*. The distribution of f(S) across the *x* the score of utterance (assigned by the system) is not uniform due to the dependency between the features extracted for speaker verification and features used for spoof detection. If we can show that $f(S)$ and $g_i(S)$ are independent, then we can declare that by setting the spoof detection threshold of the false positive error (probability of falsely deciding that a sample input is counterfeit when it has been natural) and false negative errors (probability of falsely deciding that a sample input is natural when it has been counterfeit) of the entire population is the same as every interval of *x*[1]. Otherwise we should specify *fp(x)* and *fn(x)* with the thresholds set for each interval of *x*. Without independence however we can set a global threshold but the global *fn* and *fp* may not

[1] If two variables are independent: P(X|Y)=P(X) which means that the distribution of $f(S \mid x)$ is the same as $f(S)$. The required condition here is that $f(S \mid x)$ should be the same for all values of *x*.

accurately determine the FAR and FRR of the combined system and module (yet we are still able to 'test' the system with certain thresholds for the verification system and the module and by a dataset).

If we assume that the detection block has the false positive error of $fp(x)$ where x is the score the verification system assigned to the sample, the overall false rejection error of the system is (f_G is the probability distribution of genuine speakers):

Equation 8-3

$$FRR(T) = \int_{-\infty}^{T} f_G(x)dx + \int_{T}^{\infty} f_G(x).fp(x).dx$$

Similarly (f_I is the probability distribution of genuine speakers):

Equation 8-4

$$FAR(T) = \int_{T}^{\infty} f_I(x) fn(x).dx$$

Rewriting FRR (*T*):

Equation 8-5

$$FRR(T) = \int_{-\infty}^{T} f_G(x)dx + \int_{T}^{\infty} f_G(x).fp(x).dx =$$

$$\int_{-\infty}^{\infty} f_G(x)dx + \int_{T}^{\infty} f_G(x).(fp(x)-1).dx = 1 - \int_{T}^{\infty} f_G(x).(1-fp(x)).dx$$

we can observe that FRR is increasing and if thresholds (*th(x)*) are chosen so that *fp* is constant over the values of score (*x*), then: $\lim_{T->-\infty} FRR(T) = fp$ meaning that since the FRR is ascending it is (for all thresholds) values higher than *fp* which means that the false positive error of the spoof detection block decides that minimum overall error of the system.

The FAR falls significantly, which may cause the system to have no equal error rate (for all threshold values the FAR may fall below the FRR).

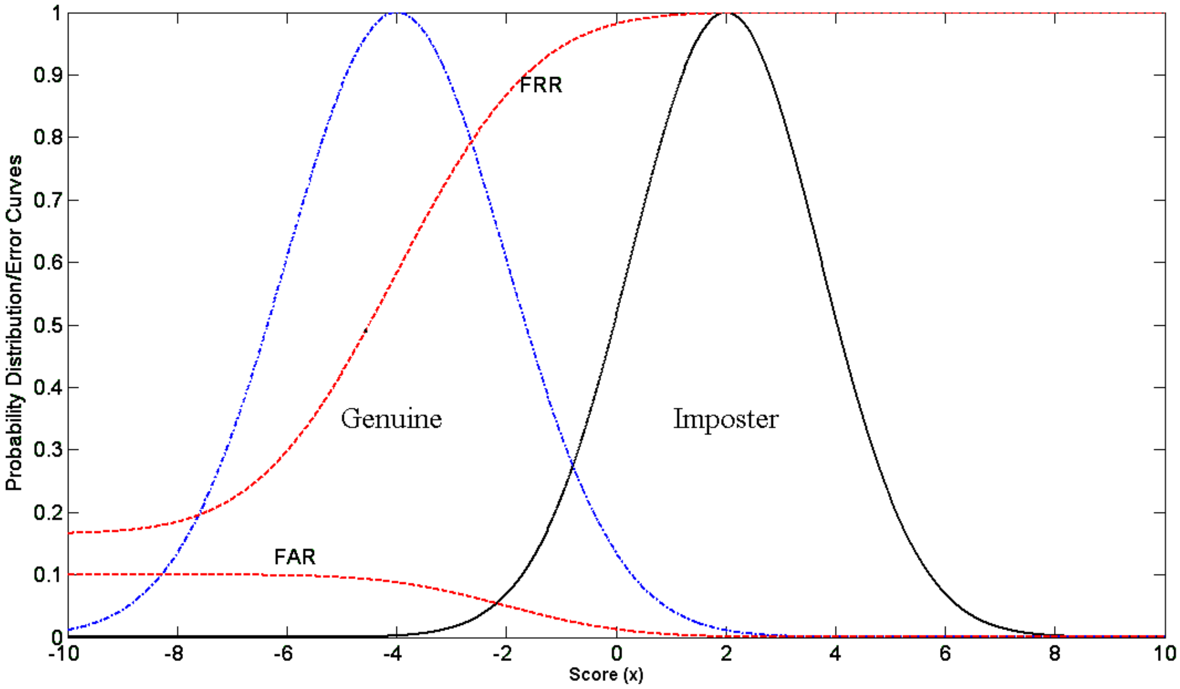


Figure 8-1 Influence of Spoof Detection Block on System's FAR and FRR when False Positive is High

From the perspective of the verification system, the definition of the spoof detection block can be made by specifying the $fp(x)$ and $fn(x)$ as a function of threshold. Alternatively it could be shown that $f(S)$ and $g_i(S)$ are independent and the threshold *th* can globally set with single values of *fp* and *fn*. Another way of characterizing the spoof detection block is by specifying the likelihood values it assigns to the cases in the authentic and spoofed subsets. Consequently, the evaluation system can combine the verification scores and these likelihood (of being spoofed) values to set a suitable threshold for each unit. The problem of selecting single values for *fp* and *fn* or of assigning the values to the cases is that the final FAR and FRR may be somehow different from what we expect and could only be determined by a joint evaluation of two systems.

8.3 Possible Approaches to Spoof Detection for Speech

Seven approaches to spoof detection for speech signals are suggested in this work. Each of these approaches focuses on one aspect of speech manipulation to make the detection of an altered signal possible.

1. Discontinuity Detection: Most synthetic methods are based on joining authentic parts of speech (with or without adaptation of pitch and other characteristics). Since the parts of speech, for example triphones, are authentic themselves, the simplest way (if not the only chance) to detect the spoofed signal is by identifying the discontinuity at the points of concatenation. For conversion techniques (such as the one we have developed) which work on fixed or variable frames, the concatenation effect could be present at the edges of the frames. This technique and many of its variations are intensively explored in this chapter.

2. Non-Speech Rejection: As specified in chapter 4, Common Criteria's biometric evaluation methodology-BEM-identifies the risk of a weak template: "a template created from a noisy, poor quality, highly varying or null image, which typically has a higher FAR than other templates" (p. 33). This category, however, is more general than the one mentioned in BEM and may be independent of the template. A verification system strives to reduce the false rejection error related to channel and noise effects. The relaxation and change in the verification process happening due to an estimation of the SNR ratio, could pose a high risk to the system. For these reasons, a noisy or highly varying speech sample, which may be produced by simpler spoofing techniques and does not show the characteristics of speech, should be identified and rejected. Non-speech rejection algorithms based on spectral characteristics can assign a score of being speech-like to each time slice of the signal and sum up these scores over the entire duration of signal.

3. Altered User/Global Model Scores: The manipulated speech signal might receive lower global and user model scores (from speech and speaker models) due to the fact stated under approach number 2,above. Nevertheless the difference between these two scores could still be high and surpass a threshold. This is the rationale behind the third and fourth suggestions made in this chapter for spoof detection. In other words, the third approach is the same as the second approach but implemented in a feature domain in which being speech-like is determined by the features and the models estimated on the population present in training (global or world models).

Table 8-1 Suggested Methods for Spoof Detection

Method	Applies to	Description
Discontinuity Detection	Synthesis and Conversion	Based on finding discontinuity in features/signal characteristics on time axis
Non-Speech Rejection	Conversion only	Consists in Identifying non-speech (spectral) patterns in each time slice, analysed over the entire recording's duration
Altered U/G Model Scores	Conversion only	Relates to comparison of user and global scores. While their difference may be still high the absolute values of user and global GMM scores may have altered due to conversion
Altered HMM Scores	Synthesis only	In synthesis the same phenomena could be encountered for HMMs (altered scores of all competing HMMs)
Inconsistency Detection (GMM)	Conversion only	Based on the fact that manipulation of speech distorts speaker recognition scores obtained from different algorithms
Inconsistency Detection (HMM)	Synthesis only	Based on the fact that manipulation of speech distorts speech recognition scores
Time Domain Speech Evaluation	Synthesis and Conversion	In contrast with non speech rejection these methods work on the changes of signal characteristics to check if they exhibit temporal speech patterns over the time (but not in each time slice)

4. Altered HMM Scores: The GMM scores of concatenative speech synthesis are almost the same as GMM scores of its authentic parts of speech from which it is constructed. Therefore the synthesised speech (by concatenation) does not get a low score based on GMM models. But due to the joining process, the same effect described in approach 3, above, could be observed for all HMMs (lower HMM scores regardless of difference between rival HMMs).

5. Inconsistency Detection (GMM): The fifth approach proposed and developed in this chapter is based on the assumption that a conversion technique employs a feature set and a model to manipulate speech (to make it sound like a target speech). The manipulated speech gets higher scores from the target user and global GMMs and lets the impostor break into the system.

Despite that, the conversion technique does not have control over all of the characteristics of the signal and can not consistently alter all the scores assigned to the counterfeit signal. Therefore, different models (based on different features) which normally output similar or correlated scores for natural voice produce more uncorrelated scores for converted speech. We elaborate more on this approach in 8.7.

6. Inconsistency Detection (HMM): For concatenative synthesis, inconsistency detection among the HMM scores can be used in a similar way to approach 5 for spoof detection.

7. Time Domain Speech Evaluation: This approach is similar to the second approach in that both seek to determine whether or not the signal exhibits the characteristics of natural speech. In contrast, it operates on the signal as a whole and can be applied in conjunction with the second method.

One of the most prominent features of speech is its pitch or the fundamental frequency for voiced segments. Natural speech demonstrates a smooth flow in fundamental frequency. The number and ratio of voice and unvoiced parts and the duration of these parts can also be used for evaluation of how speech-like the signal is. A warning, however, should be given that for noisy speech many of the above features could be estimated with a high distortion causing a high false rejection rate for noisy signals. Fundamental frequency will be one of the features used in this chapter for discontinuity detection. However use of it can go beyond the analysis of abrupt changes, and may include the accepted range or its normal flow for a particular speaker.

8.4 Previous Work on Detection of Spoofed and Synthesised Speech

Despite the irrefutable need for spoof detection, this topic has surprisingly attracted little attention in comparison with other areas of speech processing. One crucial problem in this area is the dependence of results on the synthesis techniques which are normally developed by the same research group as that focusing on spoof detection. The results are therefore difficult to generalise. This fact calls for the preparation of a standard database of spoofed speech as highlighted in chapter 4.

Additionally, most of the related research targets objectives other than spoof detection. The closest objective to spoof detection in the literature is quality evaluation of synthesised speech by use of distance measures. This objective has a perceptual and psycho-acoustic dimension to it which differentiates it from automatic spoof detection. Despite that, the distance and discontinuity measures developed for this purpose are also helpful for spoof detection and take us further along the road.

Incorporating pitch information into the decision is one of the potential ways to reduce synthesis risks. With this aim, Masuko et al. (2000) used HMM models to evaluate the effectiveness of pitch data for the rejection of synthesised speech. They carried out experiments on systems with and without employment of pitch information both for synthesis and verification. While there was some reduction in the error rates especially when the synthesis algorithms did not use pitch data, and the verification system did, they stated that "pitch information is not necessarily useful for the rejection of synthetic speech, and it is required to develop techniques to discriminate synthetic speech from natural speech" (Masuko et al., 2000, p. 1). They also noted the problem of adjustment of the decision threshold which yielded high FRRs.

The rest of the relevant research discussed here mainly pertains to the analysis of the quality of synthesised speech by calculation of distance measures.

In speech synthesis a 'target cost' and a 'join/concatenation cost' are assigned to each candidate unit. The target cost is a weighted sum of differences between prosodic and phonetic parameters of the target and that of the candidate unit (Pantazis et al. 2005). It shows how close the context and the unit are. On the other hand, join/concatenation cost is a cost assigned to the successive units based on the degree to which those adjacent units are matched and can smoothly be concatenated. Various parameters can be used to define this cost function such as the difference between energy, spectral characteristics, and fundamental frequency or a combination of all.

Stylianou and Syrdal(2001) carried out a two step test comprising of a perceptual test in which listeners had to report whether or not there was a discontinuity in the signal and a second step involving calculation of twelve distance measures They observed that the highest rate of prediction of discontinuity was obtained by the Kullback-Leibler distance (Kullback and Leibler, 1951) on the FFT-based power spectra and the second by the Euclidean distance between MFCCs. The best distance measure predicted only 37% of the audible signal discontinuities at a false alarm rate of 5%. For absolute difference of pitch around the concatenation points the detection rate was only 19.981%. Wouters and Macon however in 1998 in different settings had found that the Euclidean distance on mel-LPC-based cepstral coefficients was a good predictor of perceptual scores. They had also reported that little improvement was achieved by inclusion of delta features in the distance measure (Wouters and Macon, 1998). In another paper published that year, Klabbers and Veldhuis (1998) reported that the Kullback-Leibler distance offers good correlation with listening experiments when used with the LPC based spectra. Their research was confided to five vowels. Vepa et al (2002, p.1) with the assumption that "spectral discontinuities are particularly prominent for joins in the middle of diphthongs" focused on such joins in the sentence. They had one concatenation point in each sentence in the middle of one of five diphthongs. They proposed weighted sums of the distance metrics of various spectral features instead of a single distance metric.

On the opposite front, there is a continuous stream of effort aiming at improving the perceived smoothness of synthesised speech. With the hypothetical widespread use of speaker verification systems these efforts may be directed toward spoof perfection. Just as examples, Plumpe et al. (1998a) enhanced their HMM-based synthesis system with a smoothing technique based on use of delta coefficients and tested their technique by a subjective preference test which demonstrates the perceptual effectiveness of the method. They also evaluated eight possible changes in their speech synthesis engine to determine the cause of quality degradation (Plumpe and Meredith, 1998b). These eight changes were results of altering the system in three dimensions of acoustics (natural or synthetic units), pitch and phoneme duration. They observed that the pitch generation component of the engine had the largest effect on quality of the speech.

Among the recent works on discontinuity detection Pantazis et al. (2005) approach seems to be most promising. Pantazis proposed two sets of features for discontinuity detection. The first set of features was based on non-linear harmonic modelling of speech signals which was developed by Stylianou (1996) throughout his PhD. The coefficients of this type of modelling (amplitude of harmonics and first derivative) were used for discontinuity detection. The second set of features were based on separation of speech signal into AM and FM components around centre frequencies based on Maragos et al. method (1992). They reported better AM+FM coefficient results for discontinuity detection, while the AM and FM components each worked worse than harmonic model's coefficients. By the whole set of features (Harmonic parameters, AM, and FM) and linear discrimination they reported a detection rate of 56.35% (at 5% false alarm) which showed a significant improvement over previously published results on the same database achieved by the LPC based spectrum and the Kullback and Leibler distance (defined later in this chapter).

This is notable, however, that in these studies the point of concatenation is known and there is one concatenation point in the phrase which makes the analysis easier. The effectiveness of these measures for synthesis detection in large phrases with many unknown points of concatenation needs further analysis. The distance measures introduced here will be the basis for a comprehensive discontinuity detection study for spoof detection presented in section 8.8.

8.5 Design and Purpose of Spoof Detection Experiments

The main goal of the spoof detection experiments in this chapter is the evaluation of the success of these initiatives in the reduction of false acceptance errors and the consequences these methods may have on false rejection rates. The issues around threshold setting which were discussed in 8.2 will be demonstrated through experiments.

Three categories of spoof detection techniques are examined in this chapter: rejection of low score recordings, consistency analysis and discontinuity detection.

To put the results from different sources into one picture speech data was organised into 8 subsets. The first subset comprises of clean data used in normal verification experiments (Subset B). The second subset is a noise contaminated subset with white noise at SNR of 15dB. The third subset consists of sentences in hand-trimmed retold subset of IVIE. The fourth subset consists of hand-trimmed read passages and the fifth one consists of solo sentences from the Chain Corpus. These five subsets constitute the genuine portion of data. Since the hand-trimmed retold passages have discontinuities in the signal (due to the manual removal of silences) they are not used in the discontinuity detection experiments.

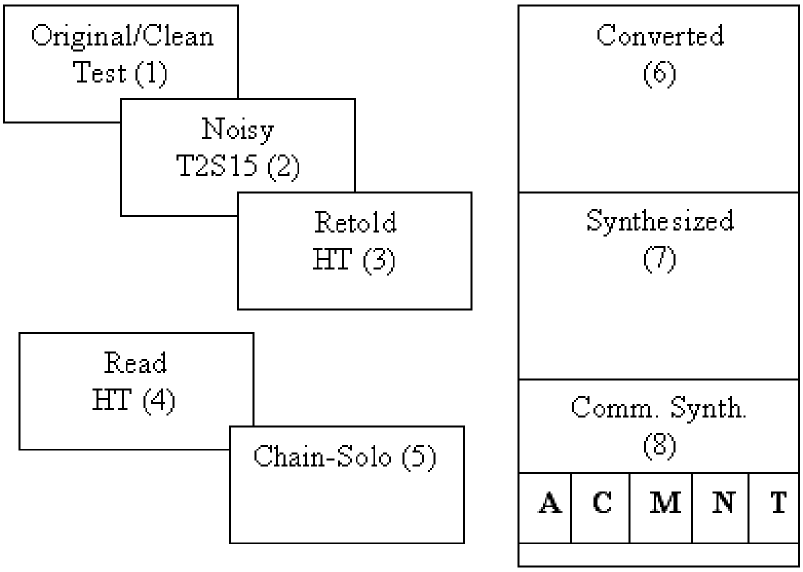


Figure 8-2 Subsets Used in the Spoof Detection Experiments (T=AT&T Natural Voice, A=Acapela BrightSpeech and Elan, C=Cepstral Voices, N=Nuance RealSpeak and M=Microsoft Engine)

Three spoofed subsets were made based on the algorithms developed in chapter six as well as data collected from commercial speech synthesis systems. Subset 6 consists of 210 converted sentences based on GMM-Y in chapter 6 (38 Gaussians, 33 filters in bank, 20 coefficients, Hamming windows of 25ms). Subset 7 includes 210 sentences synthesised based on HMM-Y in chapter 6 (3 by 60 Gaussians in HMM, 32 filters, 14 coefficients with derivatives, Hamming windows of 13.75msec). In addition to data from the two algorithms described in chapter 6, synthesised speech from five commercial speech synthesis engines[1] with maximum of 129 samples, was collected and placed in subset 6. Apart from Microsoft engine for which the sentences are generated the rest of the sentences are hand-trimmed from sample sentences provided by the developing companies for demonstration purposes and may have better quality compared to the average of all possible sentences the engines produce.

The evaluation of spoof detection techniques by these three types of algorithms is the subject of next three sub-sections.

[1] These engines were AT&T Natural Voice, Acapela BrightSpeech and Elan, Cepstral Voices, Nuance RealSpeak and Microsoft Engine.

8.6 Rejection of Low-Score Speech and its Relation to Adverse Conditions

8.6.1 Description of the Approach

Acoustic mismatch despite normalization and compensation techniques causes a shift of error curves towards lower thresholds, which means that the previously set thresholds result in very high false rejection rates (and very low false acceptance rates). This finding suggests that the lower thresholds need to be applied in the interest of convenience.

On the contrary the spoofing techniques cause the false acceptance rates to move in the opposite direction on the threshold axis. The security of the verification system at the thresholds set with the aim of convenience of users will be largely at risk.

Two hypotheses are tested in this section with regard to the spoofed speech generated in chapter 6. The first hypothesis is based on the third and fourth approaches, listed above, proposed for spoof detection which assumes lower scores (by individual global and user model despite the high difference) for spoofed signals. We will determine whether or not our generated speech follows that pattern. The second hypothesis is that converted voice and synthesised voice act as noisy or highly varying patterns. The hypothesis suggests that similar to the noise and channel effects the error inflicted by spoofing is the result of transformation in FAR and FRR curves and not a similarity caused by the spoofing algorithm. In other words, any algorithm that could distort the scores could inflict some error on the system. The relation between verification in unfavorable conditions and in face of spoofing attempts is examined under this hypothesis. A holistic approach to both issues is adopted here and the effect of decisions in favour of one side on another is analysed.

8.6.2 Comparison of Scores of Noisy and Spoofed Signals

The scores of GMM (24 coefficients, 64 Gaussians) and HMM (16+d coefficients, 30 states, 2 Gaussians per state) models both for user models, and global models (rival models for HMM) were calculated for original (clean), noisy, converted and synthesised sentences. The results are shown in Figure 8-3 for GMM scores and Figure 8-4 for HMM scores.

From the GMM figure (8-3) it is clear that noisy sentences have received lower GMM scores (the difference between user and global model scores) compared to the other three sets of original, converted and synthesised sentences.

In contradiction to the first hypothesis proposed above, the individual user and global model scores are higher for spoofed signals. This is due to the fact that in conversion algorithms we have run the gradual score maximization method for every frame, some of which would normally be silence or given a low score. It has made a more speech-rich signal, with higher average scores assigned by both models. Similarly for synthesised speech the choice of high score segments has produced speech-rich outputs with higher scores (The straight lines indicate the average of values within the set).

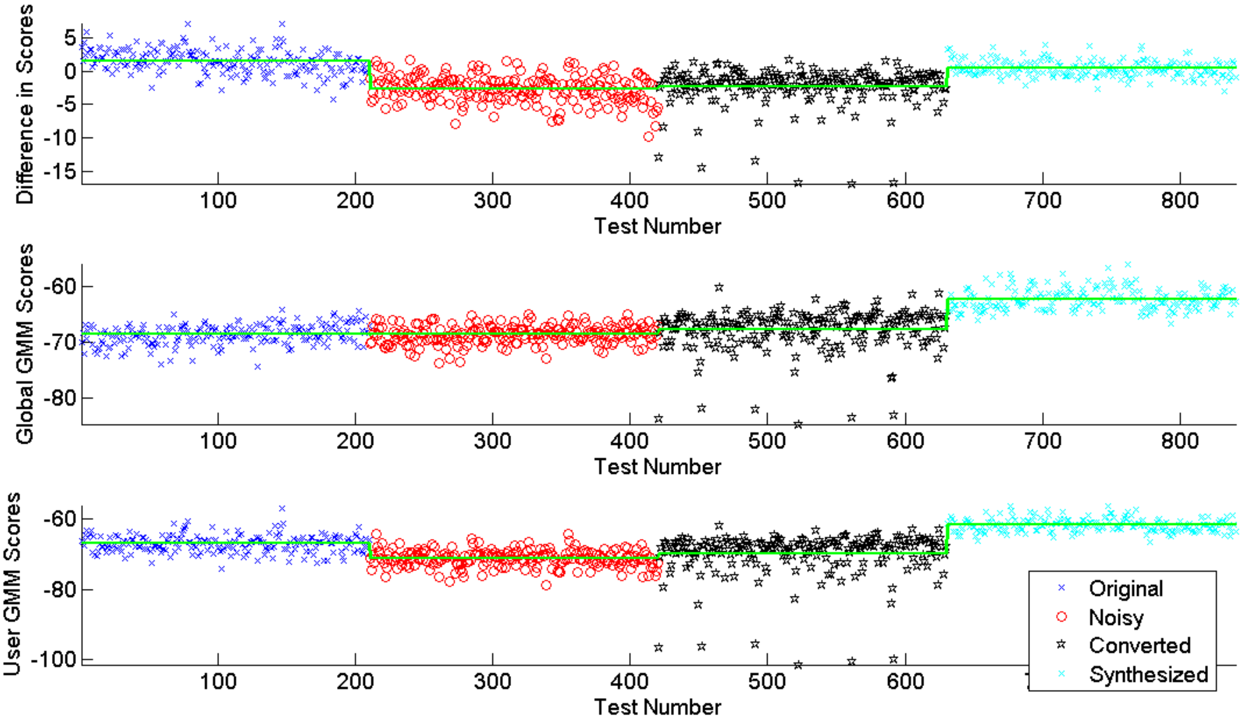


Figure 8-3 Global and User GMM Scores for 4 Subsets

The linear classification error[1] (the overlap between the subsets) for converted and synthesised speech against original (clean) speech at false alarm rate of 5% was 53.5% and 97.3% respectively, reiterating that the spoofing methods have been extremely successful at least against the speaker verification module (15.2% and 34.8% equal classification error rates respectively).

[1] Classification errors reported here are overlap errors meaning that a varying threshold is chosen for two sets and two types of errors are calculated. This error represents the overlap between two sets and the potential for classifying them. When the train and test sets for classification are different, for example when one third of data is used for training classifier's parameter and two third for testing, it is specified in the text.

The HMM scores do not show as much spoofing success as GMM scores. Nevertheless the scatter diagram shows that the individual scores of HMMs are higher (refuting the hypothesis) and that the overall HMM score is on the level of noisy speech.

The linear classification error for converted and synthesised speech against original (clean) speech at false alarm rate of 5%, was 45.4% and 27.8% respectively (11.9% and 9.5% equal classification error rates respectively).

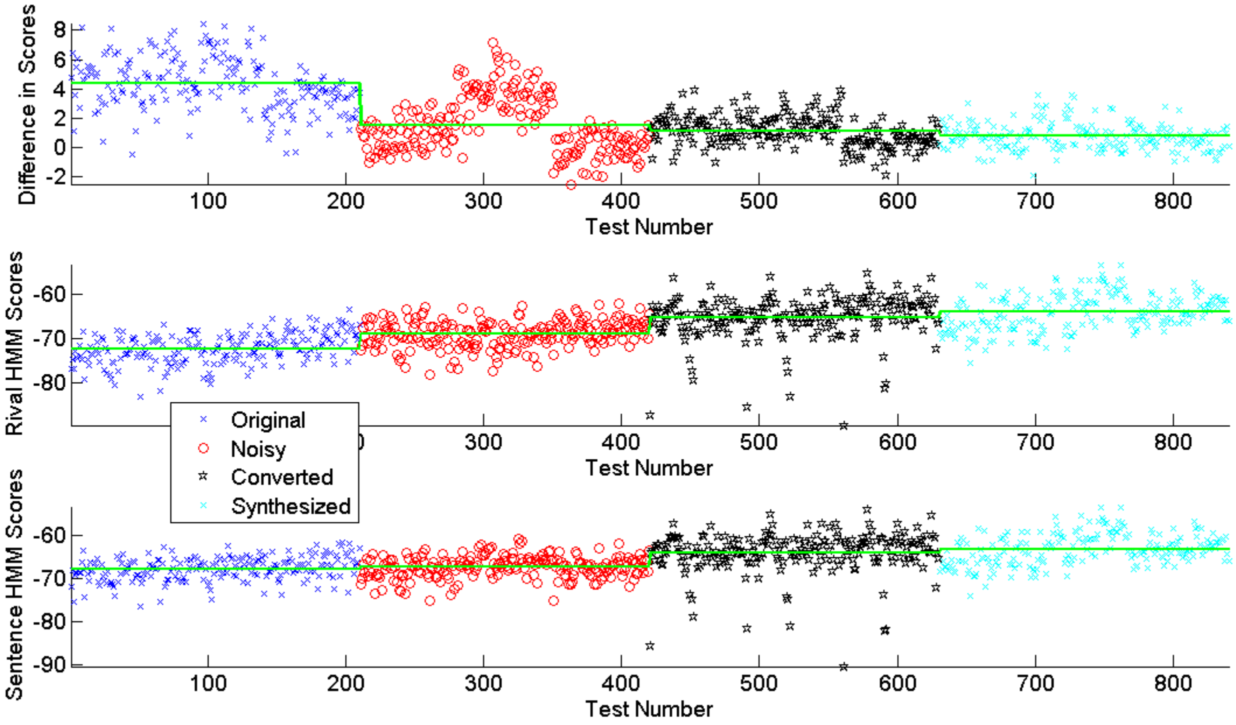


Figure 8-4 True and Rival Sentences HMM Scores for 4 Subsets

Despite the comparatively low classification errors, when the threshold is set at the EER or at the secure point, the incurred errors become as high as the ones reported in chapter 6.

Figure 8-5 demonstrates this fact by showing the content verification scores for rival-sentences (instead of the same but noisy sentences). The aim of the HMM threshold is to discriminate between prompted sentences and other sentences. Based on the figure it could be noted that setting the threshold where FAR and FRR rates are both low (for example around 0) causes acceptance of a large portion of synthesised and converted speech.

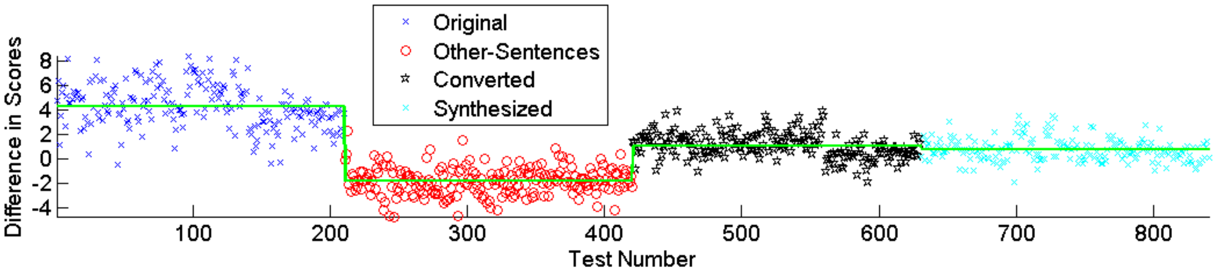


Figure 8-5 HMM Scores of Prompted Sentence against Average Scores of Rival Sentences (Other Sentences)

8.6.3 Illustration of the Problem of Setting Decision Threshold

In this section, the problem associated with the threshold setting is illustrated. This section only focuses on GMM models and voice conversion.

Figure 8-6 displays the false acceptance and rejection rates for normal conditions[1], noisy conditions and when voice conversion is carried out. It can be seen that the effect of voice conversion has been positive on the GMM scores shifting the spoof FAR curve to the right. By contrast, the noise has caused the shift of both FAR and FRR curves to the middle (FAR to the right and FRR to the left).

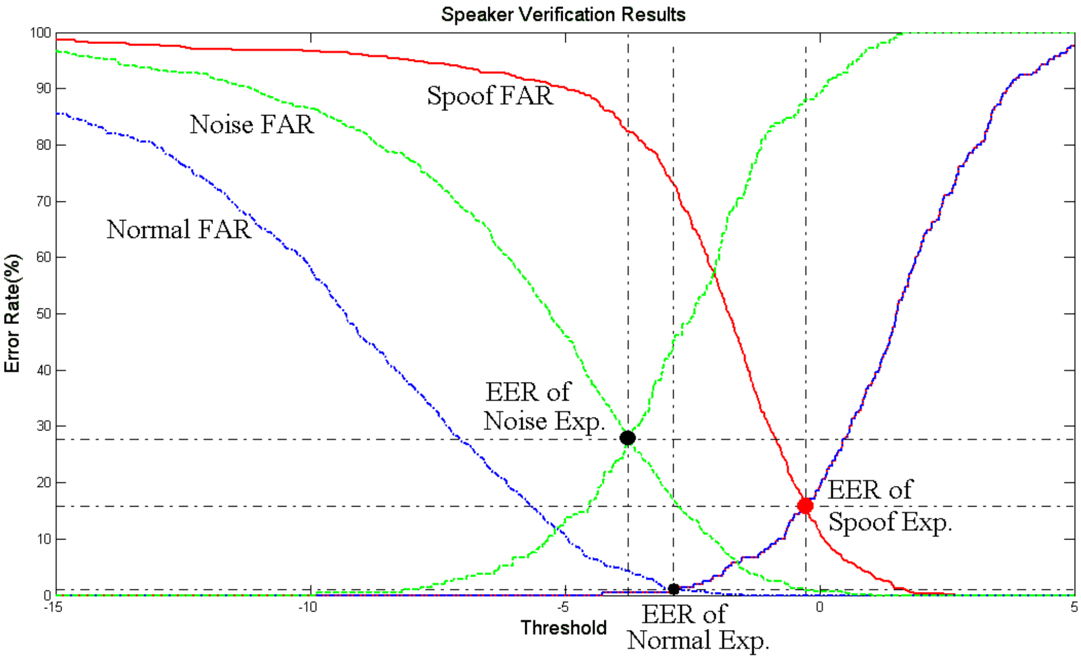


Figure 8-6 Error Curves for Normal, Noisy and Spoof-Related Experiments

[1] GMM models with 64 Gaussians and based on 24 MFCC coefficients.

Even though for speaker verification the EER of converted speech is close to 15% the false acceptance rate at the previously set (EER) thresholds is over 70%. If we incline toward convenience and set lower thresholds in order to accept a portion of noisy signals, we are accepting the risk of a very high false acceptance rate (reaching 80% and higher).

The normalization and compensation techniques try to bring the noise curves closer to clean (normal) curves. The discussion above shows that alongside pursuing this goal we need spoof detection blocks to reduce the spoof FAR.

8.6.4 Early Discussion and Conclusion

So far we have observed that rejection of under-scoring cases, in terms of either user score or global model score, fails to reject spoofed signals generated by our algorithms. As an early conclusion, the synthesised and converted speech developed in chapter 6 did not show a noise-like-pattern and individual GMM and HMM scores were higher than natural speech because of modification to, or choice of, ‘speech-like’ segments.

The analysis, however, showed that the development of more accurate and strict content verification modules could act as a line of defence but inevitably raises the question of trade-offs between security and convenience and as showed in case of noise contaminated speeches can cause rejection of a large portion of genuine speech signals.

Although these methods may be useful in rejecting low-complexity spoofing methods, we should seek more robust algorithms for the rejection of counterfeit speech.

8.7 Consistency Analysis for Detection of Converted Speech

8.7.1 Description of the Approach

This section concentrates on the consistency of the scores assigned by the various algorithms-which output correlated scores for natural voice as a tool for detection of converted speech. The underlying assumption behind this approach is that manipulated speech exhibits incoherent characteristics from different perspectives which are captured by different features.

Let's assume that the speech signal can be split into frames of $F_1,....,F_T$ and each algorithm assigns a sequence of scores to the frames:

Equation 8-6

$$s_j^t = P(f_j(F^t) \mid \Pi_j)$$

$f_j(.)$ is the feature extraction function of method *j* and Π_j is the model for method *j. t* is an index denoting the frame number. We will have values of *s* obtained from different algorithms and models over time.

If the consistency hypothesis is true, the values of *s* show a higher degree of correlation for natural speech compared to converted speech. In simple terms, when a segment's score is high, it is high for all the algorithms and when it is low is low for all of them.

For the evaluation of correlation and consistency, several measures can be used such as mutual information and difference in normalised values. The correlation coefficient is a reliable measure for the analysis of the linear association between variables employed in this chapter. The correlation coefficient between two variables *X* and *Y* is:

Equation 8-7

$$R_{XY} = R(X,Y) = \frac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)\operatorname{var}(Y)} = \frac{E\{(X-\mu_X)(Y-\mu_Y)\}}{\sqrt{E\{(X-\mu_X)^2\}.E\{(Y-\mu_Y)^2\}}}$$

We define the overall inconsistency for the sequence of scores as:

Equation 8-8

$$C(S) = 1 - \frac{1}{N} \sum_{(i,j)\in P} R(s_i, s_j)$$

where *S* is the scores' matrix whose elements are $S_{j,t} = s_j^t$ and s_j is the *j*-th row of the matrix containing values of scores from method *j* over the time, *N* is the number of members of *P* in which *P* is the set of pairs of features *(i,j)* chosen for calculation of consistency (feature pairs whose correlation coefficients are going to be calculated).

8.7.2 Choice of Features and Models for Consistency Analysis

Several sets of models and features were tested for consistency analysis. The experimental results showed that the models which have the same number of components but work on various numbers of features produce better results. The experimental results presented in this chapter are obtained based on six pair of features and models as follows:
M1: Models using 32 Gaussians and 24 coefficients
M2: Models using 32 Gaussians and 16 coefficients

M3: Models using 32 Gaussians and 16 coefficients + derivatives
M4: Models using 32 Gaussians and 24 coefficients + derivatives

M5: Models using 64 Gaussians and 24 coefficients + derivatives
M6: Models using 64 Gaussians and 16 coefficients + derivatives

P={(1,2),(3,4),(5,6)} which means that consistency values are calculated on pairs of 1 and 2, 3 and 4, and 5 and 6.

There were two models for each feature/model set: the user model and the global model. It was also observed that it increases the discrimination power if global models are used as a separate source of scores similar to user models. The results will be presented in the next section.

We will calculate the consistency values for five subsets of authentic and spoofed signals and determine the classification error rates. After that, the classification threshold decided by a portion of test subsets will be applied to the verification task in which the spoof detection block is also active.

8.7.3 Results of Consistency Analysis

Figure 8-7 displays the consistency measure for five subsets of original, noisy, converted, synthesised and retold speech. The plot certainly shows that the consistency hypothesis holds a degree of validity. It is notable that subsets consisting of original sentences, synthesised sentences and even retold (hand-trimmed) sentences have received high consistency values. In accordance with the expectation since the consistency works on the basis of frames and synthesised sentences are built from untouched segments of speech (chosen by HMMs) they have been successful in receiving high consistency scores.

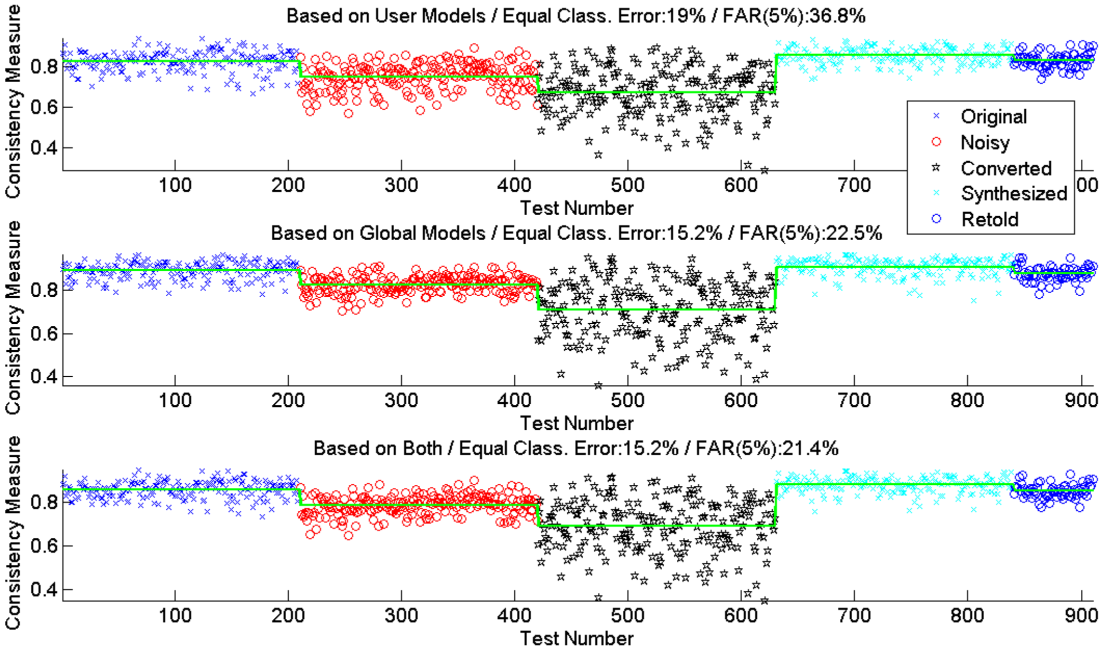


Figure 8-7 Distribution of Consistency Measures for 5 Subsets

The values presented in the figure are calculated based on three combinations of models. In the first subplot (on the top) the consistency measures are calculated only based on the user model scores. In the subplot in the middle the values are assigned only by global models and finally in the third subplot (below) both user and global models were used (correlation coefficients were calculated on 12 streams of scores instead of 6).

The equal classification error rates and the false acceptance rate at a false rejection rate of 5% (false alarm of 5%) are specified in the figures. Results indicate that use of global models or global models in addition to user models brings about better discrimination than use of user models alone.

Finally, we will try to use the spoof detection block in conjunction with the verification system to determine whether it can have a positive impact on the verification results. Due to the limited amount of data, the calculation of *fp* and *fn* functions as functions of score will not be accurate. Instead we will choose a single value for false positive (false rejection rate or false alarm) and calculate the overall FAR and FRR of the system.

One third of data in the original and spoof subsets (70 sentences) were used for threshold setting (classification training set). The threshold was set so that the false positive is around 5%. The rest of the data was used for verification in the presence of the spoof detection block. GMM models with 24 coefficients and 64 Gaussian components were employed for verification.

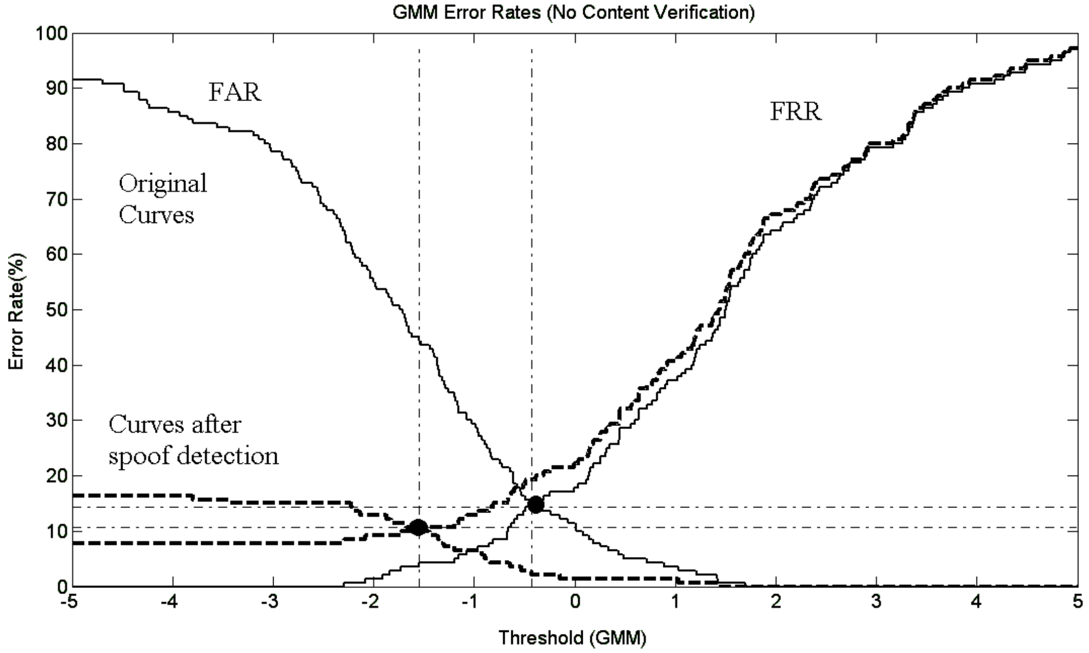


Figure 8-8 Verification Results when Spoof Detection by Consistency Analysis is Used and when it is not used

Figure 8-8 displays the results of verification with (solid lines) and without (dashed lines) the spoof detection block. The false rejection rate has slightly gone up due to the use of the spoof detection bock and it is less than 10%. By the same token, false acceptance errors have drastically fallen implying that the consistency analysis has been successful in the detection of converted signal and has imposed little false rejection rate on the system. Whether this FRR is tolerable for the system is a matter of choice and will be discussed in more detail at the end of this chapter and in the next chapter.

8.7.4 Early Discussion on Consistency Analysis

The results presented here demonstrated the efficacy of consistency analysis for the detection of manipulations of the speech signal. The values of the consistency measure were different for different subsets and followed the expected pattern, which states that there is a potential for detection of manipulation in speech by consistency analysis. Nevertheless, the hidden assumption of classification was that we already have some data from the conversion algorithm (one third of the subset used for calibrating the detection block).

While we were unable to describe fully the spoof detection block, we could empirically evaluate the combination of the system and detection block and determine the overall FAR and FRR values when the spoof detection block was active. The spoof detection block can be characterised by the values it assigns to the cases in the subsets. In that way the controller layer can combine these two values and decide about the threshold in combination with the verification system.

While the results may look promising, it should be noted that activating the spoof detection block brings about a constant FAR of 10-20% and FFR of 5-10% which are not normally tolerable. Therefore, we should devise some methods to use the detection blocks as guiding tools and which give advice on reviewing the suspicious speech signals. This will be discussed in the remainder of this chapter and in full detail in the next chapter.

8.8 Discontinuity Detection and Distance Measures

8.8.1 Description of the Approach

This part of the chapter deals with investigating the possibility of spoof detection by means of employing discontinuity and distance measures. A very large pool of features and distance functions are implemented and tested on speech signals for all the abovementioned subsets.

The discontinuity values are assigned to the feature vectors extracted from consecutive frames of speech by a distance measure. The final discontinuity value for the entire signal is either the average of these discontinuity values or top *M* values[1].

If we assume that the signal consists of the frames of $F_1,....,F_T$ and $f_j(.)$ is the feature extraction function *j*:

Equation 8-9

$$D_j = \frac{1}{T-1}\sum_{t=1}^{T-1} d(f_j(F_t), f_j(F_{t+1}))$$

d is a distance metric for example Euclidean distance. Instead of all *T-1* distance values, top *M* values of *d*(.,.) could be used to calculate the final discontinuity measure.

The rationale behind choosing the max values of the discontinuity in the signal is that (like in speech synthesis) the points of concatenation could be limited and might not be spread across the entire signal. By use of all distances we make the share of those limited points small in the final measure.

The next section introduces the features used for discontinuity analysis and the subsequent one, reviews the distance functions employed in this investigation. Finally the results of experiments on efficiency of discontinuity analysis for spoof detection will be presented in 8.8.5 and a discussion will ensue in 8.8.6.

[1] For the experiments in this chapter, when top values are reported 1/7 of the distances (with highest values) are taken into account. This is indicated by the tag ‘Top’ in the experiments.

8.8.2 Features Used for Discontinuity Analysis

This study makes use of several features for discontinuity analysis. These features are:

1. Spectrum calculated by Fast Fourier Transform (FFT)
2. Spectrum calculated by Linear Predictive Coding (LPC)
3. Mel-Frequency Cepstral Coefficients (MFCC)
4. Mel-Frequency Cepstral Coefficients with Derivatives (MFCC+d)
5. Fundamental Frequency (F0)
6. AM and FM Decomposition Coefficients (AM, FM)
7. Wavelet Features (WAV)
8. Energy of the Signal (ENRG)

The labels in parentheses are those which will be used in the text especially in the result section to refer to these features. More information about the methods used for calculation of these features follows:

Spectral features (1 & 2,above)

The spectrum of signal could be estimated by both FFT transform and LPC transform. Through LPC analysis an all-pole transfer function (filter) for the spectrum is estimated and the frequency components are calculated based on this transfer function. As we will see later in the distance discussion, the normalised and non-normalised spectrum could be used for distance estimation.

Cepstral Features (3 & 4,above)

Feature three and four are well-known cepstral coefficients without and with first derivatives.

5: Fundamental Frequency

Fundamental frequency (F0) in this study is estimated by a pitch calculation algorithm based on the sub-harmonic to harmonic ratio (SHR) proposed by Sun (2002) and by the code provided by the author. The F0 values were calculated only in the 'voiced' parts of the speech which are marked by having energy higher than an energy threshold which was a multiple of the energy of frame at the start of the signal.

6: AM/FM Decomposition Features

The AM and FM decomposition around each center frequency was carried out using the algorithms described in Pantazis et al. (2005) and by the code kindly provided by Yannis Pantazis for this study. The features had outperformed previous features of perceptual discontinuity analysis. In AM/FM decomposition the components have to be calculated around a center frequency. While there are many methods for choosing the center frequency and filtering the signal with the purpose of highest similarity with their work the signal was filtered using 20 Gabor filters with the filters uniformly distributed from 250Hz to 5000Hz. The window size used for AM/FM analysis was 300 frames (18.75ms). For AM and FM features the sum of Euclidean distances of AM/FM components on the edges of the frames and added for all the filters in the filterbank (20 pairs of AM and FM) was used as the discontinuity measure. Contrary to the studies pertaining to perceived discontinuity such as the one mentioned above, we do not know the point of concatenation and the discontinuity measures such as AM and FM are calculated on non-overlapping frames.

7: Wavelet Features

Continuous wavelet transform (CWT) provides a powerful analysis tool with perfect resolution for decomposition of non-stationary signals in time and frequency. There are several wavelet basis functions from which we can choose the most suitable one for particular types of signals. This is an advantage which a short term Fourier transform with decomposition based on underlying Sinusoidal functions lacks. In continuous wavelet analysis, we can focus on specific spectral areas by proper selection of scales. Due to the continuous nature of CWT the values per scale are calculated over time therefore events such as abrupt changes are clearly presented in wavelet transform.

CWT coefficients for a real signal *s(t)* and wavelet basis $\varphi(t)$ are defined to be:

Equation 8-10

$$C(a,b) = 1/\sqrt{a}\int_{-\infty}^{+\infty} s(t)\varphi(\frac{t-b}{a})dt$$

where $\varphi(t)$ is the wavelet basis function (that could be chosen from an enormous number of choices including Morlet, Daubechies, Haar, Mexican Hat, etc.), *a* is the scale value and *b* is the shift value. The wavelet energy coefficients are simply the squared values of CWT:

Equation 8-11

$$S(a,b) = |C(a,b)|^2$$

The main idea behind using wavelet transform was that due to the continuity and better time-frequency resolution it is capable to spotting the abrupt changes in time.

Figure 8-9 displays a signal with two rough concatenation points in the time domain, its continuous wavelet transform (CWT) coefficients and the distance and envelope (max estimate) of the distance. The distance is Euclidean distance between consecutive values of CWT. The change in the max estimate of the distance (the max filter used returns the max value of a 64 point sliding window) is noticeable at the concatenation points.

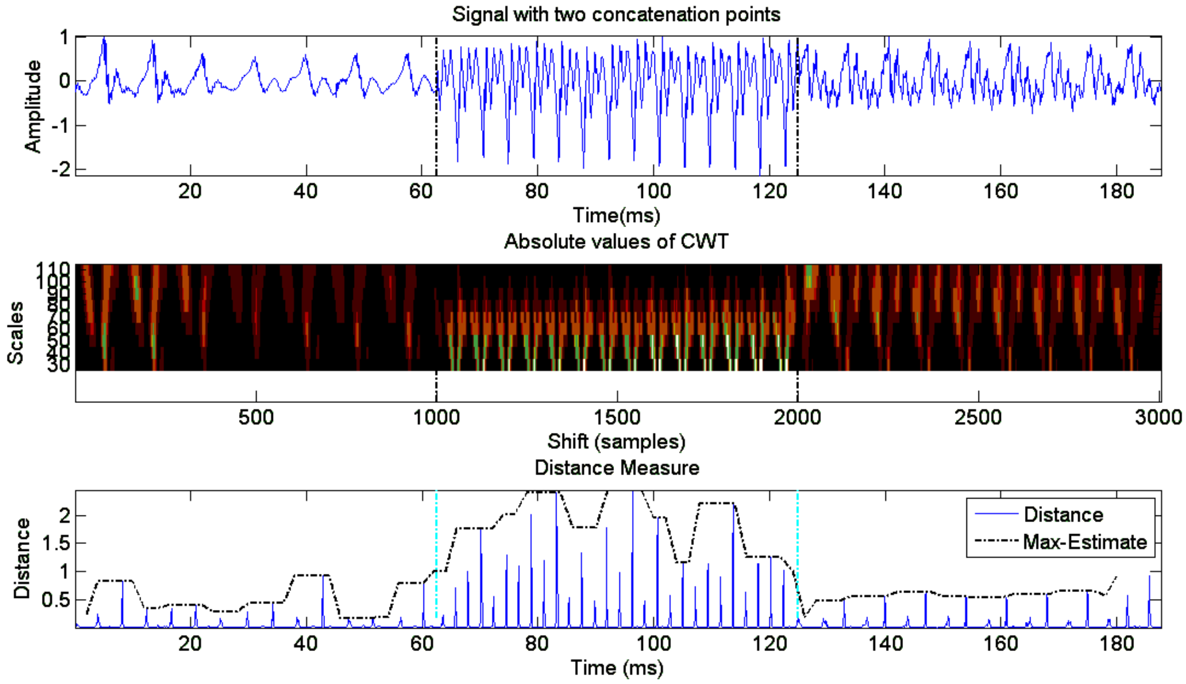


Figure 8-9 Signal with 2 Concatenation Points, its Wavelet Transform, and Frame by Frame Distance Values

The phenomenon shown in this figure does not necessarily happen in elaborate synthesis algorithms which use pitch adjustment and choose the parts of speech from large databases in order to minimise the concatenation cost. Nevertheless perfect synthesis requires plenty of data from the user.

Similar to other methods, wavelet features could be calculated on non-overlapping frames and the values of CWT coefficients could be averaged for the entire frame. The problem with this approach is that the concatenation points lie within the frames and the abrupt change could be left unnoticed at the frame boundaries.

The time-frequency decomposition provided by CWT allows us to extend the number of points for which we calculate the discontinuity to almost every point (It is possible for other features but for many of them requires repeating the computation at each point again).

The results reported for wavelet analysis in this chapter are based on two methods:

1. In the method labeled WAV1 the distance measures are simply the difference in the average values of $S(a,b)$ between consecutive frames. *S(a,b)* is first calculated on the entire signal[1] and for each frame wavelet feature vector, consists of elements having average of energy of wavelet per scale in the frame of 16ms (average of *S(a,b)* over all values of *b* in the frame for each *a*). For two consecutive frames the distance is:

Equation 8-12

$$D(i) = d(W(i+1), W(i))$$

where

Equation 8-13

$$W(i) = \frac{1}{N} \sum_{b=i*N+1}^{(i+1)*N} S(a,b)$$

D is the distance between frame *i* and its next frame, *d(.)* is a distance function from the ones introduced in the next section. *W* is the wavelet vector for the frame *i* (starting from 0) and *N* is the length of frames in terms of samples.

1. In the method labeled WAV2 the wavelet vectors are calculated as explained above but for frames which slide one sample at a time. This is not computationally expensive when the wavelet features are calculated beforehand for the entire signal (this is similar to applying a linear smoothing filter of length 256 to the values of CWT for scales). After that for frames of 16 ms (256 samples) the distance between feature-vectors of all adjacent frames are calculated i.e. the distance of the frame starting at sample 1 to the one starting at sample 257, sample 2 to 258 and so on. In this method we are taking into account all the points in time.

Two wavelet basis functions of Morlet (morl) and Daubechies-2 (db2) are used in the feature extraction.

[1] The scales used where 16 scales from 2 to 77 which translate to frequency range of 168-6500Hz for morlet and 138-5333Hz for db2 wavelet basis functions. Due to inverse proportionality of scale and frequency the density of scales is higher for lower frequencies.

8: Energy Feature

Signal energy (sum of squared values of signal amplitude) per frames of 16ms was calculated on non-overlapping frames throughout signal.

8.8.3 Distance Functions

Two main distance functions are used in this study. The first distance measure is the Euclidean distance between two vectors defined as:

Equation 8-14

$$d(X,Y) = \sqrt{(X-Y).(X-Y)}$$

where *X* and *Y* are two vectors and "." operator indicates inner product.

The second distance is Kullback–Leibler (KL) distance/divergence which has been widely used for assessment of perceived discontinuity and was mentioned in the review of previous works. The KL distance of *Y* from *X* is defined to be:

Equation 8-15

$$d_{KL}(X,Y) = \sum_i x_i \log \frac{x_i}{y_i}$$

in which x_i and y_i are elements of *X* and *Y* vectors. KL distance is not a symmetric distance and was initially made for probability distributions assuming that *X* and *Y* represent pdfs and $\sum_i x_i = 1$ (and $\sum_i y_i = 1$). Therefore the values in X and Y are normalised before calculation of distance. For example, if the spectrum is used as a feature the spectrum should be normalised prior to calculation of KL-distance. There is not, however, any mathematical limitation on use of non-normalised values for KL analysis. Normalised values ignore the information about energy-change in consecutive frames. The distances based on both normalised and non-normalised values are included in this study.

In the report section the Euclidean distance is shown by EUD, KL-distance is denoted by KL and when features are normalised (to have a sum of one) it is indicated by (Z).

8.8.4 Design of Experiments

The experiments aim at determining the success of spoof detection by discontinuity analysis for converted and synthesised speech. Many subtle points have been taken into consideration to make the implications more predictive of the results in practical situations.

The results are first reported in terms of overlap errors. The false alarm or FRR is fixed at 5% and the FAR is reported for each distance measure. A total of 40 feature/distance combination is picked out from possible choices. The results report overlap of 'subset' versus 'subset' for example classification error of synthesised subset versus original/clean subset. This is an indicative of the potential of the distance measure to discriminate between subsets.

One of the main considerations about spoof detection and interpreting its results is that we should make sure that the distinction made by the detection algorithm is based on the spoof-related characteristics of speech rather than unrelated attributes of those particular subsets. For example it was noticed that some of the WAV1 wavelet distances were higher for authentic speech by which we could discriminate synthesised voices from authentic ones. Nevertheless it was in contradiction with the main logic behind the use of distance and therefore the classification results were not reported based on replacing two classes (it is always assumed that the distances are lower for authentic voices).

With the same objective, the experiments were repeated from the spoofed subsets against many authentic subsets which were: original/authentic (Subset B), read (Hand-trimmed), noisy and finally, a solo subset of the Chain Corpus. This was done to ensure that the discrimination is not based on speech-related attributes such as rate, accent or style of speech.

After feature by feature analysis, the best features in each case are selected based on the criterion of having good results for all the subsets. The result of verification in the presence of the spoof detection block is presented afterwards (except for the commercial synthesised sentences for which we can not report the results since we do not have speaker models for those sentences).

8.8.5 Experimental Results

Table 8-2 reports the classification errors when the threshold is set to yield a FRR of 5%. The results shown in the table are FARs in percent for 40 selected features. Six groups of results are presented in the columns of the table. Each column represents the experiment with one subset against another e.g. the column with the CON-O name presents the errors for detection of speech in the converted subset from those original/clean test subset. CON, SYN and CMM stand for converted, synthesised and synthesised by commercial synthesisers. O and R stand for original and read[1]. 'Top' tag indicates using of 1/7 of distances with highest values.

Many observations could be highlighted by critical examination of the results in the table:

1. Abrupt energy changes at the concatenation points which has not been taken care of in both of our spoofing methods, have created an opportunity for detection of counterfeit speech based on energy feature or related features. For example, the KL distance of FFT feature can discriminate synthesised signal from authentic ones when used with non-normalised features, but does not work equally well with normalised FFT. This is because the energy information is lost in the normalised FFT.
2. Without inclusion of read sentences the error rates could be misleading. While both the original and read subsets are authentic, the distance measures do not necessarily work for both of them implying that the distance measures has captured some inter-subset changes as a result of difference in styles of speech .
3. The energy change problem can be easily alleviated and the same pattern does not exist in the speech generated by commercial synthesisers therefore the energy is not a reliable metric in all the circumstances.
4. Keeping only the distance from top 1/7 of frames has not generally worked.
5. There are many features with low errors by which our synthesised signal can be detected. The errors for conversion are higher.
6. WAV2 features have outperformed WAV1 features indicating that the idea of sample by sample investigation has been productive.

[1] The retold subset was created for this research by manually removing the silences in the speech which would cause some discontinuity. Therefore, in this section, only the read subset will be used for which the first sentence of the Cinderella Story was hand-picked.

Table 8-2 Results of Classification (FAR at FRR of 5%) by 40 Features for 6 Pairs of Subsets

No.	Features	CON-O	CON-R	SYN-O	SYN-R	CMM-O	CMM-R
1	EUD-MFCC	35.6(%)	86.8	12.6	92.0	25.0	77.7
2	EUD-MFCC(+d)	36.9	87.5	13.0	93.2	22.7	75.7
3	KL-FFT(Z)	36.8	90.0	10.9	86.4	27.1	66.2
4	KL-FFT	10.6	62.5	0.0	0.0	41.6	95.9
5	KL-LPC(Z)	61.4	95.2	28.5	100.0	14.6	61.5
6	KL-LPC	72.5	81.1	30.6	44.8	67.0	78.4
7	EUD-FFT	12.2	71.3	0.0	10.5	75.1	100.0
8	EUD-LPC	87.7	83.0	72.7	63.6	91.5	89.0
9	EUD-FFT(Z)	32.5	30.1	9.9	8.3	96.1	94.9
10	EUD-LPC(Z)	34.4	70.5	18.1	66.8	68.0	96.6
11	ENRG	1.9	69.4	0.0	5.4	47.1	100.0
12	EUD-AM	93.2	98.1	74.9	88.0	74.1	89.3
13	EUD-FM	92.4	99.0	95.4	99.8	63.3	87.0
14	Top-EUD-MFCC	76.9	98.1	34.4	95.8	14.7	74.4
15	Top-EUD-MFCC(+d)	76.5	98.1	34.6	95.9	14.7	73.8
16	Top-KL-FFT(Z)	72.0	96.2	44.0	92.1	23.4	59.2
17	Top-KL-FFT	10.4	61.7	0.0	0.6	70.0	98.3
18	Top-KL-LPC(Z)	79.2	97.0	56.7	100.0	6.5	49.8
19	Top-KL-LPC	82.8	85.1	45.0	50.3	77.5	80.0
20	Top-EUD-FFT	30.6	76.7	1.9	29.3	98.4	100.0
21	Top-EUD-LPC	91.3	83.5	77.7	65.2	92.0	85.3
22	Top-EUD-FFT(Z)	77.1	50.7	71.6	33.9	100.0	99.5
23	Top-EUD-LPC(Z)	52.1	72.5	47.8	77.3	55.2	94.7
24	Top-ENRG	13.8	77.6	0.0	20.1	92.2	100.0
25	Top-EUD-AM	78.7	94.2	17.3	37.0	90.1	99.4
26	Top-EUD-FM	93.5	100.0	85.8	100.0	34.7	90.2
27	F0	87.1	86.6	86.8	85.7	75.2	75.0
28	Top-F0	89.4	89.8	83.3	82.0	87.1	86.5
29	KL-WAV1-morl (Z)	100.0	100.0	100.0	100.0	11.9	44.2
30	EUD-WAV1-morl	94.8	91.4	95.8	91.4	96.9	96.9
31	KL-WAV1-db2(Z)	100.0	100.0	100.0	100.0	20.0	48.9
32	EUD-WAV1-db2	92.9	83.6	95.6	89.2	97.2	96.9
33	KL-WAV2-db2(Z)	77.6	97.1	52.9	98.3	0.0	11.4
34	Top-KL-WAV2-db2(Z)	86.2	99.0	75.3	98.6	0.0	10.9
35	KL-WAV2-db2	6.3	30.9	0.0	0.0	33.9	91.8
36	Top-KL-WAV2-db2	6.2	19.6	0.0	0.0	64.9	90.9
37	KL-WAV2-morl(Z)	72.4	92.0	30.1	87.7	6.5	34.6
38	Top-KL-WAV2-morl(Z)	84.6	95.0	50.1	95.5	2.6	30.6
39	KL-WAV2-morl	6.7	31.9	0.0	0.0	37.8	91.7
40	Top-KL-WAV2-morl	9.4	25.9	0.0	0.2	61.1	92.0

To choose the best features in each case a criterion is set that the selected feature should not have a classification error (FAR at FRR of 5%) above 40% against any of the 4 authentic subsets of original, noisy, read and solo (of Chain corpus). The best four features that satisfy the criterion along with their FAR errors (in percent) are presented in Table 8-3.

It is noticeable that for synthesised and converted signals the distances based on non-normalised features have worked better (because of the energy change) while in the commercial cases the same distance measures have outperformed when calculated on normalised underlying features. ENRG itself is not among the top features due to its variability and, therefore, its poor performance on other datasets.

Table 8-3 Classification Errors of Best Distance Measures which Satisfy the Presented Criterion

Converted vs.	Original	Noisy	Read	Chain
EUD-FFT(Z)[1]	32.5	0.0	30.1	35.7
KL-WAV2-db2	6.3	5.0	30.9	31.9
Top-KL-WAV2-db2	6.2	6.2	19.6	21.0
KL-WAV2-morl	6.7	5.9	31.9	36.4
Synthesised vs.	**Original**	**Noisy**	**Read**	**Chain**
KL-FFT	0	0	0	0
KL-WAV2-db2	0	0	0	0
Top-KL-WAV2-db2	0	0	0	0
KL-WAV2-morl	0	0	0	0
Commercial vs.	**Original**	**Noisy**	**Read**	**Chain**
KL-WAV2-db2(Z)	0.0	0.8	11.4	26.4
Top-KL-WAV2-db2(Z)	0.0	0.0	10.9	22.5
Top-KL-WAV2-morl(Z)	2.6	2.3	30.6	37.2

The plot in Figure 8-10 presents three of these features. Each of the features is efficient for detection of at least one set of spoofed signal. The distance measures for all the subsets are included in the figure to ensure that the discrimination is not accidental. Another important observation in this figure is that even within the commercial synthesised signals each algorithm has its own signature and the second and third synthesisers have produced signals with higher discontinuity metrics compared to three others.

[1] EUD-FFT(Z) was chosen for the sake of having various features otherwise Top-KL-WAV2-morl had better results and satisfied the requirements.

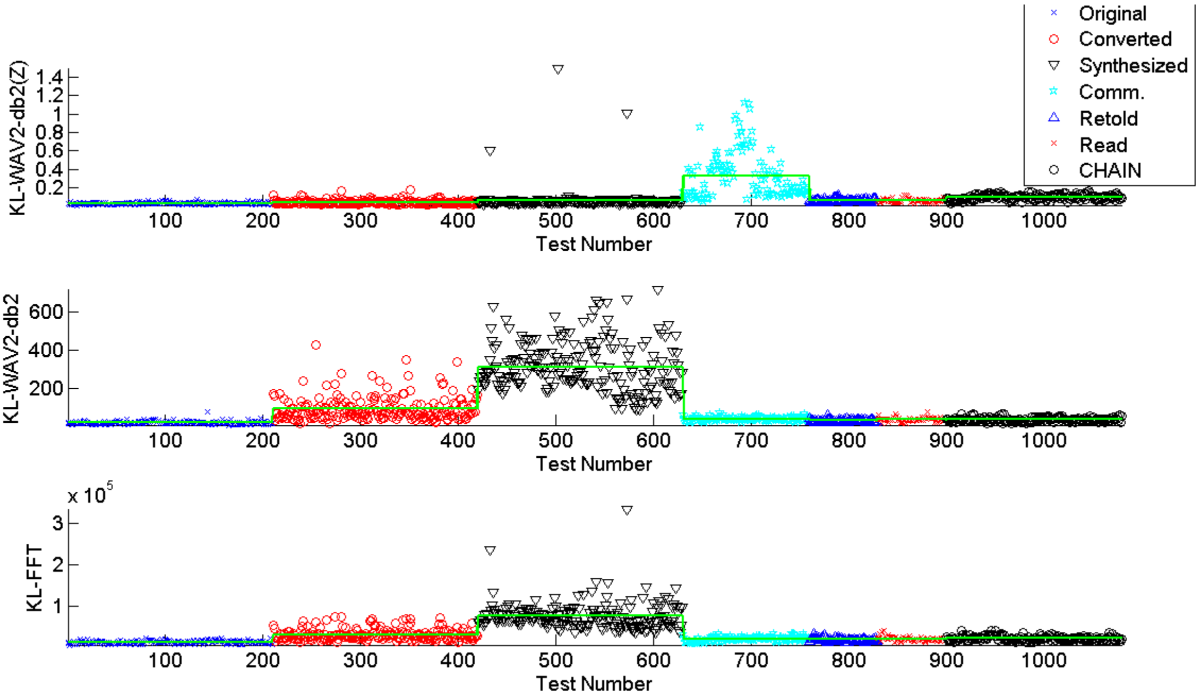


Figure 8-10 Distribution of 3 Distances for 7 Subsets

Finally we will try to use the detection block in conjunction with the voice verification system. Since the error rates are zero for synthesised signals with choosing a good classifier we will be able to reject all counterfeit cases (or close to all cases based on the choice of classifier) for speech synthesis.

The scatter diagram in Figure 8-11 displays the distribution of samples in the feature space consisting of the two features of EUD-FFT (Z) and KL-WAV2-morl used by the classifier. It is notable that accurate classification requires a rival set consisting of counterfeit cases.

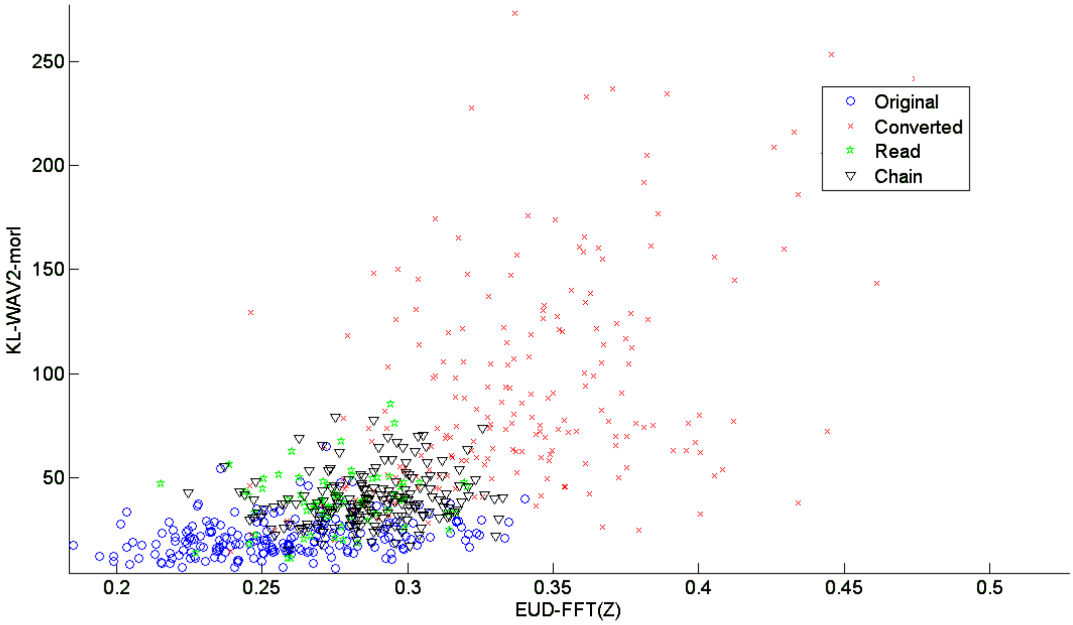


Figure 8-11 Scatter Diagram Showing Distribution of 4 Subsets in Two Dimensions Representing two of the Best Distance Measures

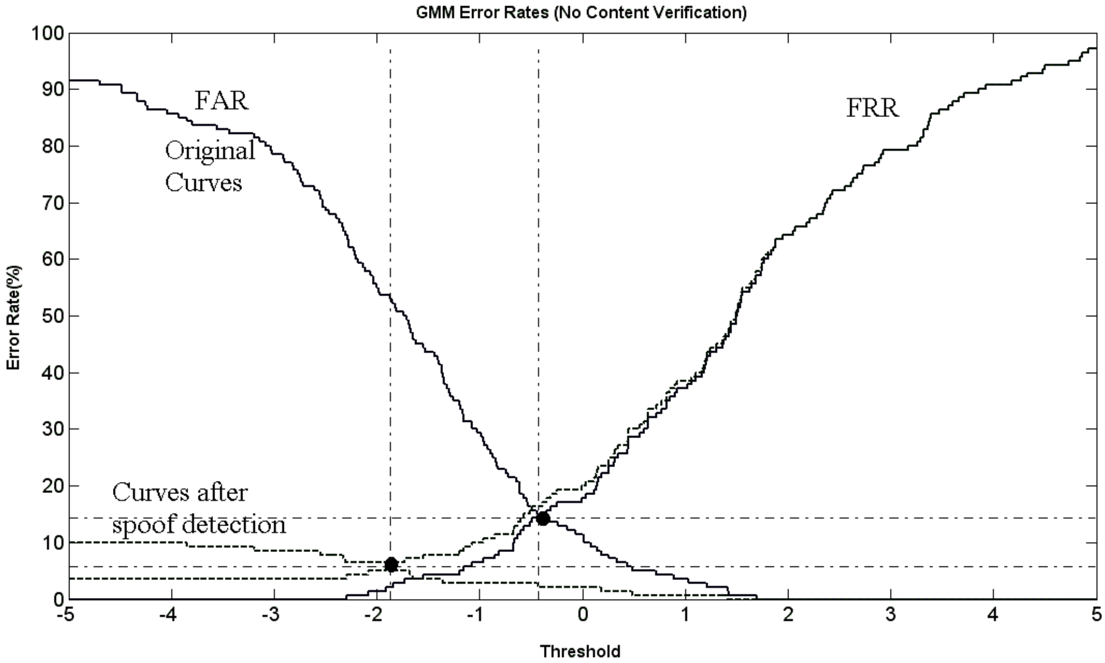


Figure 8-12 FAR and FRR Error Curves Before and After Use of Spoof Detection Block

In the case of the conversion algorithm, a quadratic classifier in two dimensions with 1/3 of data from original and synthesised subsets as training and the rest as test, has produced verification curves such as those shown in Figure 8-12. The FAR values have drastically fallen at the cost of a slight increase in FRR values. More importantly the EER point has been shifted to the left approaching the point where the previous EER of the system under normal conditions was located, making the threshold selection easier (compare with Figure 8-6).

8.8.6 Early Discussion on Discontinuity Detection

All the findings up to this section revealed that the security of voice verification systems, without a spoof detection block, is very much in jeopardy. One of the possible solutions to spoof detection is the detection of abnormal discontinuities in speech signal over the time. Previous research had shown surprisingly that all of the distance measures, even if the point of concatenation was known, would produce poor success rates for detection of perceived discontinuity. The false acceptance rates were mostly above 50% at false alarm of 5%. Only one study showed better results as a result of a combination of features and use of a linear classifier.

In this chapter, a detailed study of all distance measures was offered and some new distance measures based on wavelet transform were proposed. The results seemed promising for some subsets, and we were able to completely eradicate or considerably minimise the security risks. However the results should not be misinterpreted.

Three important points about these experiments should be considered:

Firstly, we examined a large pool of distance measures in conjunction with the available datasets. While these features work for the problems at hand there is no guarantee that the same features work properly for new ways of spoofing.

Secondly, an acceptable classification outcome could not be achieved unless data from a rival set is available. In other words, we need data from the spoofing techniques and we have the assumption that the data available from a technique exhibits coherent statistical characteristics in terms of distances.

Thirdly the fact that different features work variably for different subsets implies that the method used for dealing with each algorithm of spoof is different and constant updating of spoof detection algorithm both in terms of feeding new data and devising new measures is necessary.

These three points have significant implications for the design of a working verification system and its maintenance. These implications are addressed in greater detail in the next chapter.

8.9 Early Conclusion and Next Chapter

Voice verification is principally vulnerable to two types of unreliability of decisions: Unreliability of decisions due to a mismatch in the train and test conditions and unreliability imposed by the possibility of spoofing. Setting an acceptance threshold is almost impossible if these two types of unreliability are not treated properly.

In this chapter, a long stride towards determining various aspects of the second type of vulnerability was taken. The lessons learnt in this chapter help us evaluate the chance of using voice verification in each of the contexts mentioned before. More importantly the results give essential information about the requirements of keeping the system secure over time and in the ongoing battle on the spoofing front.

Several suggestions were made and three of them were implemented for mitigating the risk of spoof attacks. While all of them were successful to some degree they required various amount of data about the spoofing techniques.

It was shown that content verification itself is a line of defence against simple spoofing techniques which can not satisfy the conditions of speech verification. On the other hand relying heavily on the content verification brings up the same mismatch issues discussed in chapter 7 for speaker verification and can cause a large false rejection rate. Therefore making use of strict content verification modules is not a realistic suggestion.

A consistency measure was proposed and its success in the removal of the counterfeit signal was analyzed. While the results were promising, the false rejection rates necessitate restricting the use of these detection blocks, otherwise the constant false rejection of spoof detection block will be added to the system's FRR even when the verification system's threshold goes to minus infinity. Selective use of a spoof detection block based on the context of verification allows us to use this blocks only for and in the contexts in which the vulnerability exists. For example, when an operator is talking to the user on the phone at distance, with the assumption that the operator can detect a converted signal the spoof detection blocks for those recognizable spoofing techniques by human could be turned off.

A large study on the discontinuity detection measures revealed that the spoof detection techniques can take advantage of some of the flaws in the speech generation systems such as amplitude change which are not taken care of by perpetrators. The measures, however, do not work globally and each method leaves its own footprint. Success in the detection of discontinuity in signals is based on having data from spoofing modules and relies on the assumption that a technique displays consistent statistical characteristics in terms of distances. Constant monitoring of the system, feeding new data to the spoof detection modules and updating the types of spoofs are some of the requirements stipulated by the analysis.

The commercially synthesised speech signals normally use long portions of natural speech which may not be available to all perpetrators.. Some of the sentences used here were common sentences for which the synthesiser has been optimised, had had enough data and needed little adjustment. Prompting uncommon sentences unlike string of digits, combination of address, name etc. imposes more complexity on the synthesiser and forces it to make lots of modification in the available data in which it leaves its own footprint.

In short, and considering all the assumptions and errors discussed in this chapter, it seems that use of spoof detection blocks is inevitable but that they should be used as helping tools along with the human supervision. In other words, these blocks can pick out signals which are most likely to be spoofed for further human inspection and for the updating of detection blocks.

Development of spoof detection blocks and keeping them up-to-date is necessary for upholding the security of voice verification systems. New techniques always lie in wait. There is huge room for optimizing the way they are used and this will be discussed in grater detail in the next chapter in which we put all the discussions presented up to now into one picture and discuss their implications for the design and development of working voice-based biometric systems in various contexts.

Chapter 9 : Synthesis

9.1 Goal of This Chapter

This chapter recapitulates the main ideas discussed in the thesis, reviews the findings, and places the findings in the larger study of analysis of voice as a biometric for security and crime-related applications.

The propositions made here, along with the results reported in the previous chapter, demonstrate that despite all reliability and spoof related issues voice verification in a semi-automatic set-up can be an effective and sustainable biometric but needs to constantly evolve alongside the developments in speech science.

Section 9.2 is a brief summary of findings reported in previous chapters from which conclusions are later drawn. In 9.3, all human concerns are revisited and voice is compared with other biometrics. A list of human concerns is presented here, which provide a strong basis for subsequent stakeholder opinion surveys targeting a 'specific population' using a 'specific system' for a 'specific purpose'. On a higher level a blueprint for collaboration of service-providers and (biometric) identity providers with emphasis on separation of access permits and certification of biometric services is presented in 9.4 and it is demonstrated how this set-up could mitigate concerns related to use of biometrics and it is compatible with legal requirements and recommendations discussed in respect of identity management systems. Sections 9.5 to 9.8 draw conclusions about suitability and the problems of using voice in 4 categories laid out in chapter 3. Section 9.9 identifies the potential for future research and section 9.10 concludes the study presented in the thesis, under the title of analysis of voice and other biometrics for criminological and security applications.

9.2 A Short Summary and Review of Findings

In chapter 3, categories of identity fraud were reviewed and it was emphasised that there are cases for which a reduction of opportunity for fraud is desirable even though the eradication of identity fraud is impossible or unrealistic. Especially when long distance authentication is needed, voice owing to the availability of collection devices and for non-technical reasons, is one of the best options. Non-technical (legal, ethical, privacy related) dimensions of biometrics were scrutinised and a list of human concerns was compiled. The conclusion was that the legal frameworks allow use of biometrics with consent, and in some cases in the interests of the public but there are concerns that have to be addressed. Based on the review of literature and opportunities for use of voice, 9 groups of scenarios in 4 categories in need of further investigation and analysis were suggested.

Chapter 4 addressed the most significant problem of the use of voice for security and crime investigation. This was that, a comprehensive framework for the evaluation of the reliability of voice-based verification algorithms under various conditions especially in view of the spoof attacks, remains absent from the research literature and standards. The necessary characteristics of such a framework and an automatic evaluation system in accordance with this framework were discussed.

In chapter 5, a fairly stable voice verification system was introduced and it was shown that the system works suitably.

Chapter 6 was devoted to the development of two spoof modules and the evaluation of the verification system in the face of attacks made by these two modules. The first module used a new method of conversion based on a gradient ascent of GMMs in the cepstral domain and the second was based on an HMM based synthesis in a set-up suggested in this piece of work. Both modules were extremely successful in deceiving the verification system.

Chapter 7 revealed the other side of the evaluation coin: reliability analysis. Several mismatch factors, such as, distance and channel, coding, style and noise were analysed. In each case, some suggestions for improvement was made. Despite this, the main contribution of this thesis in this area was the classification and categorization of conditions in which the verification system should be tested. This was the basis for the evolution of the evaluation system on two dimensions of security (chapter 4 and 6) and reliability (chapter 7). As a secondary contribution and finding, the fusion of features with different spectral foci was suggested and it was shown that one fusion rule that does not need any information about the condition in which it is used, produces results compared to BGI-based fusion which needs extra information about the channel or conditions of use. The fusion algorithm proved to be robust under a variety of conditions.

Following the results in chapter 6 suggesting that spoof is successful even if little data is available about the verification system, the goal of reviving voice verification system's security by the development of spoof detection blocks was pursued in chapter 8. The initiatives were fairly successful and several methods for spoof detection were proposed in this thesis. In addition, important steps towards the development of spoof detection features and algorithms were taken. A large set of features available in the literature (for the similar task of quality evaluation), in addition to new wavelet features proposed here, were analysed, some of which showed promising results for the detection task. Despite the success of spoof detection, it should be noted that the use of spoof detection blocks could increase FRRs. The experiments also demonstrated that there are two goals that oppose each other: increasing security of the system (in face of spoof) and increasing user convenience by reducing false rejection decisions in adverse conditions.

While due to the importance of security-related use of voice verification and its chance of blooming in the future the features and models evaluated here were (inevitably) those used for this purpose (in automatic text-independent mode), the framework, subsets and details of evaluation outlined in the thesis were independent of how the voice verification works.

9.3 Review of Human Factors and their Relevance to Voice

Chapter 3 in conjunction with Appendix C made a review of all human concerns expressed in the realm of biometrics. In this section, the relevance of voice in each case is studied. The next section proposes an architectural design for collaboration of various players contributing to verification which could mitigate the concerns. As mentioned in chapter 3 the issues surrounding biometrics could not be settled once and forever. Rather, table 9-1 (which is compiled generously in favour of critics and its items bear some overlap) could be used as a checklist for the design of questionnaires and surveys for the elicitation of users' opinions and for the design of biometric systems. Table 9-1 summarises different concerns introduced in chapter 3, and presents a severity index for voice and three other biometrics (fingerprint, iris and face) in case of each item. It is contended here that the severity indices should be assigned to a specific application for specific users and for a specific mode of use. Those presented here are assigned as a result of discussions presented in chapter 3 and here.

For example, item 6 expresses the concern that use of biometrics in some cases goes beyond what is needed for accomplishing the goals. To investigate this concern, one method is to survey the target-users' opinions about whether or not the measure seems unnecessary for the task at hand after giving them alternatives from which to compare and choose. The rest of this section reviews these concerns and discusses the entries presented in Table 9-1.

Item 1 (stealing parts of body) seems to be most irrelevant to voice. This is a concern for other biometrics especially fingerprint. Medical concerns are negligible for voice (item 2). Medical information could not be extracted from voice (item 3) in contrast to DNA or iris image. Voice however could be used to extract some racial and background information if cross-referenced with sociolinguistic knowledge. Requiring biometric services to become certified by biometric certification authorities as elaborated on in 9.4 reduces such risks.

It seems that there is no ground for religious resistance to use of voice as opposed to some other biometrics.

Item 7 stipulates educating target users[1] about the true value of biometric systems and threats to their effectiveness (e.g. spoof attack possibilities) which is in conflict with the interests of biometric developers and should be enforced by third parties and monitoring agencies. Publishing the results of the evaluation (as described in chapter 4 and the following chapters) in a way that public could digest goes a long way towards alleviating this concern.

Issues related to harm to human integrity (item 8) and the branding argument[2] (item 10) seem to be exaggerated. As mentioned in Appendix C a rather reliable survey (BITE, 2006) has shown that branding analogy is not widely supported. Since voice production is a voluntary action and voice collection happens outside the boundaries of body voice is among the least controversial biometrics in this regard. Social stigma[3] associated with voice is minimal compared to biometrics such as fingerprint. It should be emphasised that despite these general statements, public opinion should be gauged for a 'specific application' in a 'specific target group'. Therefore the severity indices specified in the table only provide a rough assessment and also un-transferable from one application and context to another.

[1] Biometrics could be used in different chains. Target users are those whom the system is intended for. Target users of different systems are different. They have different needs, level of education and expectations. For example the target users of a biometric system which is installed in a university are students, staff and faculty. The users of a biometric system installed at ATMs are more diversified. It is important to attend to the needs of specific groups of users before deploying the system and when designing surveys.

[2] Branding argument simply stated that use of biometrics for human is similar to use of brands and tags for products.

[3] This stigma comes from being asked to give a voice sample. Asking someone to give voice samples is not usually associated with the fact that he or she is a suspect or criminal.

Table 9-1 A Repertoire of Human Concerns and an Estimate of the Severity of Each Item for Voice and Other Biometrics (Fingerprint, Iris and Face) [Severity: Low (L), Medium (M) or High (H)]

	Severity for	
Concern	**Biometrics (FP, I, F)**	**Voice**
1. Concerns about stealing parts of body (e.g. cut of a finger)	H, L, L	L
2. Direct Medical Concerns	L, H, L	L
3. Indirect Medical Implications (similar to sensitive data disclosure)	L, H, M	M
4. Disclosure of sensitive information	L, H, H	M
5. Religious concerns	L, L, M	L
6. Concern about being unnecessary and redundant	-[1]	-[2]
7. Ethical aspects of fear reduction (without reducing its underlying cause)	H, H, H	H
8. Concerns about transgression of human body and human integrity	L to M (all)	L
9. Social Stigmatization	L to M (all)	L
10. Ethical Concerns about value of human and human's dignity (branding argument)	L, L, L	L
11. Concerns about possibility of function creep	L, H, M	M
12. Fear of biometric tracking and pervasive surveillance	L, L, H	M
13. Fear of construction of full profile of actions from partial identities (similar to tracking)	M, M, H[3]	M
14. Fear about covert collection	L, L, H	M
15. Privacy questions	L, L, H	M
16. Depriving people of anonymity	L, L, H	L
17. Possibility of permanent ID theft	H, H, H	H
18. Fear of misuse of data	H, H, H	H
19. Costs (deployment, updating, maintenance as well as costs of fraud and wrong decisions)	H, H, H	H
20. Power accumulation and weakening of democracy	M, M, H	L
21. Law enforcement concerns including false interpretation of biometric evidence in courts	H (all)[4]	H
22. Disability and age problems	-[5]	M
23. Impact on social interactions	H, H, H	H

[1], [2] Depends on the application.

[3] This item is more related to the design of the system and its mode of use, rather than the biometric identifier used. This fact holds for items 15, 16 and 20.

[4] Dealing with technical questions about admissibility of biometric evidence in courts for other biometrics is outside the scope of this research.

[5] Beyond remit of this research.

Function creep, as the most significant issue expressed in relation to biometrics, could be avoided by design considerations detailed in the next section. Biometric certification is suggested in the next section along with the presented design to limit the risk of function creep.

Items 14-16 are addressed by governmental regulations detailed in 9.8 where speech surveillance is analysed. Concerns around privacy, compromising of anonymity and ubiquitous surveillance could all be addressed by the concept of notification as described in 9.4. In addition, if the biometric data are only used for authentication and their use and transmission is limited (as specified in 3.4.2 and detailed in the blueprint presented in 9.4) users can track the history of their biometric authentication and have control over their privacy.

Permanent ID theft (item 17) is a substantial issue over and above the use of biometrics. Voice is no exception. An advantage of voice is the possibility of text-prompting speaker verification. Nevertheless, this advantage does not eliminate this concern as the generation of a spoof signal is possible even in that case. The magnitude of this concern is determined by the vulnerability of the system to a spoofed signal. Two design considerations could alleviate this concern: notification of user as elaborated on in 9.4 and the progressive enhancement of spoof detection modules detailed in 9.5.

Item 18 (fear of misuse of data) as appeared in the literature, refers to the use of biometrics by impostors and the costs of this misuse. Together with item 19 they express a need for caution when the costs of deployment of biometric systems are estimated. Due to the need for constant monitoring of spoofing techniques, the necessity of random human checks and the development of spoof detection blocks, the cost of keeping the system up-to-date and geared-up against various threats is substantially higher than its one-time development cost. Just like credit cards, the loss inherent to use of biometrics is part of the entire cost and should be put into balance with its benefits. Section 9.5 elaborates on how a semi-automatic voice verification system should be monitored and updated in the face of spoof attempts.

Despite all the efforts to regulate use of biometrics by governments and the private sector, biometrics remains a strong tool for controlling citizens. They may contribute to power accumulation and weaken the democracy in the long run (item 20). The design considerations presented in the next section could allay the concerns in this regard contingent upon the condition that government agencies are not constantly exempted from the biometric regulations such as notification requirements.

Biometric evidence is liable to being misinterpreted in the courts (item 21). This may lead to irreversible consequences. The stance taken here is that forensic features and procedures should undergo the same tests carried out for automatic security-based speaker recognition in this thesis under various conditions and that unless proved reliable should not be employed in juridical procedures except for crime investigation and as a means of providing clues rather than proof. In addition, the procedures should be standardised to ensure uniform application in different cases and courts.

Disability problems could be reduced by use of multi-biometrics and offering options to use alternatives to biometrics (such as PINs, tokens and passwords). For voice verification, regular updating of voice models over time, and as new authentication attempts are made, could make up for gradual changes in vocal tract and speaking style over time.

Very few of the concerns discussed above could be as severe as impact of biometrics on social interactions. Similar to permanent ID theft, the severity of this concern is proportional to how the spoof attacks could happen. While the effect is inevitable its magnitude could be controlled by developing more robust verification systems, described in sections 9.4 and 9.5.

9.4 Design Considerations in Light of Findings and Human Concerns

The blueprint proposed in this section aims to incorporate the notions of consent, notice and supervision (both by external auditors and by users) into its underpinning design[1]. Likewise it helps the user deal with a previously categorised number of services and give informed consent[2]

[1] Literature's solutions to the non-technical problems of biometric authentication presented in 3.4.3 included self regulation by biometric providers, social impact assessment, requiring notice, requiring consent, providing user's access to own data for modification, confining use of data to the data collection purpose, use of data in a sectoral boundary and external supervision. Notice and consent in relation to data collection are the most important concepts studied in several scholarly articles and have become part of EU Directive 95/46/EC.

[2] A key question ahead of biometric authentication is whether it should be carried out by governments and data should be stored in large databases owned by state agencies, or people tend to trust private identity providers more than government agencies. A survey cited in Appendix C indicated that the respondents' mistrust was distributed equally between these two options. Therefore the design presented here remains neutral about this choice but proposes shaping of biometric certification authorities as independent monitoring bodies.

by transferring the burden of classification and monitoring of services to new entities named 'Biometric Certification Authorities (BCA)'[1] here.

The role of BCAs is to analyse the services which work on biometrics, to ensure that they fit in a pre-defined category of authorised biometric functions. Those categories might include training a template over raw data, saving a template in the biometric database, matching a record with a template /model or modifying a template / model to reflect a mismatched condition. For this to happen the BCA should not only inspect the inputs / outputs and the description of the function, but it should also analyse the services' internal processes to make sure that no hidden procedures are there[2]. This is the most efficient way to eliminate the possibility of function creep in the future.

A biometric service without a certificate from a BCA should not be trusted by users, service providers and other identity providers (shown in Figure 9-1).

Several entities are displayed in Figure 9-1. The role and function of each entity is described here:

1. BCA is responsible for inspection of the biometric services developed by biometric service providers (BCP). BCA creates a certificate for each service and signs it with its own private key. BCA certifies that "*Service_A* deployed to the server at *Location_X* with input *Arguments_I* and output *Arguments_O*, is developed by *BCP_Y*, and its function is categorised as *C*".

2. To avoid function creep a good practice for BCA is to differentiate between access to raw biometric data and models trained on biometric data. Biometric services usually do not need to connect to databases holding raw biometric data for verification purposes. Access to raw data should normally be restricted to template/model generation or adjustment.

[1] The growing number of biometrics and increasing need for authentication makes it almost impossible for the users to provide informed consent to each application. The need for providing only a few and easily understandable privacy preferences and policies has already been highlighted in the literature (see C.5).

[2] This is already part of the certification process of financial applications on the smart cards.

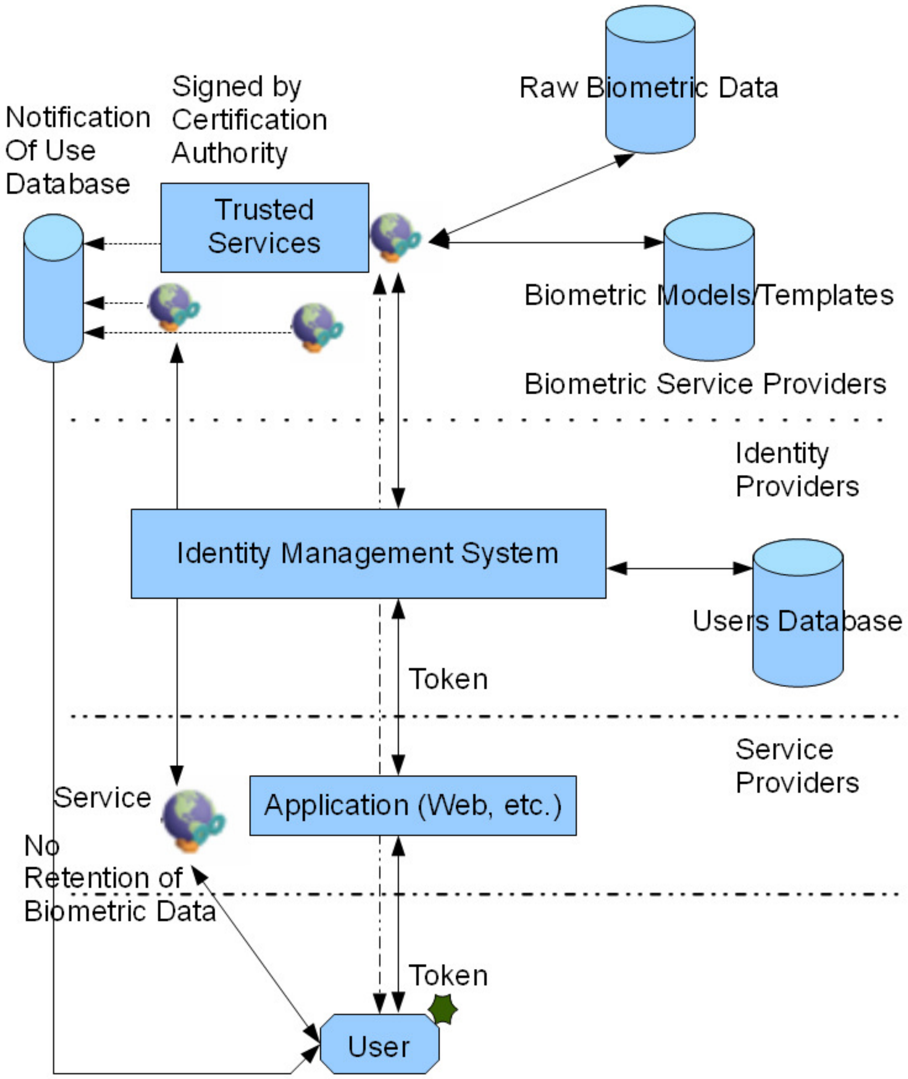


Figure 9-1 A blueprint for collaboration of service-providers and (biometric) identity providers with emphasis on separation and certification of biometric services

3. Identity providers, for example identity management systems, are service consumers of biometric services. The layers shown in the diagram are abstract. A biometric management system, contingent upon the approval of BCA, could deploy its biometric services to its own server, or use tables in its own database for biometric data, but the access to such data should be made through certified services only.

4. Service providers which need to authenticate their users should go through biometric services for this purpose. Biometric services do not apparently respond to unknown requests. Mutual authentication should take place before satisfaction of a request. Users communicate directly

with the biometric services during the authentication process[1], and the results are communicated with the identity management system or the service which has requested authentication.

5. Any use of biometric services is logged in a notification database. Users could review all the entries in the database indicating: service provider which has requested authentication, category of biometric service, and time and date of request.

The details of communication are beyond remit of this research[2]. Apart from those details, the architecture guarantees notification of use, facilitates informed consent, transfers the responsibility of monitoring[3] and certification to BCAs, and finally hinders function creep.

9.5 Automatic Speaker Recognition in Civil Context and for Crime Reduction

9.5.1 Voice Verification on Smart Cards

In this group of scenarios, user specific models are placed on a smart card and in various situations the person uses the voice instead of PIN or passwords. There are a few differences between these scenarios and the next one (long distance authentication).

The first difference is that, similar to use of PINs on cards, the card could be the only place in which the speaker models are held. Biometric data is not maintained in a central database and users have full control over where and how their information is used. The authentication is performed by card and the only result returned by card is the decision.

Chapter 7 showed that noises are detrimental to voice verification. The fact that authentication in conjunction with card, is carried out in public places (in shops, at ATMs, etc.) indicates that the noise factor is crucial. A solution to this problem is devising a uniform way of data collection by employing pre-designed identical microphones (receivers) which are placed on the mouth and insulate the vocal system from the outside (this calls for further research on hygienic implications). Use of a unique microphone reduces the channel problems as well.

Verification parameters such as thresholds are updated from time to time when the transaction goes online. Since the models are only on the card, loss of the card means loss of the models. No

[1] This is similar to internet secure messaging according to standards such as Secure Socket Layer (SSL) when credit card information is communicated (see e.g. Stallings, 2007)

[2] Security information is usually communicated through Security Assertion Markup Language (SAML) tokens to the services. Customizing SAML tokens for this purpose could be the subject of another project.

[3] The BCA certifies the deployment of an authorized biometric service to a server. It is BCA's responsibility to make sure constantly that the service at that location is not changed after inspection and deployment.

security threat is attached to this possibility, yet the data collection and model training (which requires collection of tens of seconds of voice) has to be repeated.

Use of biometric templates on cards (instead of in a central database) reduces concerns 4, 6, 11, 12, 13, 14, 15, 16 and 23. On the other hand there is no advantage to the use of voice compared to other biometrics except for concern 1 (actual harm), 9 (social stigmatization) or medical concerns (2) which are minor issues.

In unsupervised authentication (e.g. at ATMs.) the risk of spoof mandates the use of a spoof detection design presented in the next section.

All told, the facts illustrate that there is little advantage to the use of biometrics instead of a PIN on card, in civil context and for low risk applications. For more sensitive applications (such as passport control or those requiring negative authentication) other biometrics such as iris and finger-print are preferred.

9.5.2 Long-Distance Speaker Authentication

Allowing long distance speaker verification is the most pronounced advantage of voice over other biometrics. This advantage is closely related to the presence of ubiquitously available collection devices and that the voice is the main modality of communication. Requiring any change in this set-up for example restricting types of data collection devices harms this advantage.

Three specific cases could be discussed in this group of scenarios involving long distance authentication:

1. A user (with the aim of using a service) goes thorough a verification process by mobile phone, or on the internet e.g. a person calls the customer service of a communications company and tries to change his usage plan. Authentication is performed in a supervised mode and is part of a longer conversation with an agent.

2. A similar scenario involves the same process except that the authentication is unsupervised e.g. as part of an online purchase on the internet (by a credit card).

3. The last case involves electronic monitoring of curfewees by making random calls to their designated place.

A discussion on the technical and non-technical implications of this research in those cases is presented here.

Experiments in chapter 6 demonstrated that spoof attacks are significant threats to the security of voice verification systems. In Chapter 7 it was established that many adverse conditions require re-adjustment of the threshold in favour of convenience. In chapter 8, both factors were discussed and spoof detection blocks were developed. While automatic spoof detection blocks were partly successful in recognizing manipulated speech, each of them imposed (cumulative) false rejection errors which were unendurable for the verification system. The system design depicted in Figure 9-2 is proposed to enable taking advantage of spoof detection blocks and at the same time mitigating the inconvenience factor.

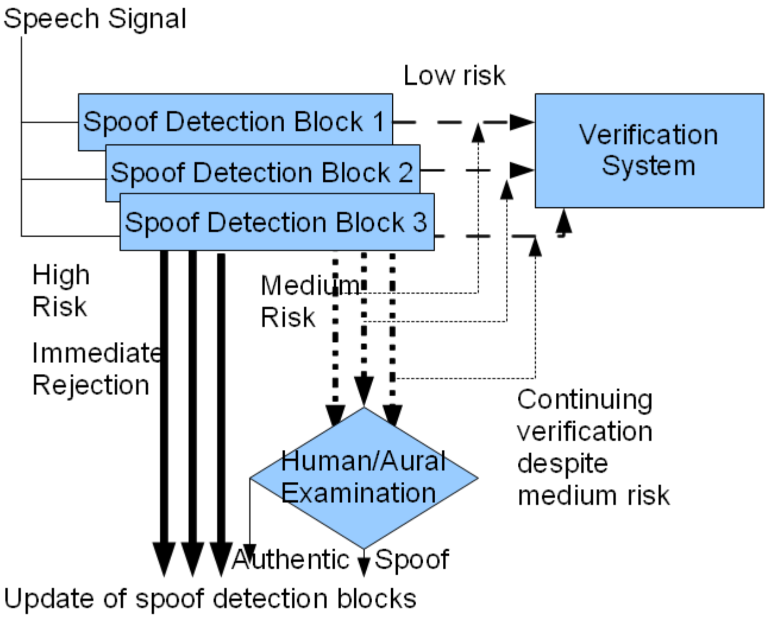


Figure 9-2 A blueprint for using and updating spoof detection blocks in collaboration with the voice verification system

In this design, instead of one rejection/acceptance threshold, two thresholds are employed by and set for the system. The result of spoof detection is any of:

- High risk (leading to immediate rejection)
- Medium risk (for which verification is continued but the signal is stored for further human examination)
- Low risk (for which verification is continued without retention of the verification signal)

High risk signals, in addition to low risk signals which are later decided to be spoof through expert / humnan examination, are used for developing better spoof detection blocks and adjustment of classification thresholds as demonstrated in chapter 8.

From the system performance perspective immediate FAR and FRR verification errors are decided by the first threshold (high risk which leads to immediate rejection).

The above scheme incorporates the concepts presented in chapter 8 and opens doors to exploiting the power of human supervision in a selective and adjustable manner in the fight against fraud.

Several other lessons were learnt through the experiments and investigations that are summarised here for this group of scenarios:

1. In chapter 4, the need for subset based evaluation was emphasised. The obvious yet important question to be answered here, is how the overall verification errors are determined after such an evaluation. The answer is that there is no reliable one-value FAR and FRR for the entire system. It is the frequency of use of the system in different conditions that decides the incurred errors. For example, if the system is used in 80% of the cases in noisy environments, the actual error rates are decided by the errors of the noisy subsets[1].

2. It is notable that human supervision could be modeled in the same evaluation model used so far. Based on this view, human interaction during verification is seen as a spoof detection procedure.

3. Experiments, especially those reported in chapter 8, hinted that security is not a life-time claim for a verification system. Systems which are purported to be secure need to go through the experiments outlined in chapters 4 to 8. In addition, there is no end to the list of possible types of spoof attacks and with the design proposed in Figure 9-2, a verification system should continually be enhanced over time. In view of the above fact, the costs that have to be taken into consideration for development, use and maintenance of such systems are different in nature and could be placed in a variety of categories. Those costs include:

- Cost of implementation
- Hidden cost imposed on users due to the time they have to spend to enroll in the system and the time spent during verification attempts
- Cost of losing credibility with the users

[1] It is still possible to report a one value error assuming that prevalence of use is known. The overall FAR for example is determined as follows: $FAR = \frac{\sum_{i=1}^{N} P_i FAR_i}{\sum_{i=1}^{N} P_i}$ where P_i is the prevalence or frequency of occurrence of signals in conditions similar to subset *i* and FAR_i is the FAR for subset i.

- Cost associated with false acceptance (misuse of data)
- Cost of surveying human perception for items in table 9-1
- Costs of educating people in the use of the new technology
- Costs of system's impact on social interactions caused by possibility of use of ordinary data for conducting biometric spoof attacks (leading to increase in fear of crime)
- Cost of system maintenance
- Costs of development of spoof detection blocks
- Costs of human supervision and keeping the spoof detection blocks up-to-date

4. It is worth mentioning that there are 3 concepts that are discussed in the context of biometrics security but do not provide any advantage in the area of security of verification system in the sense addressed in chapter 4 onward. These concepts are biometric encryption[1], biometric watermarking[2] and revocable biometrics[3]. **Regardless of the efficiency of these methods they provide absolutely no defence against spoof attacks, where the signal that reaches the receiver is already manipulated.**

5. The choice between using either text-independent or dependent verification is crucial. Up to this point, it was argued that text-independent speaker verification (in text-prompting set-up) is more secure since it rules out replay attack. Despite that and if we question the security of unsupervised verification (then for supervised verification) text-dependent and independent verification provide the same level of security: if the verification is in person and supervised, the chance to play a piece of speech or use a conversion device is zero. In long distance supervised voice verification the agent on the phone could detect the change in the voice or replay of a piece of speech. If a text-dependent recognition system could offer a considerable advantage, it might still be considered.

6. The semi-supervised systems prompt for a phrase in each verification attempt. Building the system on prompt of 'a string of digits' should strongly be avoided since a simple digit concatenation module with a very limited database of samples of speech from the speaker

[1] Biometric encryption involves use of a biometric sample as a key for carrying out encryption/decryption of data (Adler 2008).

[2] Watermarking involves embedding hidden messages in the signal. Watermarking could be helpful to check if the device used for data collection is the same as the one installed in the first place.

[3] Revocable biometrics is a title given to biometric data stored in a distorted form rather than raw form. During the enrollment the biometric is distorted and stored in the database. During verification the input biometric record undergoes the same distortion process and comparison is performed on the distorted data (see Adler 2008).

(composed of only 10 digits) could enable an effective spoof attack in this case. Owing to the natural silence between digits, the detection of spoofing is virtually impossible. The chosen phrases should facilitate spoof detection by having enough speech components and bearing continuity in their structure.

6. The features discussed in this thesis were all derived from spectral components and in spite of differences, shared the same underlying building blocks for extraction. We know that these features are vulnerable to noise and channel distortion. Further research should be directed towards detection and extraction of cues in speech. Speaker specific cues are less vulnerable to being lost in noise or being affected by channel. Other features such as chaotic features (e.g. Lyapunov exponent) that are inherently different from spectral features could provide additional information and reduce spoof threats. Nonetheless, since these features should undergo academic scrutiny and be discussed in academic journals and in the texts accessible to the public, spoof techniques which could target these methods will be developed as time goes by.

7. The systems that ensure recording of a piece of speech provide a higher level of security compared to those which require sending a speech file. The algorithms implemented in chapter 7 showed that producing fake features is much easier than producing a signal which has those features.

8. There is an individuality element in the evaluation of speaker recognition which was not analyzed in this work. While the population may undergo some changes, the changes for one individual may be more drastic. When a population's FRR is for instance 10% in a mismatch condition, it does not necessarily mean that FRR is 10% for each speaker. One speaker may be unable to be verified in such circumstances at all. The statistical analyses for each speaker (as done in forensic speaker recognition) are necessary to evaluate the speaker specific effects.

9. This study calls for the collection of a standard spoof database consisting of various subsets and spoof simulation blocks with the aim of facilitating the evaluation of spoof detection algorithms and the security of voice verification systems.

10. Subsequent to the completion of evaluation tests the results should be translated into comprehensible recommendations to the public. Educating people is an indispensable part of the deployment process. The system developers should clearly specify where the verification should

not be attempted (e.g. in noisy places such as subways[1], etc.) and what kinds of suspicious communications with strangers should be avoided.

11. In line with the last recommendation and as biometrics may give the false sense of security without offering the real benefit, it is essential to ensure that necessary information about the real extent of the security provided and the truth about the possibility of imposture is passed on to the users.

12. No experiment was conducted on mimicry in this research. Generalization of the experiments carried out by Lau et al. (2005)-reported previously-has a disappointing message for unsupervised (or even semi-supervised) verification[2]. While the results could be disputed in many ways[3] they imply that either mimicry detection should be carried out or the strength of the system against mimicry should be improved[4].

13. When identity management systems are used, it is essential to make sure that a) the identity management systems generate tokens which carry no information about actual user data; b) service providers do not receive biometric data; and c) biometric service providers go through the tested, logged and certified channels for accessing the biometric data (certified services).

14. Unless suggestions are made for the detection of purposely altered voice, there is little or no opportunity for use of voice in applications which call for negative authentication especially at distance.

9.5.3 Voice Signature

A mention of the voice signature was made in chapter 3, and a proposition for voice signature (which satisfies the requirements of electronic signature) is presented in Appendix L.

It is shown in the appendix that the use of voice signature, according to this scheme, has two great advantages over other biometric signatures:

[1] The underground!

[2] They showed that mimicry is successful if either speakers are close in terms of GMM models / utterances or imitators use linguistic knowledge.

[3] They chose the closest speaker by use of the same modeling technique. This hardly represents any real situation. The imitators were allowed to hear the verification phrase from the true speaker 3 times. These arrangements do not reflect the practical conditions. Their results show that the FAR for male linguist is not as high as female linguist which is similar to the results obtained for the closest speaker (female linguist imposed higher FAR) which suggests that the high FAR should be attributed other factors.

[4] By use of various features for verification and methods for inconsistency detection and by adding other security measures (e.g. testing knowledge of the speaker) to deter random attempts of mimicry.

1. Voice signature could be attached to the document and contain some information about the contents of the document in a way that it is impossible to be realised by most of the biometrics.

2. It satisfies the requirements of an electronic signature for being ‘in sole possession of the owner’. Unlike fingerprints and most biometrics, after the signature is revealed to a party, it can not be attached to any other document except the one for which it is generated because of the difference in the message content.

Spoof threats are minimal in the case of a voice signature since the audio message could be verified aurally[1]. In addition, noise, channel and style are under control in this case.

Therefore, this application has a promising future is civil and legal contexts.

9.6 Automatic Speaker Recognition in a Forensic Context

Two groups of scenarios were placed in this category: conviction or elimination of the suspects and identification by lay-persons and through voice parades.

The contribution of this thesis to these scenarios is the creation and testing of the evaluation framework. The evaluation framework was independent of the underlying process of assigning a score to a recording. Regardless of whether expert opinion, semi automatic approaches, automatic approaches or even aural verification by a lay-person is used for verification, the result could always be quantified. Hence, these methods could and need to undergo the same evaluation processes which were outlined in chapters 4 to 8.

As described in chapter 3, and the appendices, some of the forensic methods, especially those employed by experts, have not gone through rigorous scientific and statistical tests.

It is noteworthy that the experiments reported in this thesis were carried out on a ‘population’. For forensic speaker verification or identification the subject is an ‘individual’. Therefore, instead of the genuine/impostor populations the subsets should be built for individuals/impostors.

Experiments in chapter 7 revealed that if the GMM approach, along with cepstral coefficients, are used, the mismatch in noise, coding, channel, distance of recording and style of speaking, creates distortion in the probability distribution functions which leads to the production of unreliable likelihood ratios. The distortion was so evident that running any statistical test to show that two distributions were the same was unnecessary.

[1] In spite of the possibility of aural detection, all methods and discussions presented in chapter 8 for automatic spoof detection are relevant to this case.

When the populations' score distribution changes it implies that the individuals' score distributions have necessarily changed (if the population's distribution of scores does not change, it does not mean that the individual's scores have not changed. Individuals' scores may change in a way that the overall statistics for the population remains the same). This fact re-affirms that on an individual level, the adverse conditions are detrimental to the reliability of decisions.

Experiments with Dendrograms in chapter 5 demonstrated that even for the very clean recordings collected in a controlled environment (which is hardly available in practical conditions) when the number of sentences from the speaker decreased to 2 sentences in each segment, wrong decisions in the association task (finding two closest segments from a set of segments from 8 speakers) started to appear.

These facts suggest that the use of forensic evidence as 'proof' faces two hurdles: firstly the proof of concept is not yet established for the forensic methods (the procedures have not gone through rigorous tests), and secondly, available data in real situations usually does not meet the requirements of a reliable decision and do not permit reporting trustable likelihood ratios.

Formant features, as the most favoured acoustic forensic features, are affected by channel, noise, coding and style of speaking. For text-independent analysis, the results cited in chapter 3 and the appendices (Kinnuen (2004) and Becker et al. 2008) show that formants do not produce better verification results than cepstral features in normal conditions.

To recapitulate this topic, it should be pointed out that the forensic methods (including lay person identification) should undergo the same evaluation processes in different conditions outlined in this thesis, make certain in which of those conditions they withstand tests, how they prove similarity of conditions (match of training and test conditions), and how much data is necessary in minimum in reliable condition of work. In light of the above mentioned facts, it is admissible that concern 21 (about interpretation of biometric evidence in courts) in this domain is valid because not-adequately-tested forensic methods could contravene the rights of the suspects and lead to verdicts that do not reflect the truth of the situation.

9.7 Speech for Crime Investigation

Two groups of scenarios were introduced and placed in this category in chapter 3: determining the speaker' profile and deciphering contents of a piece of speech. The contribution of this research to these areas was minimal except for non-technical dimensions. Therefore, this section will be short.

With regard to determining the speaker's, profile it should be noted that such information (race, gender and social background) is sensitive information and, if excluding exceptional conditions, they should not be used for access control or prohibition. In crime investigation, on the other hand, speech processing provides strong tools to assist crime investigators. As specified in chapter 3, wrong information if not used as evidence in the court could be at worst misleading.

Deciphering contents of a piece of speech is in the realm of speech recognition and has had less discussion in this thesis. Modeling techniques discussed in chapter 3 and its associated appendices (such as HMM modeling) along with the same cepstral features are widely used for speech recognition. Therefore, automatic speech recognition faces the same challenges in adverse conditions that were elaborated on in chapter 8 but use of voice, in both scenarios put in this category is technically possible and has a relatively bright future.

9.8 Speech for Surveillance

Speech-based surveillance is usually carried out to reduce crime or fraud. Therefore, applications which are of surveillance descent could also be placed in the first category. Due to the importance of non-technical issues surrounding surveillance a whole section and category is devoted to voice surveillance.

Two groups of scenarios were placed in this category:

- Eavesdropping / wire-tapping (any interception of communications)
- Audio surveillance for public safety and security

Before dealing with surveillance issues, it is instrumental to try to discriminate between goals and levels of audio surveillance and interceptions. Based on whether the surveillance system aims at extraction of information about the content of speech, supplementary data about the speech / speaker or supplementary data about sounds and their causes the conceivable cases could be presented as follows:

1. Content of speech (with or without identification of people in the conversations) is analysed (automatically or by a person) through interception of communications or surveillance

2. Speech is examined for the existence of a number of sensitive words or phrases (e.g. the vocabulary employed for talking about a particular subject e.g. terrorist activity)

3. Abnormal style of speaking is identified (high pitch, loud, arrogant speaking implying violence e.g. violence against children and women)

5. Auxiliary information is extracted from a particular conversation without attention to its content, e.g. social background or race information. The data could be used in statistical analysis (e.g. the demographics of a population), for access control or putting the system in an 'alert' status.

6. Changes in the patterns of sounds and noises is monitored (examples include triggering warnings about gun shots, window smashing, falling of a person). This type of surveillance could go hand in hand with video surveillance. Subsequent to the recognition of specific sounds the video recording is started or the videos are set aside for further investigation.

7. Speaker (and not the content of speech) matters and is identified. Auxiliary information about the location, date and time of the conversation is recorded (examples include electronic monitoring of offenders and those under house arrest).

In the above cases, the technical questions are speech, speaker and noise / sound pattern recognition. Most problems in this category are non technical and fall into the surveillance related concerns introduced in Table 9-1.

There is normally little justification for voice recording in a civil context especially when the 'content of speech' is recorded and analysed. UK's ICO code of conduct introduced in chapter 3 (which complies with DPA Act 1998) states that the recording of voice for surveillance is normally prohibited.

If voice verification could be justified, ICO's recommendations for video recording could be customised and applied as expanded on here:

1. The level of sound quality that will be necessary to achieve the specified purpose should be identified and adjusted. For example, for sound recognition (gun shots or a car crash) the data could be filtered in a way that human speech is unperceivable in the recordings.

2. The purpose of audio recording should be specified and justified. The code of conduct suggests making distinction between monitoring, detection, recognition and identification. Similarly, for voice, the distinction between seven audio surveillance levels presented above is helpful for making justifications about the set-up and procedures.

3. There is no difference between voice and audio in terms of documentation, requirements for storing / access, disclosure, retention, letting people know, subject access requests and other responsibilities outlined in the ICO's code of conduct.

4. Similar to video surveillance, continuous monitoring (e.g. monitoring of a workforce) should only be used in very exceptional circumstances, for example, where sound recording could prevent a hazard.

5. Devices installed for preventing and detecting crime should not be used for non-criminal matters.

As for now, it seems that in a civil context, the only application that seems to be justified and is less controversial is audio-dependent warning systems which are triggered by changes in sound patterns.

Covert and directed surveillance conducted by state agencies are normally governed by different rules and regulations. Interception of communications in many ways and for numerous reasons is permitted by law in the UK as described in Regulation of Investigatory Powers Act (2000). In the scenarios that notification does not conflict the purpose of interception (and does not provide opportunities for criminals to avoid identification) use of the scheme expanded on in 9.5 is extremely instrumental in reducing consequences of public surveillance. The scheme mandates that whenever biometric services are used for identification, the information about the process (comprising the service consumer and service provider) is logged in a way that enables the person to check and be aware of it. The impression of being in control of one's own data and processes could be contrasted with the impression of being permanently under surveillance by unknown parties and in different situations without being aware of it.

9.9 Areas of Further Research

Not every issue related to voice verification has apparently been undertaken here. Numerous research questions are still there for further investigation and understanding. These will be discussed here in two groups; one technical and the other, non-technical issues.

Technical issues in need of further attention are:

- Reducing the effect of channel / noise and proposition of better features capable of withstanding mismatch conditions are among the matters that deserve constant attention. Cue-based recognition, which requires identification of subtle differences in individuals' speech and proposition of features to capture those cues, could go a long way toward reducing channel effects and spoof threats. Translation of human authentication procedures into these cues and studying the process of human voice verification merits further research. Combination of cue-based and statistical voice verification, could limit the room for spoof, as for being successful in this case, the counterfeit speech needs to exhibit both the overall statistical resemblance and specific cues.
- While interference (simultaneous speech) posed a less negative effect compared to noise when the main speech was dominant, there are still several research questions that need to be answered in the area of speech separation. The examples include adaptation of ICA for cases with two non-identical channels and speech separation when only one channel is available. While several methods are suggested for these tasks in some cases the speaker recognition results are not yet reported after separation.
- Introduction of other types of features which are less dependent on spectral components e.g. chaotic features and combinations of them with available features, is another area of improvement which could affect both security and reliability of voice verification.
- Researchers in the domain of speaker recognition should always keep an eye on the developments in speech science, especially speech synthesis and conversion. The evaluation system should always be kept up-to-date and spoof simulation and spoof detection modules should be developed alongside the advances in speech science.
- Only two of the methods proposed for spoof detection in chapter 8 were elaborated on and developed in this piece of work. Research on the success rate of spoof detection, based on the other proposed ideas, needs to be undertaken. This includes the use of various features

for inconsistency analysis, proposition of new metrics for discontinuity detection and devising techniques for telling speech and non speech signals apart.

- Clarification of the technical details concerning the implementation of the blueprint presented in Figure 9-1 in compliance with available security and web-service standards is the topic for another project.
- If multi-biometrics are going to be used in addition to voice (e.g. lip reading) the same level of scrutiny, especially concerning spoof possibilities, should be exercised for other biometrics.
- Recognition of special patterns of sound and the detection of emotions is of great value to sound-dependent surveillance.
- Detection of mimicry is a must if voice is going to be used for negative authentication. Unless we attain features which could not be voluntarily affected by people, the only chance to block off this type of security breach is through mimicry detection and little has been done in this area.
- Complete evaluation of the systems making use of other normalization techniques (different from global model normalization) in all of the conditions outlined in this thesis needs to be undertaken.
- Forensic methods should undergo the same tests that were carried out here for automatic verification by acoustic features. Formant dynamics, combination of formants and cues in formants are great sources of information about the speaker's identity. Yet the main question that needs to be addressed is how the formant-based analysis (which normally requires manual and laborious hand-labeling of data) could be used for automatic recognition.
- While the studies offered here were conducted on a 'population' the same analyses should be carried out for 'individuals' in forensic speaker recognition.

Several non-technical issues are worth close attention:

- The question of whether the use of biometrics for surveillance and the interception of communications is justifiable in various cases is and will be there forever. A clear definition of where privacy turns into an undeniable right (rather than an interest) which the majority could not contravene is necessary.
- Devising innovative methods to elicit the public's opinions on the items presented in Table-1, quantifying the opinions and comparing the severity of concerns with benefits gained

through use of biometrics (especially for surveillance) is vital. This is an invaluable step towards standardizing the use of biometrics in both technical and non-technical domains.

- Responsibilities of biometric certification authorities need to be clarified as well as rules governing their actions. Necessary regulations in this regard should be discussed and passed.
- Solid strategies for dealing with the consequences of false acceptance errors need to be developed in carefully designed frameworks. As mentioned before, these errors, as well as spoof possibilities, are as inherent to the utilization of biometrics as they are to the use of credit cards.
- Further research is required to gauge the effect of the use of biometrics on human interactions with a close examination of the psychology of human behaviour when spoof attacks are brought into the picture.

9.10 Final Comments

This research demonstrated that despite the obstacles on the road of the use of voice verification especially the severity of spoof threats, contingent upon adopting proper designs and suitable strategies, voice enabled authentication systems could live a healthy and prosperous life in the future.

Nevertheless, it could never be emphasised enough that voice is not an immaculate identification tool and the expectation about the reliability of decisions made through analysis of voice as a piece of evidence in all applications mentioned hitherto should be moderated.

The best way to attain a realistic picture of how reliable such decisions are is through a comprehensive evaluation strategy such as the one elaborated on here. This is crucially essential in the case of forensic speaker recognition and when evidence is going to be brought into court.

On the other hand, the decision about the merits of the introduction of a voice-based security measures should only ensue full understanding of the various technical and non-technical costs involved in the process of development and maintenance of the system including those related to continuous system evaluation and (manual and automatic) spoof detection; in other words the costs of upholding the arms race between the system owners and impostors.

It could be concluded that due to the possibility of spoof and importance of convenience, in the future, voice verification is only likely to be used in supervised and semi-supervised applications in the design proposed in this chapter and co-evolution of spoofing techniques and methods of detection of attacks maintains the balance between the two sides so that remote unsupervised voice authentication will not qualify as the sole security measure for unconstrained use unless a degree of risk is tolerable for the system.

Bibliography

Abdulla, W. & Kasabov, N. (1999), 'The Concepts of Hidden Markov Model in Speech Recognition', Technical report, Technical Report TR99/09, NK Department of Knowledge Engineering Lab Information Science Department University of Otago New Zealand.

Adami, A.; Mihaescu, R.; Reynolds, D. & Godfrey, J. (2003), 'Modeling prosodic dynamics for speaker recognition', Proc. ICASSP, 788--791.

Adler, A. (2008), 'Biometric System Security', *Handbook of Biometrics*, 381--402.

Adoul, J.; Mabilleau, P.; Delprat, M. & Morisette, S. (1987), 'Fast CELP coding based on algebraic codes', Proc. ICASSP, 1957--1960.

Alderman, E. & Kennedy, C. (1997), *The right to privacy*, Vintage Books.

Alexander, A.; Botti, F.; Dessimoz, D. & Drygajlo, A. (2004), 'The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications', *Forensic science international* **146**, 95--99.

Amano-Kusumoto, A.; Hosom, J. & Shafran, I. (2009), 'Classifying clear and conversational speech based on acoustic features', *Interspeech 2009, Brighton.*

American Board of Recorded Evidence (1999), 'Voice Comparison Standards', Last Retrieved 2010-04-01, <http://www.forensictapeanalysisinc.com/Articles/voice_comp.htm>.

Anderson, D.; Sweeney, D. & Williams, T. (1994), *Introduction to statistics: concepts and applications*, West Group.

Antoniou, G.; Sterling, L.; Gritzalis, S. & Udaya, P. (2008), 'Privacy and forensics investigation process: The ERPINA protocol', *Computer Standards & Interfaces* **30**(4), 229--236.

APACS (2006), 'Fraud: The Facts 2006', APACS, <*www.apacs.org.uk*>

Aronowitz, H.; Burshtein, D. & Amir, A. (2004), 'Speaker indexing in audio archives using test utterance Gaussian mixture modeling', *Interspeech-2004*, 609--612.

Auckenthaler, R.; Carey, M. & Lloyd-Thomas, H. (2000), 'Score normalization for text-independent speaker verification systems', *Digital Signal Processing* **10**(1-3), 42--54.

Aungles, A. & Cook, D. (1994), 'Information Technology and the Family', *Information Technology & People* **7**(1), 69--80.

Barras, C. & Gauvain, J. (2003), 'Feature and score normalization for speaker verification of cellular data', Proc. ICASSP', Citeseer, 49--52.

Baumer, T.; Maxfield, M. & Mendelsohn, R. (1993), 'A comparative analysis of three electronically monitored home detention programs', *Justice Quarterly* **10**(1), 121--142.

Becker, T.; Jessen, M. & Grigoras, C. (2008), 'Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models', *Proceedings of Interspeech 2008 Incorporating SST 2008*, 1505--1508.

Beyer, W. (1968), *Handbook of tables for probability and statistics*, CRC Press.

Biometric Information Technology Ethics (BITE) (2006), '2nd scientific Project Meeting', Last Retrieved 2010-04-27, <http://www.biteproject.org/documents/BITE_FINAL_CONFERENCE_PRESENTATIONS.zip>.

Biometric Information Technology Ethics (BITE) (2006), 'Ethical and Social Implications of Biometric Identification Technologies', Last Retrieved 2010-05-01, <http://www.biteproject.org/documents/PUBLIC_CONSULTATION_REPORT.pdf>.

Biometric Technology Today (2002), 'Alternative Biometrics', *Biometric Technology Today*.

Biometric Technology Today (2006), 'Biometric standards - An update', *Biometric Technology Today* **14**(1), 10 - 11.

Black, J.; Lashbrook, W.; Nash, E.; Oyer, H.; Pedrey, C.; Tosi, O. & Truby, H. (1973), 'Letter: Reply to" speaker identification by speech spectrograms: some further observations".', *The Journal of the Acoustical Society of America* **54**(2), 535.

Boccardi, F. & Drioli, C. (2001), 'Sound morphing with Gaussian mixture models', Proc. of the 4th COST G-6 Workshop on Digital Audio Effects (DAFx01), Limerick, Ireland.

Bolt, R.; Cooper, F.; David Jr, E.; Denes, P.; Pickett, J. & Stevens, K. (1973), 'Speaker identification by speech spectrograms: Some further observations', *The Journal of the Acoustical Society of America* **54**, 531--534.

Bolt, R.; Cooper, F.; David Jr, E.; Denes, P.; Pickett, J. & Stevens, K. (1970), 'Speaker identification by speech spectrograms: A scientists' view of its reliability for legal purposes', *Journal of the Acoustical Society of America* **47**(2), 2.

Bonafonte, A.; Hoge, H.; Kiss, I.; Moreno, A.; Ziegenhain, U.; van den Heuvel, H.; Hain, H.; Wang, X. & Garcia, M. (2006), 'TC-STAR: Specifications of language resources and evaluation for speech synthesis', Proc. of LREC Conf.

Bonastre, J.; Bimbot, F.; Bo\'eb, L.; Campbell, J.; Reynolds, D. & Magrin-Chagnolleau, I. (2003), 'Person authentication by voice: a need for caution', Eighth European Conference on Speech Communication and Technology.

Botti, F.; Alexander, A. & Drygajlo, A. (2004), 'On compensation of mismatched recording conditions in the bayesian approach for forensic automatic speaker recognition', *Forensic science international* **146**, S101--S106.

Bregman, A. (1990), *Auditory scene analysis*, The MIT Press.

Britanak, V.; Yip, P. & Rao, K. (2007), *Discrete cosine and sine transforms: general properties, fast algorithms and integer approximations*, Academic Press.

Broeders, A. (2001), 'Forensic speech and audio analysis forensic linguistics', Proc. 13th INTERPOL Forensic Science Symposium', 16--19.

Brookes, M. (2004), 'VOICEBOX', Last Retrieved 2010-5-30, <http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html> .

Broun, C.; Zhang, X.; Mersereau, R. & Clements, M. (2002), 'Automatic speechreading with application to speaker verification', IEEE International Conference on Acoustics Speech and Signal Processing.

Brown, G. & Wang, D. (2005), 'Separation of speech by computational auditory scene analysis', *Speech enhancement*, 371--402.

Bunnel, T. (2000), 'Voicing Source', Last Retrieved 2010-04-09, <http://www.asel.udel.edu/~bunnell/aip-cd/source.htm>.

Byrne, C. & Foulkes, P. (2007), 'The'mobile phone effect'on vowel formants', *International Journal of Speech Language and the Law* **11**(1), 83.

Cabinet Office (UK) (2002), ‘Identity Fraud, A Study’, Last Retrieved: 2010-04-17, <www.statewatch.org/news/2004/may/id-fraud-report.pdf>.

Caliński, T. & Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in Statistics-Simulation and Computation* **3**(1), 1--27.

Cameron, K. (2005), 'The Laws of Identity’, Last Retrieved 2010-04-20, <http://www.identityblog.com/stories/2005/05/13/TheLawsOfIdentity.pdf>.

Cardoso, J. (1998), 'Blind signal separation: statistical principles', *Proceedings of the IEEE* **86**(10), 2009--2025.

Carey, M.; Parris, E.; Lloyd-Thomas, H. & Bennett, S. (1996), 'Robust prosodic features for speaker identification’, Fourth International Conference on Spoken Language Processing.

Carotti, E.; De Martin, J.; Merletti, R. & Farina, D. (2007), 'Compression of surface EMG signals with algebraic code excited linear prediction', *Medical Engineering and Physics* **29**(2), 253--258.

Carreira-Perpinan, M. (2000), 'Mode-finding for mixtures of Gaussian distributions', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(11), 1318--1323.

Carreira-Perpinan, M. (2001), 'Continuous latent variable models for dimensionality reduction and sequential data reconstruction', PhD thesis, University of Sheffield.

Champod, C. & Meuwly, D. (2000), 'The inference of identity in forensic speaker recognition', *Speech Communication* **31**(2-3), 193--203.

Charkani, N. & Deville, Y. (1997), 'Optimization of the asymptotic performance of time-domain convolutive source separation algorithms’, ESANN, 273--278.

Chazan, D.; Hoory, R.; Cohen, G. & Zibulski, M. (2000), 'Speech reconstruction from mel frequency cepstral coefficients and pitch frequency’, IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING, IEEE; 1999.

Chen, H. & Jain, A. (2008), 'Automatic Forensic Dental Identification', *Handbook of Biometrics*, 231--251.

Chen, W.; Hsieh, C. & Lai, E. (2004), 'Multiband Approach to Robust Text-Independent Speaker Identification', *Computational Linguistics and Chinese Language Processing* **9**(2), 63--76.

Christmann, K. & Rogerson, M. (2004), 'Crime, fear of crime and quality of life identifying and responding to problems', Sheffield Hallam University, Centre for Regional Economic and Social Research.

Clarke, R. (1994), 'Human identification in information systems: Management challenges and public policy issues', *Information Technology & People* **7**(4), 6--37.

Clarke, R. (2001), 'Biometrics and privacy', Last Retrieved 2010-04-25, <http://www.rogerclarke.com/DV/Biometrics.html>.

Clarke, R. (2006), 'Introduction to Dataveillance and Information Privacy, and Definitions of Terms ', *Last Retrieved 2010-04-25, <http://www.rogerclarke.com/DV/Intro.html>.*

Clauß, S. & Kohntopp, M. (2001), 'Identity management and its support of multilateral security', *Computer Networks* **37**(2), 205--219.

Clifford, B.; Rathborn, H. & Bull, R. (1981), 'The effects of delay on voice recognition accuracy', *Law and Human Behavior* **5**(2), 201--208.

Cole, R.; Mariani, J.; Uszkoreit, H.; Zaenen, A.; Zue, V. & others (1997), *Survey of the state of the art in human language technology*, Citeseer.

Colibro, D.; Vair, C.; Castaldo, F.; Dalmasso, E. & Laface, P. (2006), 'Speaker recognition using channel factors feature compensation', EUSIPCO-2006.

Common Criteria (2009), 'Common Criteria for Information Technology Security Evaluation: Part 1: Introduction and general model', Version 3.1, Last Retrieved 2010-03-21, <http://www.commoncriteriaportal.org/thecc.html>.

Common Criteria Biometric Evaluation Methodology Working Group (2002), 'Common Criteria, Common Methodology for Information Technology Security Evaluation', Version 1.0, Last Retrieved 2010-03-21, <http://www.cesg.gov.uk/policy_technologies/biometrics/media/bem_10.pdf>.

Council_of_Europe (1950), 'Convention for the Protection of Human Rights and Fundamental Freedoms', Last Retrieved 2010-04-25, <http://conventions.coe.int/treaty/en/Treaties/Html/005.htm>.

Coventry, L. (2004), 'Fingerprint authentication: The user experience', *presented at the DIMACS Workshop on Usable Privacy and Security Software*.

Cummins, F.; Grimaldi, M.; Leonard, T. & Simko, J. (2006), 'The CHAINS corpus: Characterizing individual speakers', Proc of SPECOM, 431--435.

Damper, R. & Higgins, J. (2003), 'Improving speaker identification in noise by subband processing and decision fusion', *Pattern Recognition Letters* **24**(13), 2167--2173.

Dass, S.; Nandakumar, K. & Jain, A. (2005), 'A principled approach to score level fusion in multimodal biometric systems', Audio-and Video-Based Biometric Person Authentication, Springer, 1049--1058.

Daugman, J. (2000), 'Biometric decision landscapes', *The Computer Laboratory, Cambridge University*.

Daugman, J. (2004), 'Combining multiple biometrics', *The Computer Laboratory, Cambridge University*.

Deoras, A. & Hasegawa-Johnson, M. (2004), 'A factorial HMM approach to simultaneous recognition of isolated digits spoken by multiple talkers on one audio channel', Proc. ICASSP, 861--864.

De-Wet, F.; de Veth, J.; Boves, L. & Cranen, B. (2005), 'Additive background noise as a source of non-linear mismatch in the cepstral and log-energy domain', *Computer Speech & Language* **19**(1), 31--54.

Dimitriadis, C. & Polemi, D. (2006), 'An identity management protocol for Internet applications over 3G mobile networks', *Computers & Security* **25**(1), 45--51.

Drygajlo, A.; Meuwly, D. & Alexander, A. (2003), 'Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition', Eighth European Conference on Speech Communication and Technology.

Durbin, R.; Eddy, S.; Krogh, A. & Mitchison, G. (1998), 'Markov chains and hidden Markov models', *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, 46--79.

Ekblom, P. (1999), 'Can we make crime prevention adaptive by learning from other evolutionary struggles?', *Studies on Crime and Crime Prevention* **8**, 27--51.

Ekudden, E.; Hagen, R.; Johansson, I. & Svedberg, J. (1999), 'The adaptive multi-rate speech coder', Proc. IEEE Workshop on Speech Coding, 117--119.

Endres, W.; Bambach, W. & Flosser, G. (1971), 'Voice spectrograms as a function of age, voice disguise, and voice imitation', *The Journal of the Acoustical Society of America* **49**, 1842.

Eriksson, A. (2005), 'Tutorial on forensic speech science', Interspeech, Lisbon, Portugal.

Eskénazi, M. (1992), 'Changing Speech Styles: Strategies in Read Speech and Casual and Careful Spontaneous Speech.', ICSLP 92, 755--758.

Europay, Master, Visa (EMV) (2000), *Integrated Circuit Card, Specification for Payment Systems, Book*

3, Application Specification, EMVCo.

European_Group_on_Ethics_in_Science_and_New_Technologies (EGE) (2000), 'Citizens' rights and the new technologies: a European challenge', Last Retrieved 2010-04-23, <http://ec.europa.eu/european_group_ethics/docs/prodi_en.pdf>.

European_Parliament (2000), 'Charter of Fundamental Rights of the European Union', *Official Journal of the European Communities,* Last Retrieved 2010-04-23, <http://www.europarl.europa.eu/charter/pdf/text_en.pdf>.

Fant, G. (1973), 'Acoustic description and classification of phonetic units. Reprinted in Speech Sounds and Features', Cambridge: MIT Press.

Fant, G.; Liljencrants, J. & Lin, Q. (1985), 'A four parameter model of glottal flow', *STL-QPSR 4/1985.*

Farrus, M.; Wagner, M.; Anguita, J. & Hernando, J. (2008), 'How vulnerable are prosodic features to professional imitators?', Proc. IEEE Workshop on Speaker and language Recognition, Stellenbosch, South Africa.

Faundez-Zanuy, M.; Hagmuller, M. & Kubin, G. (2006), 'Speaker verification security improvement by means of speech watermarking', *Speech Communication* **48**(12), 1608--1619.

Foulkes, P. & French, J. (2001), 'Forensic phonetics and sociolinguistics', *In Mesthrie, R. (ed.) Concise Encyclopedia of Sociolinguistics. Amsterdam: Elsevier Press.*, 329--332.

Fukada, T.; Tokuda, K.; Kobayashi, T. & Imai, S. (1992), 'An adaptive algorithm for mel-cepstral analysis of speech', Proc. ICASSP', 137--140.

Furnell, S. & Evangelatos, K. (2007), 'Public awareness and perceptions of biometrics', *Computer Fraud & Security* **2007**(1), 8--13.

Furui, S. (2001), *Digital speech processing, synthesis, and recognition*, CRC.

Furui, S. (2005), '50 years of progress in speech and speaker recognition', Proceedings of the 10th International Conference on Speech and Computer (SPECOM)', 1--9.

Gamma, E.; Helm, R.; Johnson, R. & Vlissides, J. (1995), *Design patterns: elements of reusable object-oriented software*, Addison-wesley Reading, MA.

Ghahramani, Z. & Jordan, M. (1997), 'Factorial hidden Markov models', *Machine learning* **29**(2), 245--273.

Gonzalez-Rodriguez, J.; Drygajlo, A.; Ramos-Castro, D.; Garcia-Gomar, M. & Ortega-Garcia, J. (2006), 'Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition', *Computer Speech & Language* **20**(2-3), 331--355.

González-Rodriguez, J.; Ortega-Garcia, J.; Martin, C. & Hernández, L. (1996), 'Increasing robustness in GMM speaker recognition systems for noisy and reverberant speech with low complexity microphone arrays', Fourth International Conference on Spoken Language Processing', Citeseer.

Gordon, A. (1994), 'Identifying genuine clusters in a classification', *Computational Statistics & Data Analysis* **18**(5), 561--581.

Gordon, G. & Willox Jr, M. (2004), 'Identity fraud: A critical national and global threat', *Journal of Economic Crime Management* **2**(1), 1--48.

Górriz, J.; Puntonet, C.; Ramirez, J. & Segura, J. (2005), 'Statistical Tests for Voice Activity Detection', ISCA Tutorial and Research Workshop (ITRW) on Non-Linear Speech Processing.

Grabe, E.; Post, B. & Nolan, F. (2001), 'The IViE corpus', *Department of Linguistics, University of Cambridge.*

Grey, G. & Kopp, G. A. (1944), 'Voiceprint identification.', *Bell Telephone Laboratories Report*, 1--14.

Griffin, P. (2004), 'Predicting Performance of Fused Biometric Systems', Identix Research, September.

Grijpink, J. (2001), 'Privacy Law :: Biometrics and privacy', *Computer Law & Security Report* **17**(3), 154--160.

Grijpink, J. (2005), 'Two barriers to realizing the benefits of biometrics-A chain perspective on biometrics, and identity fraud', *Computer Law & Security Report* **21**(2), 138--145.

Guillemin, B. & Watson, C. (2006), 'Impact of the GSM AMR Speech Codec on Formant Information Important to Forensic Speaker Identification', Proceedings of the 11th Australian International Conference on Speech Science & Technology', 483--487.

Haigh, J. & Mason, J. (1993), 'A voice activity detector based on cepstral analysis', Third European Conference on Speech Communication and Technology.

Han, R.; Zhao, P.; Gao, Q.; Zhang, Z.; Wu, H. & Wu, X. (2006), 'Casa based speech separation for robust speech recognition', Ninth International Conference on Spoken Language Processing (ISCA).

Hartigan, J. (1975), *Clustering algorithms*, Wiley New York.

Heck, L. (2004), 'On the deployment of speaker recognition for commercial applications', Proc. Odyssey Speaker Recognition Workshop.

Heck, L.; Konig, Y.; S\'f6nmez, M. & Weintraub, M. (2000), 'Robustness to telephone handset distortion in speaker recognition by discriminative feature design', *Speech Communication* **31**(2-3), 181--192.

Helsinki University of Technology (2005), 'The FastICA package for MATLAB', Last Retrieved 2010-07-15, <http://www.cis.hut.fi/projects/ica/fastica/>.

Hollien, H. & Schwartz, R. (2000), 'Aural-perceptual speaker identification: problems with noncontemporary samples', *International Journal of Speech Language and the Law* **7**(2).

Holm, S. (2003), 'Individual use of acoustic parameters in read and spontaneous speech', Proceedings of Fonetik 2003', Citeseer, 157--161.

Holmes, J. & Holmes, W. (2001), *Speech synthesis and recognition*, CRC.

Home Office (UK) (2006), 'Updated Estimate of the Cost of Identity Fraud to the UK Economy', Identity Fraud Steering Committee, Last Retrieved: 2010-04-17, <www.theirm.org/events/documents/12TwelfthMeeting-Scan003.pdf>.

Home Office Circular (2003), 'Advice on the use of voice identification parade', Last Retrieved 2010-04-02, <http://www.homeoffice.gov.uk/about-us/publications/home-office-circulars/circulars-2003/057-2003/>.

Hyvarinen, A. & Oja, E. (2000), 'Independent component analysis: a tutorial', *Neural Networks* **13**(4-5), 411--430.

Information Commissioner's Office (ICO) (2008), 'CCTV Code of Practice', Information Commissioner's Office, Last Retrieved 2010-04-27, <http://www.ico.gov.uk/upload/documents/library/data_protection/detailed_specialist_guides/ico_cctv final_2301.pdf>.

International Biometric Group (IBG) (2005), 'Biometrics & Privacy', Last Retrieved 2010-04-27, <http://www.biteproject.org/documents/BITE_FINAL_CONFERENCE_PRESENTATIONS.zip>.

International Biometric Group (IBG) (2006), 'Spoof 2007—High-Level Test Plan.', *New York, NY.*.

Israel, S.; Irvine, J.; Cheng, A.; Wiederhold, M. & Wiederhold, B. (2005), 'ECG to identify individuals', *Pattern Recognition* **38**(1), 133--142.

Jackson, W. (2007), 'Under attack: Common Criteria has loads of critics, but is it getting a bum rap', *Government Computer News*.

Jain, A. & Ross, A. (2003), 'Information fusion in biometrics', *Pattern Recognition Letters* **24**(13), 2115--2125.

Jain, A. & Ross, A. (2008b), 'Introduction to Biometrics', *Handbook of Biometrics*, 1--22.

Jain, A.; Dass, S. & Nandakumar, K. (2004), 'Soft biometric traits for personal recognition systems', Proceedings of International Conference on Biometric Authentication.

Jain, A.; Flynn, P. & Ross, A. (2008a), *Handbook of biometrics*, Springer.

Jang, G. & Lee, T. (2003), 'A maximum likelihood approach to single-channel source separation', *The Journal of Machine Learning Research* **4**, 1365--1392.

Jin, Q.; Schultz, T. & Waibel, A. (2007), 'Far-field speaker recognition', *IEEE Transactions on Audio Speech and Language Processing* **15**(7), 2023.

Jones, G. & Levi, M. (2000), 'The value of identity and the need for authenticity', *Foresight Crime Prevention Panel Essay, Turning the Corner.*

Jones, J. (2007), 'Simplifying Web Services Development with the Decorator Pattern', Oracle Technology Network, Last Retrieved 2010-03-25, <http://www.oracle.com/technology/pub/articles/jones-owsm.html>.

Jurafsky, D. & Martin (2000), *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall.

Jurafsky, D. & Martin, J. (2009), *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, Prentice Hall, Second Edition.

Kajarekar, S.; Bratt, H.; Shriberg, E. & de Leon, R. (2006), 'A Study of Intentional Voice Modifications for Evading Automatic Speaker Recognition', IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006', 1--6.

Kalman, R. (1960), 'A new approach to linear filtering and prediction problems', *Journal of basic Engineering* **82**(1), 35--45.

Karlsson, I.; Banziger, T.; Dankovicova, J.; Johnstone, T.; Lindberg, J.; Melin, H.; Nolan, F. & Scherer, K. (2000), 'Speaker verification with elicited speaking styles in the VeriVox project', *Speech Communication* **31**(2-3), 121--129.

Katz, J. & Lindell, Y. (2008), *Introduction to modern cryptography*, Chapman & Hall/CRC.

Kersta, L. (1962), 'Voiceprint identification', *The Journal of the Acoustical Society of America* **34**, 725.

Kinnunen, T. (2004), 'Spectral features for automatic text-independent speaker recognition', *Ph. Lic. Thesis, Univ. of Joensuu, Dept. of Computer Science*.

Kinnunen, T.; Chernenko, E.; Tuononen, M.; Franti, P. & Li, H. (2007), 'Voice activity detection using MFCC features and support vector machine', Int. Conf. on Speech and Computer (SPECOM 2007), Moscow, Russia', Citeseer, 556--561.

Kinnunen, T.; Karkkainen, I. & Franti, P. (2001), 'Is speech data clustered?-statistical analysis of cepstral features', Seventh European Conference on Speech Communication and Technology.

Kinser, L.; Frey, A.; Coppens, A. & Sanders, J. (2000), *Fundamentals of Acoustics*, Wiley, New York.

Klabbers, E. & Veldhuis, R. (1998), 'On the reduction of concatenation artefacts in diphone synthesis', Fifth International Conference on Spoken Language Processing', Citeseer.

Koenig, C.; Bartosch, A. & Braun, J. (2009), *EC competition and telecommunications law*, Kluwer Law Intl.

Koolwaaij, J. & Boves, L. (1999), 'On the use of automatic speaker verification systems in forensic casework', Audio-and Video-based Biometric Person Authentication', 224--229.

Koutras, A.; Dermatas, E. & Kokkinakis, G. (1999), 'Blind signal separation and speech recognition in the frequency domain', 6th International Conference on Circuits and Systems.

Koutras, A.; Dermatas, E. & Kokkinakis, G. (2000), 'Blind speech separation of moving speakers in real reverberant environments', IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING', Citeseer.

Koutras, A.; Dermatas, E. & Kokkinakis, G. (2001), 'Improving simultaneous speech recognition in real room environments using overdetermined blind source separation', Seventh European Conference on Speech Communication and Technology.

Krause, J. & Braida, L. (2004), 'Acoustic properties of naturally produced clear speech at normal speaking rates', *The Journal of the Acoustical Society of America* **115**, 362.

Krawczyk, S. (2005), 'User Authentication using On-Line Signature and Speech', Master's thesis, Michigan State University.

Kuenzel, H. (1994), 'On the problem of speaker identification by victims and witnesses', *Forensic Linguistics* **1**(1), 45--58.

Kullback, S. & Leibler, R. (1951), 'On information and sufficiency', *The Annals of Mathematical Statistics* **22**(1), 79--86.

Kunzel, H. (2001), 'Beware of the telephone effect": The influence of telephone transmission on the measurement of formant frequencies', *Forensic Linguistics* **8**(1), 80--99.

L., H. (1998), 'Automatic Personal Identification Using Fingerprints', PhD thesis, Michigan State University.

Laheld, B. & Cardoso, J. (1994), 'Adaptive source separation with uniform performance'(2)'Proc. EUSIPCO', Citeseer, 183--186.

Lamel, L. & Gauvain, J. (2000), 'Speaker verification over the telephone* 1', *Speech Communication* **31**(2-3), 141--154.

Lau, Y.; Tran, D. & Wagner, M. (2005), 'Testing Voice Mimicry with the YOHO Speaker Verification Corpus', Knowledge-Based Intelligent Information and Engineering Systems', Springer, 15--21.

Lau, Y.; Wagner, M. & Tran, D. (2004), 'Vulnerability of speaker verification to voice mimicking', Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on', 145--148.

Lazarick, R. & Cambier, J. (2008), 'Biometrics in the Government Sector', *Handbook of Biometrics*, 461--478.

Liberty Alliance Project, 'Liberty Glossary v2.0', Last Retrieved 2010-04-07, <http://www.projectliberty.org/liberty/resource_center/specifications/liberty_alliance_specifications_support_documents_and_utility_schema_files/liberty_glossary_v2_0/>.

Lindberg, J. & Blomberg, M. (1999), 'Vulnerability in speaker verification-a study of technical impostor techniques', Sixth European Conference on Speech Communication and Technology.

Lindh, J. (2004), 'Handling the Voiceprint Issue', *FONETIK 2004*, 72.

Liu, F.; Stern, R.; Huang, X. & Acero, A. (1993), 'Efficient cepstral normalization for robust speech recognition', Proceedings of ARPA Speech and Natural Language Workshop', Morgan Kaufmann, 69--74.

Liu, Y. (2007), 'Introduction to biometrics from a legal perspective', *Last Retrieved 2010-04-22, <http://www.uio.no/studier/emner/jus/jus/JUR5630/v08/undervisningsmateriale/Introduction_to_biometrics_from_a_legal_perspective-1.ppt>.*

Lu, X. & Dang, J. (2007), 'An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification', *Speech Communication*.

Luettin, J.; Thacker, N. & Beet, S. (1996), 'Speaker identification by lipreading', Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on.

M., Kamsika (2003), 'A MATLAB demonstration of Independent Component Analysis', *Undergraduate project dissertation*.

Macpherson, L. (2009), 'UK Photographers Rights Guide V.2', *Last Retrieved 2010-04-25, <http://www.sirimo.co.uk/wp-content/uploads/2009/05/ukphotographersrights-v2.pdf>.*

Maghiros, I.; Punie, Y.; Delaitre, S.; Lignos, E.; Rodriguez, C.; Ulbrich, M.; Cabrera, M.; Clements, B.; Beslay, L. & Van Bavel, R. (2005), 'Biometrics at the frontiers: Assessing the impact on society', *EUROPEAN COMMISSION, JOINT RESEARCH CENTRE and INSTITUTE FOR PROSPECTIVE TECHNOLOGICAL STUDIES (eds.), Technical Report Series*.

Mair, G. (2005), 'Electronic monitoring in England and Wales: evidence-based or not?', *Criminology and Criminal Justice* **5**(3), 257.

Maltoni, D.; Maio, D.; Jain, A. & Prabhakar, S. (2003), *Handbook of Fingerprint Recognition. 2003*, Springer, New York.

Mansfield, T.; Kelly, G.; Chandler, D. & Kane, J. (2001), 'Biometric product testing final report, issue 1.0', *National Physical Laboratory of UK*.

Maragos, P.; Kaiser, J. & Quatieri, T. (1992), 'On separating amplitude from frequency modulations using energy operators', Proc IEEE Int. Conf. ASSP, San Francisco, CA', 1--4.

Martin, A. & Przybocki, M. (2000), 'The NIST 1999 speaker recognition evaluation--an overview', *Digital Signal Processing* **10**(1-3), 1--18.

Martin, A.; Doddington, G.; Kamm, T.; Ordowski, M. & Przybocki, M. (1997), 'The DET curve in assessment of detection task performance', Fifth European Conference on Speech Communication and Technology', Citeseer.

Martin, E. & Law, J. (2009), *A dictionary of law*, Oxford University Press, USA.

Masthoff, H. (1996), 'A report on a voice disguise experiment', *Forensic Linguistics* **3**, 160--167.

Masuko, T. (2002), 'HMM-Based Speech Synthesis and Its Applications.

Masuko, T.; Hitotsumatsu, T.; Tokuda, K. & Kobayashi, T. (1999), 'On the security of HMM-based speaker verification systems against imposture using synthetic speech', Sixth European Conference on Speech Communication and Technology.

Masuko, T.; Tokuda, K. & Kobayashi, T. (2000), 'Imposture using synthetic speech against speaker verification based on spectrum and pitch', Sixth International Conference on Spoken Language Processing', Citeseer.

Matsumoto, T.; Matsumoto, H.; Yamada, K. & Hoshino, S. (2002), 'Impact of artificial gummy fingers on fingerprint systems'(1)'Proceedings of SPIE', Citeseer, 275--289.

McDougall, K. (2007), 'Dynamic features of speech and the characterization of speakers: Toward a new approach using formant frequencies', *International Journal of Speech Language and the Law* **13**(1), 89--126.

Miller, I.; Freund, P. & Johnson, P. (1965), 'Probability and Statistics for Engineers', Prentice Hall, Inc., Englewood Cliffs, NJ.

Milligan, G. & Cooper, M. (1985), 'An examination of procedures for determining the number of clusters in a data set', *Psychometrika* **50**(2), 159--179.

Milner, B. & Shao, X. (2006), 'Clean speech reconstruction from MFCC vectors and fundamental frequency using an integrated front-end', *Speech Communication* **48**(6), 697--715.

Moore, E. (1920), 'On the reciprocal of the general algebraic matrix', *Bull. Amer. Math. Soc* **26**, 394--395.

Moreno, P. & Stern, R. (1994), 'Sources of degradation of speech recognition in the telephonenetwork', 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.

Morrison, G. (2008), 'Forensic speaker recognition using likelihood ratios based on polynomial curves fitted to the formant trajectories of Australian English ai', *Manuscript submitted for publication*.

Müller, S. (2005), 'JMATLINK 1.3', *Last Retrieved 2010-5-30, <http://jmatlink.sourceforge.net/index.php>* .

Murphy, P. (2004), 'The Bayes Net Toolbox for Matlab', Last Retrieved 2010-5-30, <http://code.google.com/p/bnt/> .

Nakadai, K.; Matsuura, D.; Okuno, H. & Tsujino, H. (2004), 'Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots', *Speech Communication* **44**(1-4), 97--112.

Nakadai, K.; Okuno, H. & Kitano, H. (2003), 'Robot recognizes three simultaneous speech by active audition', IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION', Citeseer, 398--405.

Nandakumar, K. (2005), 'Integration of multiple cues in biometric systems', Master's thesis, Michigan State University.

Nandakumar, K.; Chen, Y.; Jain, A. & Dass, S. (2006), 'Quality-based score level fusion in multibiometric systems', ICPR 2006. 18th International Conference on Pattern Recognition.

National Biometric Security Project (2007), '2007 Annual Report on the State of Biometric Standards', Last Retrieved 2010-03-21, <http://www.nationalbiometric.org/docs/2007_annual_report_on_the_state_of_biometric_standards.pdf> .

Nixon, K.; Aimale, V. & Rowe, R. (2008), 'Spoof Detection Schemes', *Handbook of Biometrics*, 403--423.

Nolan, F. & Grigoras, C. (2007), 'A case for formant analysis in forensic speaker identification', *International Journal of Speech Language and the Law* **12**(2), 143.

Nordstrom, T.; Melin, H. & Lindberg, J. (1998), 'A comparative study of speaker verification systems using the Polycast database', Proc. ICSLP'98', Citeseer, 1359--1362.

Olesen, M. (1995), 'A Speech Production Model including the Nasal Cavity', PhD thesis, Aalborg University.

Olsen, T. & Mahler, T. (2007a), 'Risk, responsibility and compliance in Circles of Trust-Part I', *Computer Law & Security Report* **23**(4), 342--351.

Olsen, T. & Mahler, T. (2007b), 'Identity management and data protection law: Risk, responsibility and compliance in Circles of Trust-Part II', *Computer Law & Security Report* **23**(5), 415--426.

Openshaw, J. & Masan, J. (1994), 'On the limitations of cepstral features in noise', 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.

Oppenheim, A.; Schafer, R. & Buck, J. (1989), *Discrete-time signal processing*, Prentice Hall Englewood Cliffs, NJ.

Orphanidou, C.; Moroz, I. & Roberts, S. (2003), 'Voice morphing using the generative topographic mapping', *Proceedings of CCCT '03* **1**, 222--225.

Orphanidou, C.; Moroz, I. & Roberts, S. (2004), 'Wavelet-based voice morphing', *WSEAS Transactions on Systems* **3**(10), 3297--3302.

Ortega-Garcia, J.; Cruz-Llanas, S. & Gonzalez-Rodriguez, J. (1999), 'Facing Severe Channel Variability in Forensic Speaker Verification Conditions', Sixth European Conference on Speech Communication and Technology.

Pantazis, Y.; Stylianou, Y. & Klabbers, E. (2005), 'Discontinuity detection in concatenated speech synthesis based on nonlinear speech analysis', Ninth European Conference on Speech Communication and Technology.

Patrick, A. (2004), 'Usability and acceptability of biometric security systems', *Lecture Notes in Computer Science*, 105--105.

Pelecanos, J.; Navratil, J. & Ramaswamy, G. (2006), 'Addressing channel mismatch through speaker discriminative transforms', IEEE Odyssey 2006: The Speaker and Language Recognition Workshop, 2006', 1--6.

Pellom, B. & Hansen, J. (1999), 'An experimental study of speaker verification sensitivity to computer voice-altered impostors', ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing- Proceedings', 837--840.

Penrose, R. & Todd, J. (2008), 'On best approximate solutions of linear matrix equations'(01)'Mathematical Proceedings of the Cambridge Philosophical Society', Cambridge University Press, 17--19.

Perrot, P.; Preteux, C.; Vasseur, S. & Chollet, G. (2007), 'Detection and Recognition of voice disguise', Proceedings, IAFPA 2007, The College of St Mark & St John, Plymouth, UK.

Phua, K.; Chen, J.; Dat, T. & Shue, L. (2008), 'Heart sound as a biometric', *Pattern Recognition* **41**(3), 906--919.

Plumpe, M. & Meredith, S. (1998a), 'Which is More Important in a Concatenative Text to Speech System-Pitch, Duration, or Spectral Discontinuity?', The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.

Plumpe, M.; Acero, A.; Hon, H. & Huang, X. (1998b), 'HMM-based smoothing for concatenative speech synthesis', Fifth International Conference on Spoken Language Processing.

Porter, D. (2004), 'Identity fraud: the stealth threat to UK plc', *Computer Fraud & Security* **2004**(7), 4--6.

Potamianos, G.; Neti, C.; Luettin, J. & Matthews, I. (2004), 'Audio-visual automatic speech recognition: An overview', *Issues in Visual and Audio-Visual Speech Processing*.

PrimeLife (2009), 'From H1.3.5: Requirements and concepts for identity management throughout life', Last Retrieved 2010-04-20, <http://www.primelife.eu/news/latest-news/28-heartbeat-on-identity-management-throughout-life-published>, Technical report, Privacy and Identity Management in Europe for Life.

Rabiner, L. & Sambur, M. (1975), 'An algorithm for determining the endpoints of isolated utterances', *Bell Syst. Tech. J* **54**(2), 297--315.

Rabiner, L. (1989), 'A tutorial on hidden Markov models and selected applications inspeech recognition', *Proceedings of the IEEE* **77**(2), 257--286.

Raj, B.; Singh, R. & Smaragdis, P. (2005), 'Recognizing speech from simultaneous speakers', Ninth European Conference on Speech Communication and Technology', Citeseer.

Ramachandran, R. & Mammone, R. (1995), *Modern methods of speech processing*, Springer.

Raphael, L.; Borden, G. & Harris, K. (2006), *Speech science primer: Physiology, acoustics, and perception of speech*, Lippincott Williams & Wilkins.

Ravindran, S.; Anderson, D. & Slaney, M. (2006), 'Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing', *Reconstruction* **12**, 14.

Ravindran, S.; Anderson, D. & Slaney, M. (2006), 'Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing', *Reconstruction* **12**, 14.

Reynolds, D. & Rose, R. (1995), 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE transactions on Speech and Audio Processing* **3**(1), 72--83.

Reynolds, D. (2002), 'An overview of automatic speaker recognition technology', IEEE International Conference on Acoustics Speech and Signal Processing', 4072--4075.

Reynolds, D. (2003), 'Channel robust speaker verification via feature mapping', Proc. ICASSP', 53--56.

Reynolds, D.; Quatieri, T. & Dunn, R. (2000), 'Speaker verification using adapted Gaussian mixture models', *Digital signal processing* **10**(1-3), 19--41.

Richardson, F. (2002), 'Electronic tagging of offenders: trials in England', *The Howard Journal of Criminal Justice* **38**(2), 158--172.

Richiardi, J.; Drygajlo, A. & Prodanov, P. (2006), 'Confidence and reliability measures in speaker verification', *Journal of the Franklin Institute* **343**(6), 574--595.

Roach, P. (2004), *Phonetics*, Oxford University Press.

Rogerson, M. & Christmann, K. (2007), 'Burglars and wardrobe monsters. Practical and ethical problems in the reduction of crime fear', *British Journal of Community Justice* **5**(1), 79--94.

Rose, P. (2002), *Forensic speaker identification*, London: Taylor & Francis.

Rose, P. (2006), 'Technical forensic speaker recognition: Evaluation, types and testing of evidence', *Computer Speech & Language* **20**(2-3), 159--191.

Ross, A.; Nandakumar, K. & Jain, A. (2008), 'Introduction to Multibiometrics', *Handbook of Biometrics*, 271--292.

Saastamoinen, J.; Fiedler, Z.; Kinnunen, T. & Franti, P. (2005), 'On factors affecting MFCC-based speaker recognition accuracy', International Conference on Speech and Computer (SPECOM'2005), Patras, Greece', Citeseer, 503--506.

Saslove, H. & Yarmey, A. (1980), 'Long-term auditory memory: Speaker identification.', *Journal of Applied Psychology* **65**(1), 111--116.

Savova, G.; Therneau, T. & Chute, C. (2006), 'Cluster Stopping Rules for Word Sense Discrimination', *Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together*, 9.

Scherer, K.; Johnstone, T.; Klasmeyer, G. & B\'e4nziger, T. (2000), 'Can automatic speaker verification be improved by training the algorithms on emotional speech?', Sixth International Conference on

Spoken Language Processing.

Schiller, N. & Koster, O. (1998), 'The ability of expert witnesses to identify voices: a comparison between trained and untrained listeners', *Forensic Linguistics* **5**, 1--9.

Schmidt, K. (2008), 'Computing the Moore--Penrose inverse of a matrix with a Computer Algebra System', *International Journal of Mathematical Education in Science and Technology* **39**(4), 557--562.

Schuckers, S. (2002), 'Spoofing and anti-spoofing measures', *Information Security technical report* **7**(4), 56--62.

Sedgwick, N. & Limited, C. (2003), 'The Need for Standardization of Multi-Modal Biometric Combination', *Algorithmica Limited, Cambridge, Cambridge*.

Sedgwick, N. (2006), 'Iris Pattern Matching using Score Normalisation Techniques', Last Retrieved 2010-04-27, <http://iris.nist.gov/ICE/CambridgeAlgorithmic_ICE_Brief.pdf>.

Shao, Y. & Wang, D. (2008), 'Robust speaker identification using auditory features and computational auditory scene analysis’, Proc. IEEE ICASSP', 1589--1592.

Shao, Y.; Srinivasan, S. & Wang, D. (2007), 'Incorporating auditory feature uncertainties in robust speaker identification’, Proc. ICASSP', 277--280.

Shriberg, E.; Graciarena, M.; Bratt, H.; Kathol, A.; Kajarekar, S.; Jameel, H.; Richey, C. & Goodman, F. (2008), 'Effects of Vocal Effort and Speaking Style on Text-Independent Speaker Verification', *Proc. of Interspeech, Brisbane, Australia*.

Shuang, Z.; Bakis, R. & Qin, Y. (2006), 'Voice conversion based on mapping formants’, TC-STAR Workshop on Speech-to-Speech Translation. Barcelona, Spain', 219--223.

Skowronski, M. & Harris, J. (2003), 'Improving the filter bank of a classic speech feature extraction algorithm’, IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS.

Slaney, M.; Covell, M. & Lassiter, B. (1996), 'Automatic audio morphing’, 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings.

Srinivasan, S. & Wang, D. (2007), 'Transforming binary uncertainties for robust speech recognition', *IEEE Transactions on Audio Speech and Language Processing* **15**(7), 2130.

Stallings, W. (2007), *Network security essentials: applications and standards*, Prentice Hall.

Stergiou, A.; Pnevmatikakis, A. & Polymenakos, L. (2005), 'Audio/visual person identification’, 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms.

Stevens, K.; Williams, C.; Carbonell, J. & Woods, B. (1968), 'Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material', *The Journal of the Acoustical Society of America* **44**, 1596--1607.

Stevens, S.; Volkmann, J. & Newman, E. (1937), 'A scale for the measurement of the psychological magnitude pitch', *The Journal of the Acoustical Society of America* **8**, 185.

Sturm, J.; Kamperman, H.; Boves, L. & Os, E. (2000), 'Impact of speaking style and speaking task on acoustic models’, Sixth International Conference on Spoken Language Processing.

Stylianou, Y. & Syrdal, A. (2001), 'Perceptual and objective detection of discontinuities in concatenative speech synthesis’, IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING.

Stylianou, Y. (1996), 'Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification', *These, Telcom Paris*.

Sun, X. (2002), 'Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio', *Proc. of ICASSP, Orlando* .

Sundermann, D.; Hoge, H.; Bonafonte, A.; Ney, H.; Black, A. & Narayanan, S. (2006), 'Text-independent voice conversion based on unit selection', 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings.

Tan, Z. & Lindberg, B. (2008), *Automatic speech recognition on mobile devices and over communication networks*, Springer-Verlag New York Inc.

Teunen, R.; Shahshahani, B. & Heck, L. (2000), 'A model-based transformational approach to robust speaker recognition', Sixth International Conference on Spoken Language Processing.

Thomson, J. (1975), 'The right to privacy', *Philosophy & Public Affairs* **4**(4), 295--314.

Togneri, R.; Toh, A. & Nordholm, S. (2006), 'Evaluation and Modification of Cepstral Moment Normalization for Speech Recognition in Additibe Babble Ensemble', Proc. SST', 94--99.

Tokuda, K.; Masuko, T.; Hiroi, J.; Kobayashi, T. & Kitamura, T. (1998), 'A very low bit rate speech coder using HMM-based speech recognition/synthesis techniques', IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING', Citeseer.

Tootell, H.; Alcock, C. & Cooper, J. (2003), 'Consumer Concern and Privacy: A Transition from Pre-Web to Post-Web', , *CollECTeR (Europe), National University of Ireland, Ireland*, 77--82.

Tosi, O.; Oyer, H.; Lashbrook, W.; Pedrey, C.; Nicol, J. & Nash, E. (1972), 'Experiment on voice identification', *The Journal of the Acoustical Society of America* **51**, 2030-2043.

Toth, B. (2005), 'Biometric liveness detection', *Information Security Bulletin* **10**(8).

Trivedi, J.; Maitra, A. & Mitra, S. (2005), 'A Hybrid Approach to Speaker Recognition in Multi-speaker Environment', *Lecture notes in computer science* **3776**, 272.

Ulery, B.; Fellner, W.; Hallinan, P.; Hicklin, A. & Watson, C. (2006), 'Studies of Biometric Fusion Appendix C Evaluation of Selected Biometric Fusion Techniques', *Last Retrieved 2010-04-27, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.5067&rep=rep1&type=pdf>* .

Uludag, U. (2006), 'Secure biometric systems', PhD thesis, Citeseer.

UNISYS (2006), 'UNISYS STUDY: Consumers Worldwide Overwhelmingly Support Biometrics for Identity Verification', *Press Release, Date: Monday, May 1 2006, Last Retrieved 2010-04-30, <http://www.allbusiness.com/crime-law-enforcement-corrections/law-biometrics/10562971-1.html>* .

United Kingdom (1998), *Data Protection Act 1998*, Stationery Office.

United Kingdom (2000), 'Regulation of Investigatory Powers Act 2000', Last Retrieved 2010-06-22, <http://www.opsi.gov.uk/acts/acts2000/ukpga_20000023_en_1>.

United States (1999), 'Uniform Electronic Transaction Act', *Last Retrieved 2010-06-14, <http://www.law.upenn.edu/bll/archives/ulc/fnact99/1990s/ueta99.htm>*.

Valin, J.; Yamamoto, S.; Rouat, J.; Michaud, F.; Nakadai, K. & Okuno, H. (2007), 'Robust recognition of simultaneous speech by a mobile robot', *IEEE Transactions on Robotics* **23**(4), 742--752.

Van Der Ploeg, I. (2005), 'Biometric Identification Technologies: Ethical Implications of the Informatization of the Body', Last Retrieved 2010-04-22, <http://www.biteproject.org/documents/policy_paper_1_july_version.pdf>.

Varga, A. & Moore, R. (1990), 'Hidden Markov model decomposition of speech and noise', proc. ICASSP', 845--848.

Vepa, J.; King, S. & Taylor, P. (2002), 'Objective distance measures for spectral discontinuities in concatenative speech synthesis', Seventh International Conference on Spoken Language Processing', Citeseer.

Walsh, J.; Kim, Y. & Doll, T. (2007), 'Joint iterative multi-speaker identification and source separation using expectation propagation', IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

Warren, S. & Brandeis, L. (1890), 'Right to Privacy', *Harv. L. Rev.* **4**, 193.

Watt, D. (2009), 'The identification of the individual through speech', *Language and Identities, Edinburgh University Press, Edinburgh.*

Webb, A. (1999), *Statistical pattern recognition*, A Hodder Arnold Publication.

Weintraub, M.; Taussig, K.; Hunicke-Smith, K. & Snodgrass, A. (1996), 'Effect of speaking style on LVCSR performance', Proc. ICSLP', Citeseer, 1457--1460.

Westby, J. (2005), 'International guide to privacy', American Bar Association.

Westin, A. (2002), 'The American Public and Biometrics, presented at a conference organized by the National Consortium of Justice and Information Statistics, New York City (5 November 2002).

Whitman, J. (2004), 'The Two Western Cultures of Privacy: Dignity versus Liberty.', *Yale Law Journal* **113**(6), 1151--1223.

Wiener, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*, New York: Wiley.

Williams, A. (2007), 'Veiled Truth: Can the Credibility of Testimony Given by a Niqab-Wearing Witness be Judged without the Assistance of Facial Expressions?', *U. Det. Mercy L. Rev.* **85**, 273.

Wilson, D. (2008), *Forensic Procedures for Boundary and Title Investigation*, Wiley.

Woodward, J. & Center, A. (2001), *Army Biometric Applications: Identifying and Addressing Sociocultural Concerns*, Rand Corporation.

Woodward, J. (2008), 'The Law and the Use of Biometrics', *Handbook of Biometrics*, 357--379.

Woodward, J.; Orlans, N. & Higgins, P. (2003), *Biometrics:[identity assurance in the information age]*, McGraw-Hill/Osborne.

Wouters, J. & Macon, M. (1998), 'A perceptual evaluation of distance measures for concatenative speech synthesis', in Proc. ICSLP.

Wu, Z. & Cao, Z. (2005), 'Improved MFCC-Based Feature for Robust Speaker Identification', *Tsinghua Science & Technology* **10**(2), 158--161.

Xiang, B.; Chaudhari, U.; Navratil, J.; Ramaswamy, G. & Gopinath, R. (2002), 'Short-time Gaussianization for robust speaker verification', IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS SPEECH AND SIGNAL PROCESSING', Citeseer.

Yamagishi, J.; Onishi, K.; Masuko, T. & Kobayashi, T. (2005), 'Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis', *IEICE TRANSACTIONS on Information and Systems* **88**(3), 502--509.

Ye, H. & Young, S. (2003), 'Perceptually weighted linear transformations for voice conversion', Eighth European Conference on Speech Communication and Technology.

Ye, H. & Young, S. (2004), 'Voice conversion for unknown speakers', Eighth International Conference on Spoken Language Processing.

Young, M. & Campbell, R. (1967), 'Effects of context on talker identification', *The Journal of the Acoustical Society of America* **42**, 1250--1254.

Yu, K.; Mason, J. & Oglesby, J. (1995), 'Speaker recognition using hidden Markov models, dynamic timewarping and vector quantisation', *IEE Proceedings-Vision, Image and Signal Processing* **142**(5), 313--318.

Zen, H. & Toda, T. (2005), 'An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005’, Ninth European Conference on Speech Communication and Technology.

List of Appendices

Appendix A: Biometric Principles

A.1 Biometric Systems

This appendix offers a brief overview of concepts related to biometrics and lays foundation for further arguments and discussions offered in the text.

Biometric systems serve the goal of identification by employing distinctive characteristics of a person, called "biometric identifiers". Biometric identifiers, or shortly Biometrics, should be able to-ideally-uniquely prove identity of a person whether in isolation or in combination with other biometrics. Some of the biometrics widely used or under investigation today are fingerprint, voice, palm-print, keystroke, iris image, gait, face, ear, hand geometry and signature (Jain et al., 2008a).

A look over how easily signature is accepted in the daily life as well as in legal disputes, despite that it is neither permanently unique nor impossible to forge reveals that there are numerous factors that affect suitability and public acceptance of a biometric identifier.

"Biometric systems" could be classified according to their characteristics, restrictions and applications as follows (Maltoni et al., 2003):

- Co-operative/ Not: Indicates whether the system's user co-operates with it and tries to be recognised or not (is it of user's benefit to be recognised?).
- Covert or Overt: If users know if they are being identified or not.
- Attended/ Not: Whether system is supervised and attended or unsupervised and non-attended. Systems could be supervised at the time of training (enrollment) be unsupervised at recognition phase.
- Habituated or Non-Habituated: Indicates frequency of use by users.
- Open or Closed: Single application and database or multiple applications and shared access to database.

Classified by application, biometric systems are divided into these classes (Maltoni et al., 2003):

- Used in commercial applications such as network login, ATM, PDA, distance learning

- Used in governmental and large scale security applications as in driver's license, national ID and passport
- Used in forensic applications such as identification of victims, missing children, criminals

A.2 Biometric Identifiers

Biometric identifiers are signatures of an individual; Physiological, chemical and behavioral characteristics (Jain et al. 2008b) that enable us to distinguish a person/case from other candidate people/cases whether automatically (computer recognition) or with the help of an expert. There are a number of requirements that a suitable identifier (ideally) should meet (Clarke 1994, Jain et al. 2008b, Maltoni et al., 2003):

- Uniqueness/Distinctiveness: Jain et al. (2008b) use the term 'sufficiently different across individuals' as the requirement for the biometric identifier.
- Universality: Every individual should have the trait.
- Persistence/Permenance: The characteristic should not change over the time. (More precisely, it should not change so that, we mistake a case for another or be unable to authenticate a case).
- Collectability/Measurability: The biometric data should be easily acquired by suitable devices.
- Being Circumvention Resistant: This refers to the possibility of circumventing verification or identification by means of fake or counterfeit piece of biometric evidence
- Acceptability: The target population should be willing to provide biometric data of their own to the system.

Table A-1 contains a qualitative comparison of all biometrics from (Maltoni et al., 2003) based on their own judgment. The entries are indeed disputable due to qualitative nature of the factors for example Hong in his PHD dissertation presents a similar table with different entries especially in circumvention column (1998).

Table A-1 Characteristics of Biometric Identifiers (Maltoni et al., 2003). [H=High, M=Medium, L=Low]

	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
DNA	H	H	H	L	H	L	L
Ear	M	M	H	M	M	H	M
Face	H	L	M	H	L	H	H
Facial Thermogram	H	H	L	H	M	H	L
Fingerprint	M	H	H	M	H	M	M
Gait	M	L	L	H	L	H	M
Hand-Geometry	M	M	M	H	M	M	M
Hand-Vein	M	M	M	M	M	M	L
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Retina	H	H	M	L	H	L	L
Signature	L	L	L	H	L	H	H
Voice	M	L	L	M	L	H	H

Journal of Biometric Technology Today carried out a survey of new technologies which target identification of human (2002). Among the new biometrics covered in the survey are skin composition and structure and the dermal structure underneath the fingernail known as nail-bed (in addition to gait, smell and ear which are specified up to now). Dental identification based on radiographs as a method especially useful for victim identification (as opposed to suspect identification) has been explored by Chen and Jain (2008).

All the biometric identifiers do not have the permanence and distinctiveness (uniqueness) properties. These identifiers which are labeled “soft biometrics” such as weight, height, gender and age (Jain et al., 2004) can provide ancillary data in parallel with primary biometric identifiers.

Some other biometrics have the potential for monitoring the users constantly and for the duration of a session, for example heart’s sound (Phua et al. 2008) or Electrocardiogram signal (Israel et al. 2005).

The expanding list of biometrics and ignorance about the dimension of surveillance by biometrics lends to the human concerns about intrusion upon their privacy which is covered in the text.

A.3 Biometric Identification and Verification

Biometric recognition could be of either types of verification (one to one comparison) or identification (one to many comparison). Biometric based recognition is undertaken in the phases presented in Figure A-1.

Biometric verification/identification process consists of two phases. The first phase is enrollment or training in which a new case is introduced to the identification system. The system builds a template/model on the input data and saves the relevant parameters in the database. In the first scenario (identification) when a biometric record (from a claimant) is introduced to the system its features will be compared with all available templates/models in the database and the model that receives highest score (highest similarity) will be chosen as the identified case. In the second scenario (verification) the system approves or rejects claimed identity of a case (person), based on the result of comparing its similarity score (with the claimed model) with a pre-defined threshold.

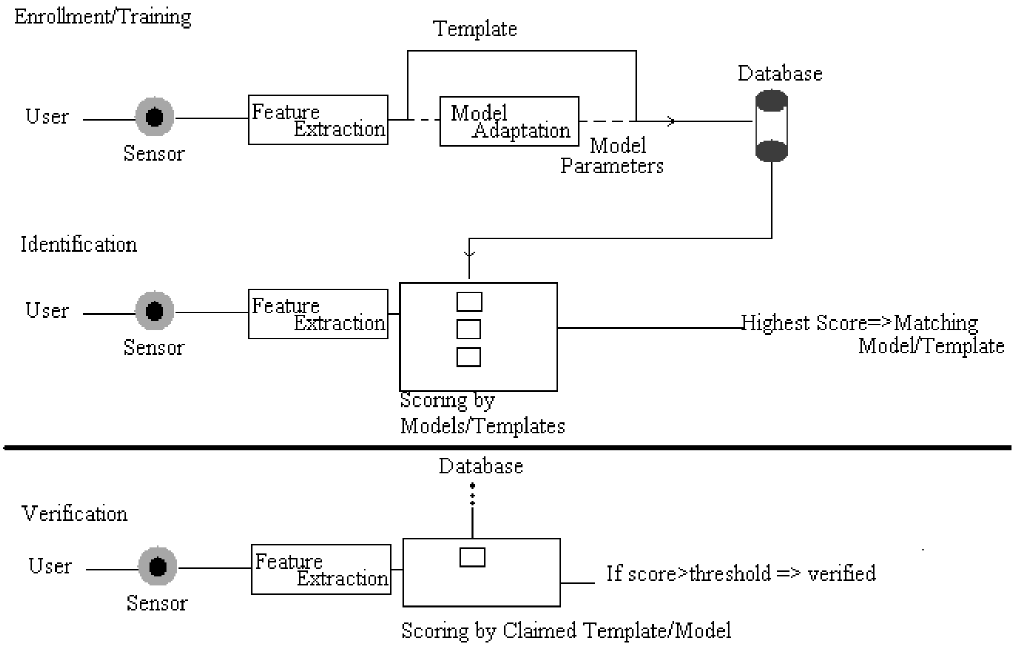


Figure A-1 Overview of a Biometric System and Identification/Verification Phases

A.4 Performance Evaluation and System Errors

Verification systems may fail to reach the right verdict in two ways. First, they may decide that a claimant is not the person that claims when actually he/she is. Second, the system may approve claimed identity of an impostor. The first error is named False Rejection Error or False Non-Match Error and the second one is called False Acceptance Error or False Match Error.

Fig A-2 illustrates the effect of choosing acceptance threshold on these errors. Two PDFs [1] are shown in this figure. The right-hand PDF shows the probability of receiving scores shown on the horizontal axis by genuine claimants. Normally true users receive higher scores when the score is a metric representing the similarity between a biometric record and a model trained over the user's training data. The left-hand PDF shows the probability of receiving the specified score by impostors.

The area shown in dark-grey corresponds to false acceptance error. Likewise the grey area on the left, shows genuine cases which receive lower scores than the pre-defined threshold and consequently are rejected (false rejection error).

It could be easily noted that there is a trade-off between False Rejection Error Rate (FRR) and False Acceptance Error Rate (FAR). Choosing higher acceptance/rejection threshold causes FAR to decrease at the price of increase in FRR and vice-versa.

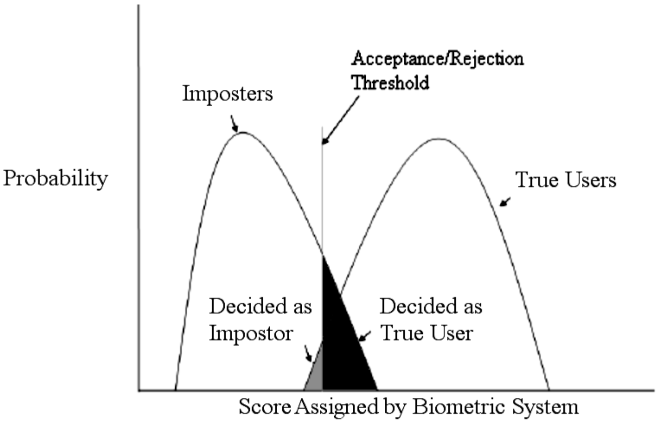


Figure A-2 Distribution of True Users and Impostors and the Effect of Setting Threshold

[1] Probability Density Functions

These two measures are similar to Type I and Type II errors in evaluation of a statistical hypothesis. If the null hypothesis is that the biometric identifier comes from the claiming user, Type I or the error of rejecting the null hypothesis while it is true is the same as false rejection error. Likewise type II error (failing to reject the null hypothesis when it is not true) is the same as false acceptance error.

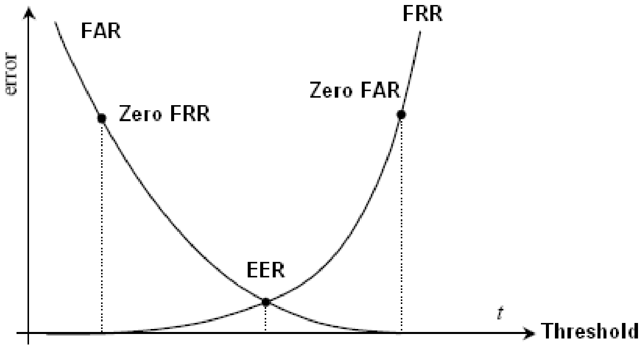


Figure A-3 FAR and FRR Curves Plotted Against Threshold (schematic error rate, arbitrary units)

The Equal Error Rate (EER) is defined as the error rate at the threshold which FAR and FRR are equal. In Figure A-2 at EER point, the threshold is set so that both shaded areas have the same area. Zero FRR is defined as the lowest false acceptance rate at which false rejection rate reaches zero and similarly Zero FAR denotes the lowest false rejection rate that its corresponding FAR is zero.

It is helpful to plot FAR and FRR curves in a diagram as in Fig A-3 which shows both values against threshold. Another standard graph used for comparison between the performance of various identifiers or different algorithms is Detection Error Tradeoff (DET) graph (Martin et al., 1997). Figure A-4 shows DET curves for verification by several identifiers (face, finger-print: chip and optical recognition, hand, iris, vein and voice) reported by Mansfield et al. (2001). Similarly Receiver Operation Curves (ROC) could be used for illustration of error rates. The main difference between ROC and DET curves is that in the latter normal deviate scales are used (Jain et al. 2008b). In ROC curves (1-FRR) (or true rejection rate) could be displayed against FAR.

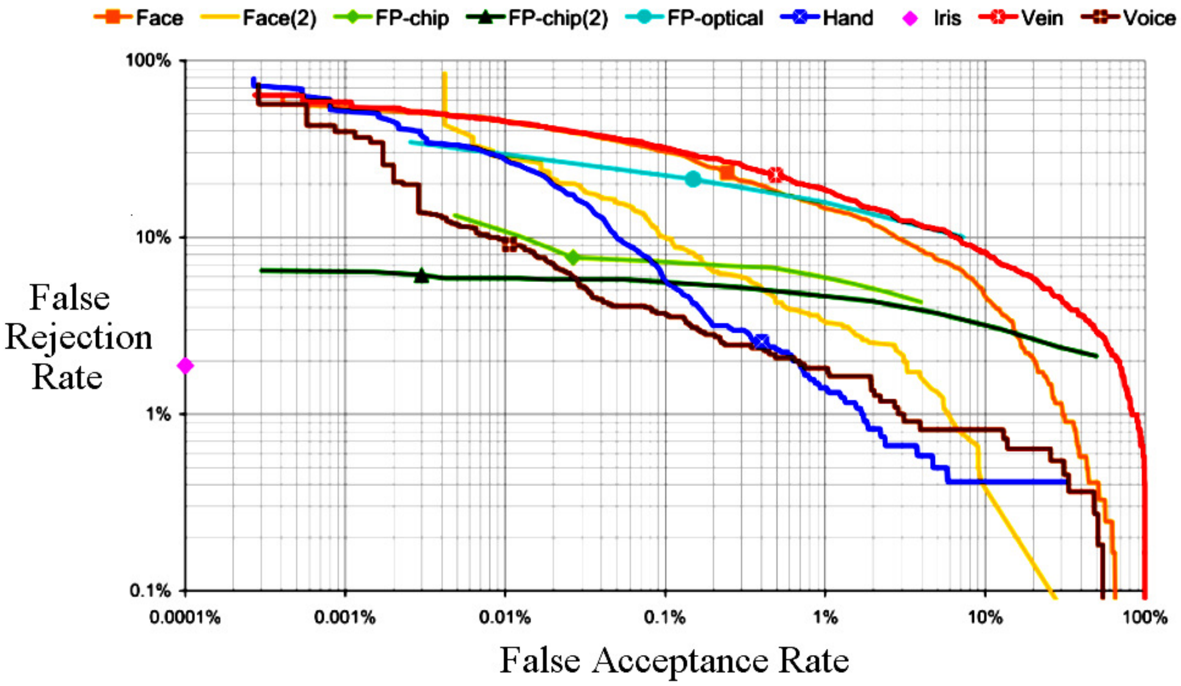


Figure A-4 Detection Error Tradeoff (DET) graph for several biometrics reported by Mansfield et al. (2001)

A.5 Biometric Fusion

Integrating biometric evidence from various sources, e.g. through use of multiple biometric identifiers, is called biometric fusion. Biometric fusion removes some of the technical and non-technical obstacles on the road of large-scale usage of biometrics. These obstacles have been discussed in the literature. For one example, Nandakumar (2005) named noise, non-universality, lack of individuality of the chosen biometric trait, absence of an invariant representation (intra-class variation) and susceptibility to circumvention among the issues ahead of unimodal biometrics. Similar list of limitations is presented by Ross et al. (2008).

One of the ethical reasons for ruling out use of biometrics in many contexts is that a portion of people-due to age and disability reasons-are unable to provide one biometric or another. This problem can be solved by allowing users to choose the means of authentication from a list of biometrics.

Use of multiple cues is not limited to employing two or more identifier. Ross et al. (2008) divided multi biometrics into these categories: multi-sensor (use of multiple sensors e.g. multiple cameras for face recognition), multi-algorithm (which involves integrating scores from various recognition algorithms), multi-instance (use of multiple instances of the same trait e.g. use of left and right index finger), multi-sample (use of multiple samples from the same biometric trait i.e. collecting data several times from the same trait), multi-modal (use of different modalities e.g. voice and face) and hybrid (a subset of all previous scenarios).

Fusion of data, scores or decisions can also occur at various stages, shaping an array of possible levels of fusions. Usually these levels are named: sensor level fusion, feature level fusion, score level fusion and decision level fusion (see e.g. Nandakumar 2005). Ross et al. (2008) contend that early integration strategies are expected to give better results but admit that it is difficult to predict performance without the real test. They also point out that negatively correlated and uncorrelated sources of information produce better results compared to positively correlated data.

The most important technical questions about biometric fusion are simple ones: What is the best type of fusion? What is the best level of fusion and what is the best algorithm for fusion?

The type of multi-biometrics is usually dictated by the type of application and the characteristics of the identifiers. The next two questions are dealt with below.

Ross and Jain (2003) after a review of possible levels of fusion, presented the results of score fusion of three modalities of fingerprint, hand geometry and face using three combination techniques: simple sum rule for score fusion, decision trees and by use of linear discriminant functions. In their experiments the sum rule outperformed the decision tree and linear discriminant classifiers.

It is a thought-provoking yet true proposition that fusion does not always yield better results. Despite the intuition that having more data gives rise to making better decisions Daugman (2000) showed that for decision level fusion the accuracy of decisions as a result of combining two biometrics, is not better than that of the stronger one[1].

Among the score fusion techniques product fusion score (PFS) which is the same method as simplified version of fusion based on Biometric Gain against Impostors (BGI) has shown promising.

Sedgwick (2003, 2006) employed concept of BGI[2] for score fusion in multi-algorithmic verification. In his work he points out that it is the optimal combiner, according to Bayesian

[1] Daugman (2000) showed that for example if the "AND" rule for decisions fusion is being used, the FAR of the weaker biometric must be made smaller than twice the EER of the stronger biometric. Otherwise a strong biometric is better off alone (in terms of total number of wrong decisions) than in combination with a weaker one when both are operating at their EER points.

[2] BGI is the ratio of the a-posteriori to the a-priori probabilities of the claimant being an impostor.

statistical theory (2003)[1] and the BGI concept allows a simple technique for fusion if the biometric measurements are statistically independent. Sedgwick stated that while this (independence) is not the case for multi-algorithmic fusion, experiments show that pattern recognition algorithms based on simple Bayesian fusion are very often highly competitive with more complicated and sophisticated approaches (2006). The equations for BGI-based fusion are presented in chapter 7 in which this method is used for fusion of scores obtained from various algorithms.

Dass et al. (2005) used likelihood ratio statistics for combination of multi-modal biometrics (fingerprint, face and hand) using two methods. In one method they assumed independence of modalities and combined the estimated marginal densities using the product rule (by similar equations to PFS and those presented and used in Chapter 7). In another method assuming the dependence between modalities they used a more complicated joint density estimation (based on Coupla models) for score fusion. Their results showed very similar performance for both methods which was better than the results of verification by the best single modality.

Nandakumar et al. (2006) proposed a likelihood ratio-based fusion scheme with the idea of dynamically assigning weights to the outputs of individual matchers based on the quality of the samples presented at the input of the matchers.

Ulery et al. (2006) in a technical report presented the results of score fusion by several methods including sum of raw scores, sum of normalised scores, linear weighted sum, product of FARs, min/max of FARs, product of likelihood ratios and logistic regression. The last two methods worked better in their experiments. In the explanation of the ‘product of likelihood ratios’ (which was the same as PFS) they stated that the technique requires knowledge of density distributions, for both genuine and impostor scores and presumes that fused scores are independent, “but it is not very sensitive to this assumption” (p. 12).

In chapter 7 simple product rule (or sum rule in log domain) is used as a bench-mark for fusion. A variety of new fusion techniques as well as the well known product fusion score (PFS) (which could be seen as modified BGI-based fusion) will be used and compared.

[1] Griffin (2004) also explains that based on the Neyman-Pearson theorem the optimal fusion is obtained by choosing decision boundaries that match equal density ratio contours. This leads to the same fusion equations offered by Sedgwick (2003, 2006), and also used as product fusion score (PFS) by Dass et al. (2005) and Nandakumar et al. (2006).

A.6 Biometrics and Spoof

The main aim of this section is to introduce spoof as the most significant threat to all biometric systems. The message put across here is that the problem of spoof is not specific to voice verification and that it could affect crucial decisions regarding the design of the system. On a higher level spoof could even question the merit of deployment of a biometric system especially for verification from a distance.

A “counterfeit sample designed to imitate a legitimate biometric submission” is called a ‘spoof’ (IBG, 2006, p. 1). Consequently Spoof attack is a type of attack on biometric system in which fake biometric data is presented to the sensors (Nixon et al. 2008). Possibility of spoof attacks raises very tough questions about suitability of biometrics for their intended purpose. Once lost biometric data is lost forever and it is not possible to revoke it.

The literature depicts the vulnerability of biometrics to spoof attacks, especially when authentication is designed to be from a distance and unsupervised. Those types of applications constitute the majority of cases which call for use of biometrics.

Matsumoto et al. (2002) in one of the first comprehensive studies on possibility of spoof for fingerprint showed that gummies (artificial fingers made out of gelatin) were accepted at a high rate by eleven fingerprint devices with optical or capacitive sensors. Their paper described the process of making artificial fingers in detail. They reported that all of the fingerprint systems accepted the gummy fingers in their verification procedures with the probability of 68-100%.

Schuckers (2002) in a review of spoof and countermeasures for various biometrics especially fingerprint contended that while someone could steal and make a copy of a key this would not discredit the use of keys. Schuckers suggested other means of reducing spoof risks such as supervising the verification/identification process, enrolling several biometric samples (e.g. several fingers), multi-modal biometrics, and liveness detection. For voice verification she cited work by Broun et al. (2002) in which they incorporated lip’s characteristics into the speaker verification task. Broun et al. had used features such as mouth width, upper/lower lip width, lip opening height/width and visibility of the tongue and teeth as visual features (2002) and reported a better recognition as a result of combination of two modalities. Similar researches are available for speaker verification by lip reading (see e.g. Luettin et al. 1996), incorporating visual cues to person identification (e.g. Stergiou et al. 2005) and speech recognition with the assistance of visual aids (see e.g. Potamianos et al. 2004).

Toth (2005) elaborated on types of liveness detection as a countermeasure against spoof in three categories (also presented by Woodward et al. 2003). These categories are using intrinsic properties of a living body, measuring involuntary signals of a living body and gauging bodily responses to external stimuli. Toth concluded that while countermeasures would be available against spoof they affect user's convenience, error rates and hardware prices.

Nixon et al. (2008) made a review of spoof attacks for iris, face and fingerprint. In short iris recognition systems are vulnerable to use of a high quality photograph of the eye. For simple face recognition systems a photograph-or even drawing of the face-can be enough to circumvent security measures. For fingerprint use of spoof material such as Silicone, Clay, Rubber, Soft plastic for simulation of finger's ridges can be a threat to the verification system. A review of previous works on vulnerability of different finger-print systems such as optical, capacitive, thermal and RF imaging systems is made by Nixon et al. (2008).

International Biometric Group (IBG) aimed at creating a spoof library by suggesting a collection of functional spoofs for fingerprint and iris systems and creating a high level test plan (IBG, 2006). It is mentioned in the plan that spoofing research demonstrates the fact that fingerprint and iris systems could be fooled using cheap and easily accessible materials. To fight back spoof attempts 'liveness detection techniques' have been proposed which examine other factors beside the biometric record itself. The plan named some of the these liveness tests including moisture detection, pulse detection and temperature gauging for finger-print and retinal light reflection and calculating the light-absorbing properties of blood, fat, and melanin for iris recognition contending that many liveness detection schemes could be defeated. For example moisture detection algorithms can often be fooled by wetting the spoof material.

It is notable that spoof is not the only category of vulnerability in biometric verification systems for example Nixon et al. (2008) mention other types of attacks on biometric systems, such as replay attack in which the attacker intercepts the communication line between the sensor and the biometric system and puts a genuine biometric record in the processing chain or a Trojan horse attack which replaces the original feature extraction device with a fake extractor which produces desired biometric information. Other categories of vulnerability are not however inherent to the biometric identification, in other words they could be avoided by available security mechanisms such as encryption and signature. Spoof on the other hand displays a flaw of the biometric-based authentication process.

Appendix B: Identity Theft and Fraud

Identity Fraud involves use of fictious or genuine (existent but someone else's) identity details to support and facilitate an unlawful activity[1]. Identity Theft occurs when a person's identity documents or details are fraudulently obtained in the commission of a crime or unlawful activity. Identity fraud generally follows identity theft or invention of identity but in some cases the genuine identity of a person is used while the true person is aware of the fraud and co-operative with the pretenders. Identity fraud can arise from the loss/theft of physical identity documents, from their improper taking from existing official/commercial files, and from simulation of being the person (Jones & Levi, 2000, p. 6).

On the other hand (according to Jones & Levi 2000): "The modern thinking on identity is that two separate equations should be satisfied. The first is to show that the individual actually exists. The second is to show that the applicant is or is not the individual they say they are" (p. 11).

An individual exists if there is a record available from him somewhere in the databases (with the 'attributed identity'). The identity is established if we can attribute one piece of evidence to that record. Biometric identifiers are only helpful in association of our physical body to those records.

While the extent of identity fraud is growing throughout the world it is known as one of the fastest growing criminal trends in the UK (Porter, 2004). One of the reports on the cost of identity fraud to the UK Economy estimates the total loss to be £1.72bn per annum (Feb 2006)[2].

Financial losses are just representatives of the threats posed by ID-fraud to the society. Perception of crime (e.g. identity theft here) is correlated with the quality of life (Christmann and Rogerson, 2004). Identity fraud can damage victims' reputation and claim so much effort and time to re-gain their credibility. Identity fraud acts as an 'enabling agent' or catalyst (Gordon and Norman, 2004) for other categories of crime and facilitates other forms of organised crime[3] such as illegal immigration (human trafficking), money laundering, terrorism and financial frauds.

[1] For detailed description and definitions see: <http://www.identitytheft.org.uk/identity-crime-definitions.asp>. Identity theft website is produced by many private and public sector bodies including UK, Home Office and UK Card Fraud Prevention Service (Last Retrieved: 2010-04-17).

[2] Home Office (UK) (2006), 'Updated Estimate of the Cost of Identity Fraud to the UK Economy', Identity Fraud Steering Committee, *Last Retrieved: 2010-04-17, <www.theirm.org/events/documents/12TwelfthMeeting-Scan003.pdf>.*

[3] Cabinet Office (UK) (2002), 'Identity Fraud, A Study', Last Retrieved: 2010-04-17, <www.statewatch.org/news/2004/may/id-fraud-report.pdf>.

A look over the categories of identity fraud helps us gain a better insight into suitability and effectiveness of biometric verification in combating it.

Figure B-1 is re-produced on the statistics provided by the Identity Fraud Steering Committee report of 2006 and enumerates categories of fraud for several organizations/departments. APACS[1] has secured the first place by a wide margin. The cost to 10 other departments are detailed in that report and are cumulated under the category of 'Others' here. The figures imply that only financial aspects of identity fraud justify devising more reliable forms of identification.

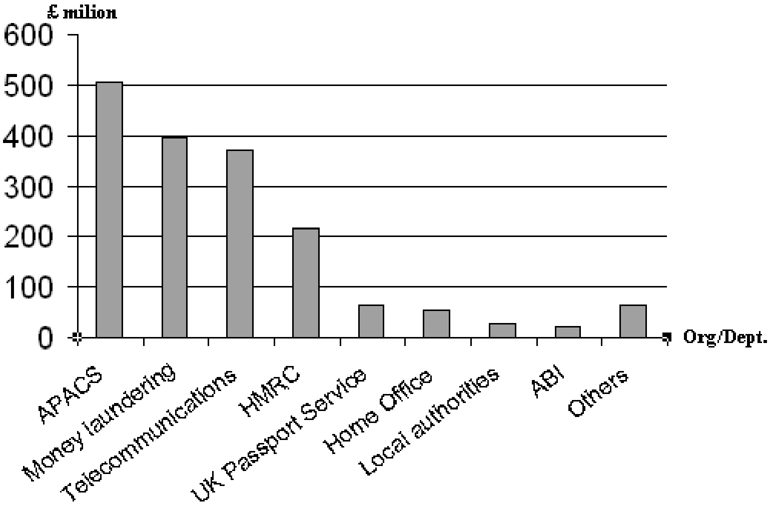


Figure B-1 Cost of ID fraud to the UK economy (the bar-chart is plotted on the figures given in the Identity Fraud Steering Committee report, February 2006)

Breakdown of APACS identity fraud related losses is presented in Figure B-2[2]. Card not present (CNP) transactions[3] have provided highest opportunity for identity fraud. Other types of financial fraud shown in the figure could also be conquered by devising methods which rely on additional factors besides knowledge and possession. APACS reports show that CNP losses continue to grow over the years.

[1] APACS (The UK Payments Association), HMRS (HM Revenue and Customs, Responsible for Direct Taxation, Indirect Taxation, Child Benefit Payment), Home Office (Immigration & Nationality Directorate) and ABI (Association of British Insurers)

[2] APACS (2006), 'Fraud: The Facts 2006', APACS, <*www.apacs.org.uk/*>

[3] CNP transactions are those which are performed in absence of both card-holder and the card such as transactions carried out over the phone, by mail or over the internet.

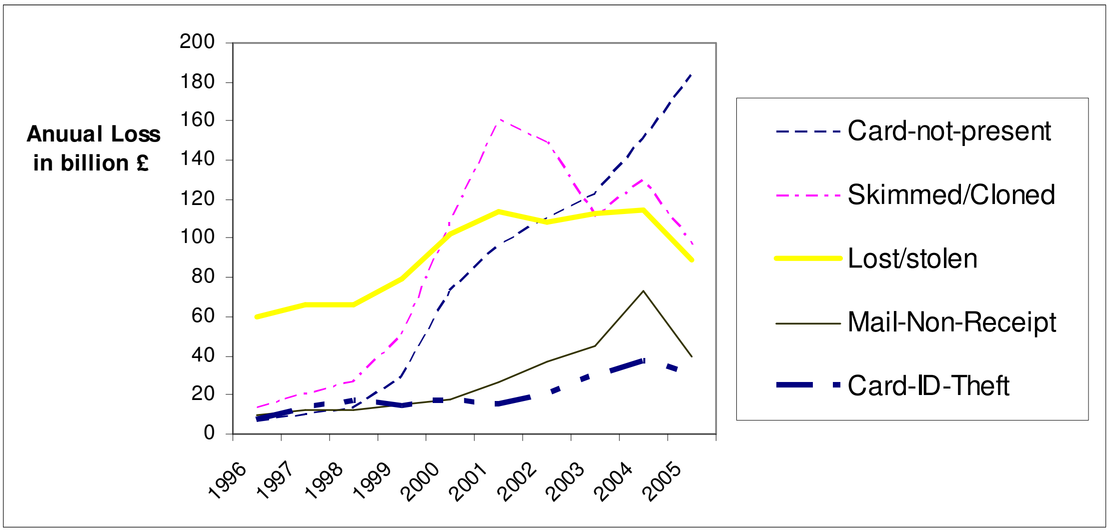


Figure B-2 Details of frauds in payment systems over a course of 9 years (the graph is drawn on the figures presented in APACS report, 2006)

As Jones and Levi (2000, p. 1) point out along with the development of ecommerce our face-to-face working relationship becomes rarer. Therefore we need new means of authentication from a distance for transactions which are not in-person (e.g. CNP transaction)

Appendix C: Non-Technical Dimensions of Biometrics

C.1 A Repertory of Human Concerns Related to Biometrics

This section presents an overview of human concerns related to biometrics.

Woodward and Center (2001) divided key socio-cultural concerns of biometrics into three categories: informational privacy (1), physical privacy (2) and religious objections (3)[1].

They discussed a number of concerns under each category which were: Function Creep[2] (1); Tracking[3] (1); Misuse of data[4] (1); Stigmatization[5] (2); Actual Harm[6] (2); Hygiene[7] (2) and Religious Objections[8] (3).

To add to the last category of concerns and name other religious objections outside Christianity resistance from those Muslim sects which consider covering women's face obligatory could be named[9]. Interestingly while face recognition for surveillance fails if the face is covered an array of biometrics could still be used for authentication purposes in such cases[10]. Even in such rare

[1] The numbers between the parentheses are the category in which they have placed the specified item in.

[2] This Concern arises from possibility of use of biometrics beyond their original purpose without the informed and voluntary consent of the participants which raises the question of whether participants will get the chance to re-assess their participation given the new mission of the system.

[3] Fear of a 'Big Brother' government able to track every individual, surveillance society, clandestine capture of biometric data and harm to the right to anonymity. The concern also arises from the possibility of making a 'complete' profile of the user using partial identities used in different situations, for example for business, leisure, education, etc.

[4] Woodward and Center expatiated on security concerns and possibility of biometric theft under this title.

[5] This topic relates to social stigma that would be associated with a biometric for example fingerprint. They point out that among program managers of voluntary private sector programs involving use of fingerprints no-one cited this stigma as a concern among participants and also foreign (Non-US) biometric programs using fingerprints reported little concern about social stigma among their populations.

[6] They mentioned that even if the biometrics would be harmless in reality the perception of harm may cause users to resist the implementation of biometric measures or be reluctant to participate in them.

[7] This topic is discussed later under health direct/indirect medical implications.

[8] Religious objections (though not widespread) exist among people in different sects of the religions. These objections in the context of Christianity are sometimes related to the excerpts from 'Revelation' referring to a time '...that no man might buy or sell, save that he had the mark, or the name of the beast, or the number of his name… (Revelation, 13:16–18.)'.

[9] BBC: Religions: Hijab refers to covering everything except the hands and face. Niqab is the term used to refer to the piece of cloth which covers the face and women who wear it usually cover their hands also. Although the majority of scholars agree that Hijab is obligatory, only a minority of them say that the Niqab is. (Last Retrieved 2010-04-25, <http://www.bbc.co.uk/religion/religions/islam/beliefs/niqab_1.shtml>). Also cited by Williams (2007).

[10] For other reasons however covering face may be problematic for the exercise of law for example Williams (2007) questions the credibility of testimony given under Niqab due to the concealing of facial expression while mentioning some solutions to the problems.

cases ensuring the person in need of identification that the biometric data is in a form that can not be used for any other purposes than specified and can not be accessed by unauthorised people could allay existing concerns to a great extent.

Woodward and Center were not the only people who compiled a list of concerns. Similar lists have been produced in the past. For example Liu (2007) enumerated several biometric concerns in a list which consisted of items such as function creep, ethical concerns, redundancy of biometrics for the task at hand (being unnecessary), fear of disclosure of sensitive information, concern about facilitating pervasive surveillance, concern about covert collection and involving lower privacy awareness, risk of hacking of central storage, question of creating a safer environment, risk of depriving people of the right to anonymity and risk of permanent ID theft.

Woodward et al. (2003) mentioned causing loss of anonymity, causing loss of autonomy, function creep, cultural related issues such as dignity and stigma and religious objections among the problems ahead of biometrics.

Maghiros et al. (2005) in a comprehensive report[1] paid attention to the possibility of spoof and emphasised that decision makers need to understand the level of security provided by the biometrics[2]. The most important points highlighted in the report that relate to human concerns are summarised:

- Function creep is a concern.
- Human factors such as age, ethnicity, gender, diseases or disabilities (including natural ageing) ought to be studied to minimise possibility of exclusion of a part of the population.
- Biometrics affects the trust model between citizens and the state.
- From economical perspective the identity fraud may become less ‘frequent’ but ‘more dangerous’ subsequent to use of biometrics.
- Fear of surveillance society is another concern.
- As biometric systems are diffused in the society concerns about ‘power accumulation’ and future use of data become important.

[1] The report was written for the LIBE committee of European Parliament in which they assessed the future impact of biometric technologies on society. They urged for new legislation when new applications become widespread in the future (following the governmental use of biometrics at the borders and after ‘diffusion effect’) and when necessary fallback procedures are defined.

[2] It is worth mentioning that the report provides a valuable insight into the biometrics. It makes recommendations in connection with the possibility of function creep, surveillance society and fallback procedures. Areas identified in need of future research in this report are research and technological development, multimodal biometric fusion and large scale field trials.

- Use of biometric evidence must become regulated in courts of law in order to protect suspects adequately.
- Contamination of the biometric sensors and radiation risks for example in case of iris recognition and retinal scanning are among the medical concerns (direct medical implications)
- There is a concern that biometric data might reveal sensitive health information (indirect medical implications). Two examples are use of the science of iridiology to divulge health information about the owner of an iris image and possibility of extracting health information through DNA analysis.

In order to extract other less-directly expressed concerns, we will look at the ethical problems surrounding biometrics, privacy issues and identity management systems as potential sources of concern.

C.2 Ethical Issues Related to Use of Biometrics

In the context of biometric several ethics-related concerns could be enumerated. Woodward et al. (2003) made a mention of the arguments against use of biometrics in which biometric authentication is compared to ‘human branding’ and ‘human tattooing’. Holding this opinion one may argue that biometrics are threats to human dignity and value.

Similar complicated views exist in the literature. To name another one, Van Der Ploeg (2005) proposed a new perception and notion of body as ‘information body’ and machine-readable body. He examined the problems of maintaining the integrity of body based on this notion. In his work he emphasised that integrity is related to inviolability of a person’s body. This mandates consent whenever the boundaries of the body are going to be passed. While skin is considered to be the traditional boundary of the body, with the notion of information body the body boundaries and the notion of body integrity become elusive concepts. Suggestion made by Van Der Ploeg clarifies the implications of this view. The most important suggestions proposed in the abovementioned work among the policy recommendations are:

- Acknowledgement that generation, processing and storing of body data touches upon body integrity
- Ethical and qualitative analysis is needed to assess how different types of body data is related to integrity of the person

- Only in specific cases involuntary collection and processing of body data is ethically justified
- Government do not have right to build databases of (virtual) bodies without consent
- Justification of use of biometrics cannot rely on the notion of individual consent (due to the lack of real choice) but should be based on ethical and human right issues

While the ethical discussions can go on as far as accusing biometrics of transgressing physical or virtual body, disrespecting the human body and harming human dignity it seems that there is not much more complaint about use of biometrics rooted in ethics. Review of several documents on the ethics of biometrics left from BITE (Biometric Identification Technology Ethics) EU funded project[1] shows that relevant non-privacy-related points mentioned in the BITE documents are few and have already been covered so far. In the BITE 2nd project meeting presentation (2005)[2] non technical issues of use of biometrics were divided into privacy issues and accessibility issues. According to the document the accessibility in biometrics has two aspects of age and disabilities. Similarly, as mentioned before, Maghiros et al. (2005) highlighted the need for attending to a portion of the population which may be excluded from the biometric users due to the factors such as age, ethnicity and disability.

In addition to those ethical concerns expressed before, I would like to draw attention to two other ethics-related aspects of use of biometrics:

Firstly, biometrics may give a false sense of security without offering the real security. Rogerson and Christmann (2007) highlighted the problematic ethic of reduction of fear of crime without changing underlying dangers and suggested that measures to reduce fear of crime should only be taken if attempts are being undertaken to reduce relevant risks. That is why I think transferring untrue sense of security, which may be taken on by biometric developers, is unethical.

Secondly, communications and social interactions in a biometric diffused society might be largely affected by the fear of biometric spoofing in the future. As we will see in the case of voice, having a limited amount of voice samples from a person allows impostors to generate counterfeit speech signals. This may be exacerbated where the use of biometrics becomes pervasive (creating the opportunity for misuse of biometric data) and data collected through daily

[1] Which aimed to launch a public debate on bioethics of biometric technology, see <http://www.biteproject.org>

[2] BITE (Biometric Information Technology Ethics) (2006), '2nd scientific Project Meeting', Last Retrieved 2010-04-27,<http://www.biteproject.org/documents/BITE_FINAL_CONFERENCE_PRESENTATIONS.zip>

social interactions could be used for the purpose of spoofing (which causes deterioration in the quality of life and increase in the fear of crime). To draw an analogy, the reader can recall how the email-related threats (such as spam emails, information gathering links, etc.) over the years have affected protocols of communication through email. Users do not open emails received from unknown senders and do not click on the suspicious links. This will happen more severely in case of biometrics contingent upon pervasive use of biometrics and possibility of spoofing.

C.3 Biometrics as a Threat to Privacy

Perhaps the most pronounced concerns about use of biometrics are those which are expressed in the realm of privacy. This section presents a short discussion about privacy and relevance of biometric surveillance / authentication to privacy.

The first step in this direction is defining what exactly the right to privacy is. Unfortunately the definition of privacy is something not easily come by. Warren and Brandeis 13 decades ago in an article on the right to privacy cited Judge Cooley drawing attention to the right "to be let alone" (1890). While several texts start from this definition, the definition seems to provide little information on the boundaries of such a right and its implementation. Woodward (2008) e.g. mentions that despite simplicity and positive appeal of this definition Ellen Alderman criticised this definition by saying that it legally provides no guidance at all. (also in Alderman and Kennedy, 1997).

Tootell et al. (2003) in an examination of the traditional notion of privacy[1] pointed out that many cultures do not have a single word for the concept of ‘privacy’ which implies the complexity of the concept.

One practical approach, then, is attempting to clarify different aspects of privacy.

Woodward (2008) explained that there are three forms of privacy respected under the US law. The Supreme Court has implicitly categorised privacy as taking three distinct forms:

1. Physical Privacy: Fourth amendment of the US constitution clearly states that: “The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable[2] searches and seizures, shall not be violated…”.

[1] as a basis for analysis of consumer concerns from pre-web to post-web

[2] Note how the word ‘unreasonable’ has provided opportunity for conducting ‘reasonable’ searches.

2. Decisional Privacy: Decisional Privacy pertains to making private choices related to marriage, procreation, education etc. This is not likely be affected by biometrics according to Woodward (2008).

3. Information Privacy: involves freedom of the individual to limit access to his/her personal information.

Thomson in a thought-provoking article scrutinised the nature of right to privacy (1975). The proposition set forth in his article is that the violation of right to privacy is actually violation of a person's other rights. Thomson stated that for example, owning a photograph brings you negative rights in respect of it, for example no-one should sell it or tear it and (more controversially) no-one should look at it.

Even if helpful, holding this opinion, raises new questions such as that do we really have the right to forbid visual/auditory access to our belongings including our bodies?

This doesn't seem to be the stance always taken by the legislators. Under the UK's laws, for example, owner of a property cannot prevent photography of it from a public place and, more amazingly, it is only the right not to be harassed that allows a person to prevent a photographer from persistent photography of him/her (see Macpherson, 2009 for photographer's rights in the UK).

Also as will be mentioned in the legal analysis of the use of biometrics, Supreme Court of the US declared in 1973 that requiring person to give voice exemplars is not a search because the physical characteristics of a person unlike the content of a specific conversation are constantly exposed to the public (Woodward, 2008).

Clarke (2006) held the opinion that privacy is an "interest" rather than a "right". He defined the privacy as 'the interest that individuals have in sustaining a personal space, free from interference by other people and organisations'.

Clarke mentioned 4 types of reasons for the importance of privacy:

1. Psychological: that people need private space. "We need to be able to glance around, judge whether the people in the vicinity are a threat".

2. Sociological: Clarke contends that under monitoring "we reduce ourselves to the appalling, un-human, constrained context that was imposed…".

3. Economical: people need to be free to innovate.

4. Political: Clarke mentioned that while people have the need to think, argue and act, surveillance[1] "chills behaviour and speech, and threatens democracy".

Clarke (2001) expatiated on several threats of biometric identification to privacy. The categories of threats elaborated on in Clarke's 2001 article were: threat to the privacy of the person, the privacy of the personal data and personal behaviour; threats due to data-sharing and multi-purpose use; denial of anonymity[2] and pseudonymity; risks of masquerade; permanent identity theft; automated denial of access; undermining democracy and freedom and finally de-humanization (considering biometrics as a method of human branding which harms human dignity).

For the sake of fairness, it should be pointed out that proponents of biometrics defend use of biometric measures in light of their positive impact on privacy. For example Woodward et al. (2003) explain that biometrics on the positive side can protect privacy by safeguarding identity and regulating access to information.

The main question left unanswered here is "what exactly 'right' to privacy implies in the context of biometrics?"

Is it an undeniable right that no-one can deprive a person of? Or is it just an interest among all human interests? Does it hinder collection of biometric data or make its justification extremely difficult?

It is helpful to defer drawing conclusion about privacy issues in the context of biometrics, until a review of legal aspects of use of biometrics is presented. This allows clarifying how the law-makers have looked at the privacy issues as well as other human concerns related to use of biometrics.

C.4 Legal Aspects of Use of Biometrics

This section looks at possibility, limitations and requirements of using biometrics from the perspective of EU, UK and the US laws.

[1] Clarke defines Surveillance as the systematic investigation or monitoring of the actions or communications of one or more persons (2006).

[2] An anonymous record or transaction is one whose data cannot be associated with a particular individual, either from the data itself, or by combining the transaction with other data. A pseudonymous record or transaction is one that cannot, in the normal course of events, be associated with a particular individual. Hence a transaction is pseudonymous in relation to a particular party if the transaction data contains no direct identifier for that party, and can only be related to them if a very specific piece of additional data is associated with it (Clarke, 2006).

Two directives govern data protection in Europe[1]. The first directive 95/46/EC is 'on the protection of individuals with regard to the processing of personal data and on the free movement of such data' and the other relates to 'the processing of personal data and the protection of privacy in the electronic communications sector' (2002/58/EC)[2].

Two concepts of data controller and data processor are defined in the 95/46/EC Directive[3].The key principles of this Directive are (Westby, 2005):

1. Notice: Data subjects (to whom data relates) should be informed of the identity of data collector and the purposes the data is going to be used for.

2. Consent: Article 7 of the Directive declares that "Member States shall provide that personal data may be processed only if the data subject has unambiguously given his consent" however this article specifies 5 other cases in which processing is possible without consent. The exceptions include the one that could be interpreted in a way that justifies use of biometrics in the interest of the public: processing is allowed only if it "…is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller or in a third party to whom the data are disclosed".

Other principles relate to consistency, access, security, onward transfer and enforcement[4].

A look over the charter of fundamental human rights in Europe is also educating as it reveals the positions and stands on the issues of dignity, privacy and data protection. The charter was proclaimed in the European Council meeting in December 2000 (European Parliament, 2000). The data protection article (article 8) of the charter on the protection of personal data states that "…data must be processed fairly for specified purposes and on the basis of the consent …Everyone has the right of access to data which has been collected concerning him or her, and the right to have it rectified….Compliance with these rules shall be subject to control by an

[1] Both accessible from the official website of European Union: <http://europa.eu/>

[2] Directive 2002/58/EC aims at harmonizing data protection standards for communication services (Koenig, et al. 2009) and is not closely related to the biometric data collection and discussions offered here.

[3] 'Controller' shall mean the natural or legal person, public authority, agency or any other body which alone or jointly with others determines the purposes and means of the processing of personal data; 'Processor' shall mean a natural or legal person, public authority, agency or any other body which processes personal data on behalf of the controller.

[4] Consistency: Data could only be used in strict accordance with the specified purpose. Access: Data subjects should have access to their data for modification. Security: Sufficient level of security. Onward Transfer: Personal information should not be transferred to a third party unless consistent with the notice. Enforcement: Member States should provide that any person who has been affected as a result of an unlawful processing operation is entitled to receive compensation (article 23) and the Member States should adopt suitable measures to ensure the full implementation of the provisions of the Directive (article 24).

independent authority". With regard to human integrity the Charter points out that (article 3) everyone has the right to respect for his or her physical and mental integrity and in the fields of "medicine and biology" the free and informed consent of the person concerned, according to the procedures laid down by law must be respected.

In accordance with 95/46/EC directive Data protection Act (DPA) 1998, governs the data protection, the responsibilities of data controllers and the rights of data subjects in the UK (Martin and Law, 2009). The act[1] comprises 8 main principles. DPA stipulates that processing of personal data should be fair and lawful. The data should be obtained only for 'specified purposes' and should not be kept for longer than necessary or processed in any manner incompatible with the specified purpose. DPA mandates taking technical and organisational measures against unauthorised or unlawful processing and keeping data up to date. The schedule 2 of the act specifies the conditions that should be met for the processing of 'sensitive data'[2]. According to DPA the processing of such data needs consent from the data subject if other conditions are not met. It is apparent that biometric data could either be counted as sensitive data or could facilitate obtaining sensitive information about the data owner.

Woodward gives a through account of limitations of use of biometrics according to the US laws (2008). The mentioned article contains many valuable details but its essential points could be summarised as follows:

- US courts have upheld numerous federal, state and municipal requirements mandating fingerprinting for employment and licensing.
- The Supreme Court in 1973 declared that requiring person to give voice exemplars is not a search because the physical characteristics of a person unlike the content of a conversation are exposed to the public. Also the court described Fingerprinting as nothing more than obtaining physical characteristics constantly exposed to the public. A similar example is presented in the case of handwriting.
- While US privacy act of 1974 regulates the collection, maintenance, use and dissemination of personal information by federal government agencies it does not govern non-state or local government agencies. It does not cover the right of non US citizens and does not govern data

[1] *Last Retrieved 2010-04-26, <http://www.opsi.gov.uk/acts/acts1998/ukpga_19980029_en_1>*

[2] Sensitive data defined in the act include the racial or ethnic origin of the data subject, his political opinions, his religious beliefs or other beliefs of a similar nature, whether he is a member of a trade union, his physical or mental health or condition, his sexual life, the commission or alleged commission by him of any offence, or any proceedings for any offence committed or alleged to have been committed.

collection by private sector. The act gives certain rights to “data subjects” and places certain responsibilities on the “data collectors” by restricting federal agencies from disclosure of records, holding them responsible for maintaining records with accuracy, granting individuals rights to access records about themselves and requiring them to establish administrative, technical and policy safeguards.

- There have been examples in the past in which despite the awareness about the impact of large-scale data collection the courts have permitted such actions in the interest of the public. For instance New York state legislature required all prescriptions for a category of drugs to be filed with the name and information about the prescribing physician, pharmacy and the patient. Following a series of actions specified in the Woodward (2008) the Supreme Court while approving the action taken concluded with a cautionary note that the court is not unaware of the threat to privacy implicit in the accumulation of vast amount of personal information which places statutory or regulatory duty on data collectors to avoid unwarranted disclosure.

While the details are of less importance to us, the information presented above and on privacy issues demonstrates that:

1. It is evident that surveillance affects our behaviour and intrudes upon our private space (even if we are in the public space).Consent is a key requirement when it comes to data collection about people. Nevertheless there are many ‘exceptions’ in the laws which justify use of surveillance in the interest of society.

2. Privacy has been treated as an “interest” rather than an “undeniable right” of which a person can not be deprived. Through the democratic processes the public’s interest can outweigh the person’s privacy interests. These democratic processes lead to (and have led to) different outcomes in different societies and cultures.

C.5 Biometrics and Issues Concerning Identity Management Systems

When biometric identification is used for access control most probably the biometric authentication module is linked with a larger scheme for storing, monitoring and managing identity-related information of the individuals in an ‘identity management system (IDM)’. Many factors such as the concept of circle of trust and existent of several identity providers lend to crucial concerns about IDMs. Here after a quick review of opinions and suggestions I propose that while all the recommendations pertaining to manipulation of 'sensitive data' by IDMs should also be applied to biometric data, there are two specific conditions which if met could mitigate the IDM-related concerns.

The main elements of IDMs for the purpose of our analysis are identity providers, service providers and the circle of trust. According to Liberty Alliance[1] models and definitions[2] (the definitions are slightly simplified here): The identity provider is the entity that manages identity information on behalf of Principals (individual users, a group of individuals, a corporation, and other legal entities). A service provider is an entity that provides services and/or goods to Principals. A circle of trust is a federation of service providers and identity providers that have business relationships based on Liberty architecture and operational agreements

Many commercial and open-source identity management systems have a feature called single-sign-on (SSO) which allows one time authentication (through the SSO module in the IDM) and multiple accesses to services and resources within the realm of SSO. Olsen and Mahler (2007a) described that as IDMs seek to solve the problem of multiple authentications (needed for various services) by making digital identities transferable across organisational boundaries they raise the concerns about the level of control users have on the data flow and their privacy.

PrimeLife[3] report on 'requirements and concepts for identity management throughout life' (2009) highlighted technical, legal/political/sociological and societal problems related to privacy protection over a long period of time (for example 40 years). The report examined seven laws of identity suggested by Cameron (2005) under the aspect of an individual's whole life-span[4]. The

[1] The Liberty Alliance (Project) describes itself as an alliance of more than 150 companies, non-profit and government organizations. The consortium is committed to developing an open standard for federated network identity that supports all current and emerging network devices. The portal for Liberty Alliance project can be found here : <http://www.projectliberty.org>

[2] Liberty Alliance Glossary Version 2.

[3] PrimeLife is an EU funded project (started March 1, 2008) which aims at addressing questions around protection of privacy in emerging Internet applications maintaining life-long privacy (http://www.primelife.eu). Its ancestor PRIME aimed to develop a working prototype of a privacy-enhancing Identity Management System (https://www.prime-project.eu/).

[4] The seven laws are worth mentioning here: 1.User control and consent; 2. Limited disclosure for limited use; 3. Justifiable parties (minimality in terms of number of parties dealing with identifying information); 4. Directed identity (The concept expanded on by Cameron is simply that we have partial identities and should have choice to direct those partial subsets toward each party. In public places-in digital domain-a smaller part of the identifiers has to be posed); 5. Pluralism of operators and technologies; 6. Human integration (which involves unambiguous human machine interaction); 7. Consistent experience across contexts; Pseudonyms act as identifiers of subjects or sets of subjects. Whereas anonymity on the one hand and unambiguous identifiability on the other are extreme cases with respect to linkability to subjects, pseudonymity comprises the entire field between and including these extremes.

seven laws and the requirements elaborated on in the report, are applicable to biometrics, especially when biometrics are considered ‘sensitive information’.

Apart from technical discussions and architectural models Olsen and Mahler (2007a) showed that the key concerns about use of identity management systems are privacy and data protection risks: “Users may be interested in knowing which of the collaborators can access their personal information how the flow of information is controlled and whether collaborators can collect personal information to create user profiles which span across multiple domains” (Olsen and Mahler, 2007a, p. 349). Olsen and Mahler (2007b) tried to address IDM related privacy and collaboration issues from the perspective of European data protection law. They contended that with the emerging notion of web services the mapping of entities in the context of IDM to two roles of controller and processor is problematic. They suggested that users should be able to choose between only a few and very easily understandable privacy preferences/policies since there is low awareness about privacy and data protection issues which is reflected in the survey results.

The concern about accumulation of partial identities which may lead to construction of user’s full profile is salient in the context of IDM. Clauß and Kohntopp (2001) described that the identity of an individual comprises of a huge amount of personal data which they called each subset of such data “partial identity”[1]. They mentioned some disadvantages of the current IDMs which included lack of sufficient privacy and security mechanisms, lack of sufficient user control (requirement of trusting the identity providers) and lack of universal and standardised approach to identity management.

Going over the IDM related articles cited here and many others in the literature demonstrates that there is a huge fear about misuse of IDMs. While the concerns are valid it is arguable that biometrics does not add to those concerns if used only for authentication and biometric data is not transferred in form of authentication tokens or any form to the service providers.

C.6 User Expectations and Concerns Reflected in Biometric Surveys

Furnell and Evangelatos (2007) at Plymouth University conducted perhaps one of the two most reliable surveys on perception, awareness and acceptance of biometrics. Based on their results voice after iris and fingerprint was the third biometric identifier that respondents were aware of.

[1] Each person uses different partial identities in different situations for example for work, leisure and shopping.

Their survey showed that except for iris and retina direct health concerns were negligible especially for voice recognition. Biometrics had received high perceived usefulness for applications such as verification of the identity of passport holders, airport check-ins and verification of the identity of credit card holders while this rate was very low for applications such as keeping track of employees work. Concern about stealth use of biometric data was high and only 4% were not concerned at all while 56% were very or extremely concerned.

One of the most important and controversial questions about use of biometrics is that whether people prefer governmental or private sector use of biometrics. Based on the survey the respondents seemed inclined to mistrust the two types of authority (private/governmental) almost equally. As for perceived reliability of voice as a biometric only 21% of respondents thought that voice is reliable or extremely reliable and only keystroke and signature received lower perception of reliability. 72% thought that fingerprint is either reliable or extremely reliable. The majority (61%) selected biometrics as their first preference, as opposed to 31% selecting secret knowledge and 10% for tokens. While they admitted that the title of survey would have caused attaining such a high percentage they cited UNISYS (2006) results in which it was shown that 66% of consumers favoured biometrics over other means of authentication.

In UNISYS survey (2006) Voice recognition was the most favoured authentication method, cited by 32 percent of respondents, followed by fingerprints (27 percent) and facial scan (20 percent).

Another large-scale biometric survey was carried out by BITE (2006) with correspondents from all over the world[1]. While the survey report is very long, and it has aimed to assess the severity of most of the concerns introduced before, the lessons learnt from the survey could be summarised as follows:

- Most people think that many ethical issues should be addressed in public as for use of remote and covert biometric recognition.
- Among civil applications, financial sector / home banking and health sector received higher likelihood of attracting biometric recognition.
- The risk of function creep was rated high by the correspondents both in civil and criminal context.
- Most respondents thought biometric data is sensitive data.

[1] One problem with the survey is that it should be considered ‘experts survey’ as 63% of correspondents defined themselves as expert in one aspect of biometrics.

- Most respondents thought that the distinction between data storage in a personal medium and a central storage is relevant for privacy.
- Most respondents thought that the branding argument (dignity-related) is not relevant.
- The majority of people thought that sometimes or always people have right to anonymity.
- Respondents thought that the risk of stigmatization of disabled people is average.

User concerns are indirectly reflected in some other works in the literature for example Patrick (2004) stated that based on the observations users assumed that a complete image of the biometric is saved, and this led to heightened concern about misuse of data and data aggregation. Patrick also cited Coventry (2004) which pointed out that the users reported significant fears that criminals might injure them in order to obtain parts of their bodies e.g. by cutting their fingers.

It is worth mentioning that support for use of biometrics is not independent of the propaganda and perception of the risks. Westin (2002) reported that the support for biometrics (while still high) had dropped from 2001 (height of terrorist events) to 2002 based on the conducted surveys (in the US). The majority of people-in Westin's reported survey of 2002-supported some crime-related use of biometrics. The approval was also high for verification of credit cards. Westin contended that two key drivers of support for biometrics are: fighting terrorism and concerns about identity fraud.

Appendix D: Speech Processing Features and Models

D.1 Source-Filter Model, Formants and Fundamental Frequency

When we hear someone's voice the information is conveyed by longitudinal waves produced by our vocal organs from the lungs to the mouth. The landmarks of vocal tract are pharynx, oral cavity, velum, tongue, and lips (Raphael et al. 2006). The air is pushed out by the lungs. During the production of voiced sounds the gap between left and right vocal cords (folds), glottis, frequently becomes narrower and wider which produces glottal pulses at the frequency known as fundamental frequency. The rest of the human vocal tract could be seen as a filter which its frequency response combined with the frequency components of glottal pulse shapes the output or sound wave. The human vocal system (or the filter in this model) is usually modeled as an acoustic tube with different number of sections (Holmes & Holmes, 2001). Each of the sections has a resonant frequency and could be modeled as a band-pass filter with pass-band around those specific frequencies.

Source-Filter model of speech production is the basis for the majority of feature extraction techniques. Based on this model, all-pole filters are used in the speech analysis using Linear Predictive Coding (LPC) (covered shortly) and for modeling voiced sounds. More complicated models are proposed for example for modeling the vocal tract and the nasal cavities together with zero and poles in the transfer function (Olesen, 1995) but were not commonly used. The source-filter model is compatible with the idea of calculation of formant frequencies as used in forensic speech processing.

To illustrate the concept of source-filter model, results of a simulation aimed at production of a vowel is displayed in Figure D-1. The source signal plotted in the upper-left subplot is provided by Bunnel (2000) based on the model of vocal flow by Fant et al. (1985)[1]. The signal represents the air pressure passing the vocal folds. The frequency components of the signal (by Fast Fourier Transform, FFT[2]) are calculated and displayed in subplot on the upper-right subplot. The fundamental frequency of the source signal (which is the frequency of vibration of vocal folds)

[1] For the source signal the wav-file on the Bunnel's website is used.

[2] See Oppenheim et al. (1989)

can be calculated from the frequency response (the constant difference between the peaks of the frequency response).

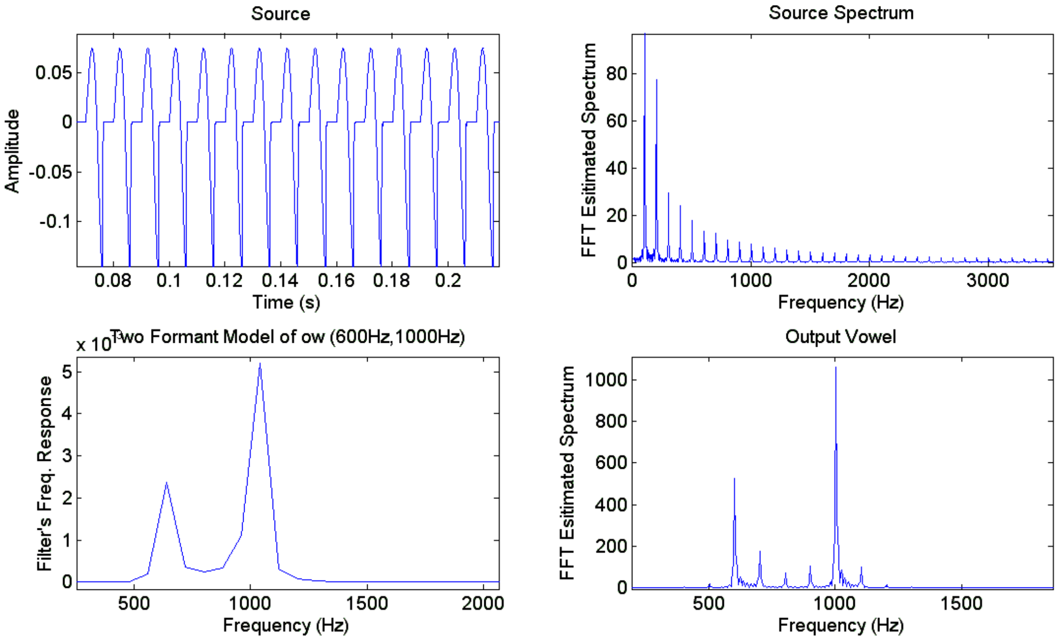


Figure D-1 Illustration of the Results of a Vowel Production Simulation

To produce vowel əʊ (shown also as ow, as in go [g ow / əʊ]) two filters[1] were used in series with pass-bands around 600Hz and 1000Hz (See Raphael et al. 2006 for position of two first formants per vowel). The frequency components of produced vowels are shown in lower right subplot. This is similar to what someone may arrive at by FFT analysis of such a vowel. From the figure, one may be able to calculate both fundamental frequencies and the formants of the vowel (resonance frequencies). The simulation above also demonstrates how easy it is to implement a formant synthesiser for misleading a voice verification system. Therefore if the verification system is only based on measurement of formants, it will be vulnerable to simple spoofing techniques.

[1] Butterworth filters of order 2

D.2 Acoustic Features of Speech

This section makes a short survey of acoustic features of speech with the emphasis on formant calculation and Mel-frequency cepstral coefficients (MFCC) which are extensively used in the literature of speech processing as well as this research[1].

Common feature extraction techniques for speech processing-both for speech and speaker recognition-are mainly based on analysis of spectral components of speech. The signal is transformed from time domain to frequency domain by a discrete transformation. Two common methods are LPC analysis and the fast Fourier transform (FFT).

FFT coefficients are representatives of frequency components for a short frame of speech and are calculated by short term Fourier transform (STFT) (see Oppenheim et al., 1989).

LPC coefficients can be calculated using Levinson-Durbin recursion algorithm (see Furui, 2001 for a detailed discussion). In LPC analysis an all pole model of the signal is adopted and it is assumed that the Z-transform of the signal (a frame of speech) can be written as:

Equation D-1

$$H(z) = \frac{1}{1 + a_1 Z^{-1} + a_2 Z^{-2} + \ldots + a_P Z^{-P}}$$

or in time domain:

Equation D-2

$$\hat{x}(n) = -a_1 x(n-1) + a_2 x(n-2) + \ldots + a_P x(n-P)$$

where $\hat{x}(n)$ is the estimate of signal $x(n)$ and LPC analysis aims at minimizing the difference between these two values.

Parameter P is the order of LPC analysis and elements $[1, a_1, a_2, \ldots, a_P]$ are LPC coefficients which are estimated through the analysis. Adoption of the all-pole model is in agreement with the source-filter model of speech production presented hitherto which assumes that the vocal tract can be modeled as a series of joint tubes with different number of sections.

By calculating the poles of equation above, we can estimate the formants for the frame of speech under investigation (Formants are the peaks in the envelope of the signal's spectrum).

The formant frequency corresponding to pole *P* will be:

[1] The explanations are just enough to allow a smooth transition to the technical discussion in the thesis text. Similar to most topics covered for literature review very densely written texts are required to cover the details about all the techniques mentioned here.

Equation D-3

$$F = a\tan(\frac{imag(P)}{real(P)})/\pi * f_s/2$$

where F is the formant frequency associated with pole P and f_s is the sampling frequency and *atan(.)* denotes the arc-tangent function.

Since the LPC coefficients are real the poles are in complex conjugate pairs therefore with LPC analysis of order P we can estimate $P/2$ formants, for example if we need to estimate 3 formants for a segment of speech, LPC analysis of order 6 has to be carried out[1]. Figure D-2 shows the P/2 poles on the upper half of unit circle. Each pole represents one formant frequency from 0 to $f_s/2$ (for angle of π radian).

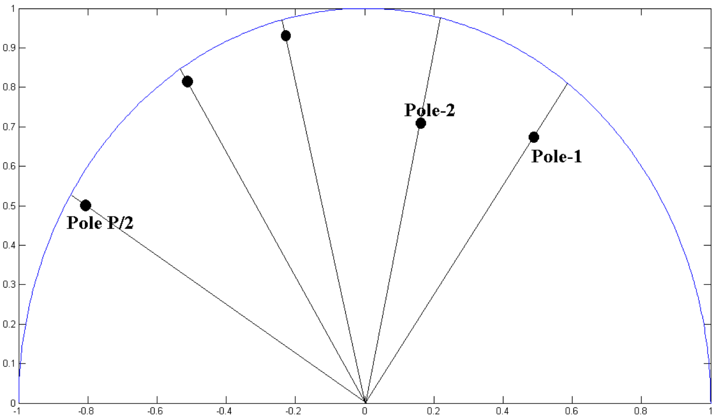


Figure D-2 P/2 Poles Used for LPC-based Calculation of Formants

In figure D-3 results of such an analysis on a short segment of a speech signal is displayed. The middle part of the word ‘meal’ consisting of vowel ‘i:’ is manually picked out from a sentence uttered by a speaker in IVIE corpus (described in chapter 5). The upper left figure shows the signal in the time domain.

The upper right plot shows the spectrum or the frequency components estimated by FFT coefficients. The energy of FFT coefficients can be used for building a spectrogram (lower-right subplot). A spectrogram is a 3-dimensional plot which displays the spectral evolution of a signal. Spectrograms show frequency and time on y and x axes and energy by different colors. The term voice-print was nothing more than representation of speech by spectrograms (Bonastre et al., 2003).

[1] This is because we need the poles on the upper half of unit circle which are lower than Nyquist frequency or half of sampling freuqncy. See Kinnuen (2004) for a detailed explanation.

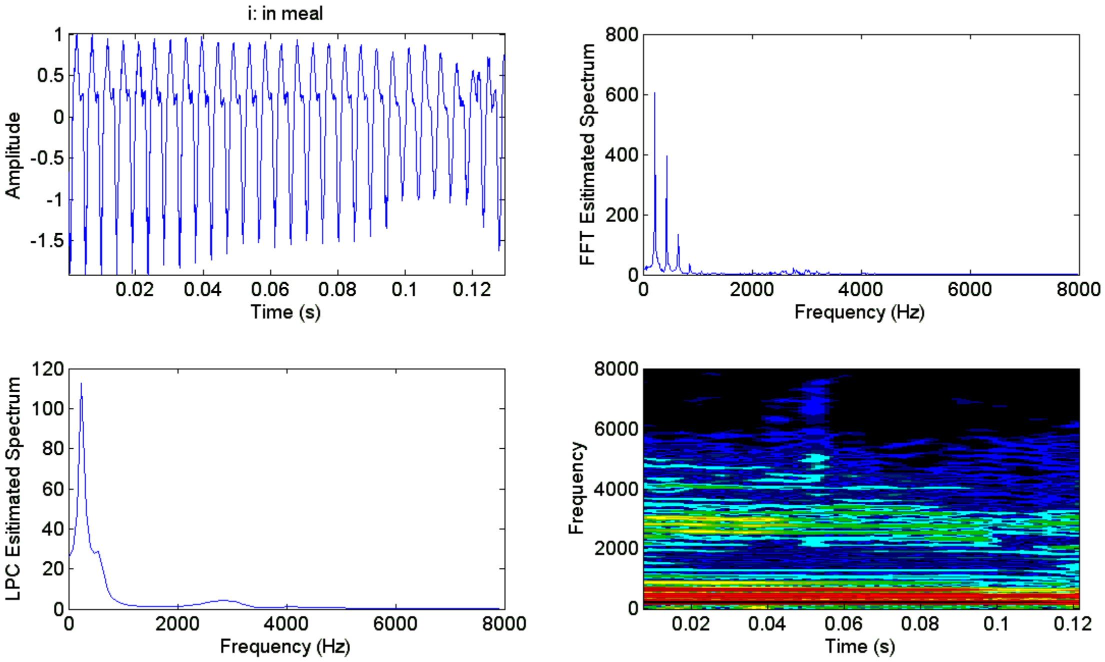


Figure D-3 Illustration of FFT and LPC Estimated Spectrum of i: in 'meal'

Formant frequencies can be calculated by experts from the spectrograms or automatically by LPC analysis as demonstrated above (also in Becker et al., 2008 and Kinnunen, 2004).

Kinnunen in his PhLic thesis has analysed and compared many features of text-independent speaker recognition based on experiments on Helsinki and TIMIT database (2004). The study seems to be the most comprehensive analysis in this area and the features explored include Line spectral frequencies (LSF), Formants, Linear Predictive Cepstral Coefficients (LPCC), Log area ratios (LAR)/Arc sine coefficients (ARCSIN) of Reflection coefficients of Levinson Durbin algorithm. The modeling technique used is vector quantization (as opposed to GMM modeling explained shortly). The results reported in normal conditions are generally good and the errors are similar and close to zero. The error rates for formants as features are slightly higher than raw LPC coefficients and both higher than cepstral coefficients.

Mel-frequency cepstral coefficients (MFCC) emerged as a result of decades of research in speech processing and were initially used for speech recognition. Many ideas such as introduction of Mel scaled filter-bank with respect to works on auditory perception (Stevens et al. 1937) and use of log-cepstrum for separation of source and filter components (see Holmes & Holmes, 2001) are incorporated into MFCC extraction technique.

The procedure of extracting MFCC feature normally consists of the following steps:

1. The signal is split into short frames (for example 20ms) usually with an overlap of around 50% (Holmes & Holmes, 2001).

2. A Window e.g. Hamming or Hanning window (Oppehheim, 1989) is applied to each frame.

3. The FFT of the frame and the energy of the coefficients are calculated.

4. A filter-bank is applied to the FFT energy coefficients. If the center frequencies of the filters in the filter-bank are distributed according to Mel-scales the coefficients will be Mel-cepstral coefficients.

The relation between Mel-frequencies and linear frequencies based on Fant's expression (1973) could be written as (Skowronski and Harris, 2003):

Equation D-4

$$f_{mel} = 2595\log_{10}(1+\frac{f}{700})$$

If the centre frequencies of the filters are uniformly distributed in the mel-frequency space they are denser in lower linear-frequency space and less dense in the higher frequencies. For the implementation of filter-banks VOICE-BOX toolbox as described in chapter 5 was used (Brookes, 1997) in this work. Figure IV displays the distribution of 29 filters in a filter bank in the middle-left subplot.

5. Logarithm of the filters' outputs is calculated. Since we can write:

Equation D-5

$$\log(|X.Y|) = \log(|X|.|Y|) = \log(|X|) + \log(|Y|)$$

If the signal is the result of multiplication of two elements with different frequency ranges, by using log operator we can separate these two signals and filter out one component. We are interested in finding coefficients which can represent the 'filter' or vocal tract (and not the excitation) and based on the source-filter model these slowly varying parameters (Holmes & Holmes, 2001) can be obtained by a low-pass filter.

6. Finally the discrete cosine transform (DCT) of the log values are calculated. There are two reasons behind use of DCT coefficients: de-correlation of coefficients and separation of source and filter. DCT transform is a good approximation of Karhunen-Loeve (see Britanak et al. 2007) transform for de-correlation (Tan and Lindberg, 2008). This is especially important when diagonal GMM models (covered shortly) are used for modeling the speaker's space which

hypothesise that non-diagonal elements of covariance matrix are zero or in other words there is no correlation between dimensions of feature space.

On the other hand DCT acts as a spectral feature and by keeping only first coefficients of DCT transform we will achieve the goal of source-filter separation described above.

Despite several excellent ideas behind MFCC feature extraction, these features are vulnerable to disturbance by noises. This topic is extensively studied in chapter 7.

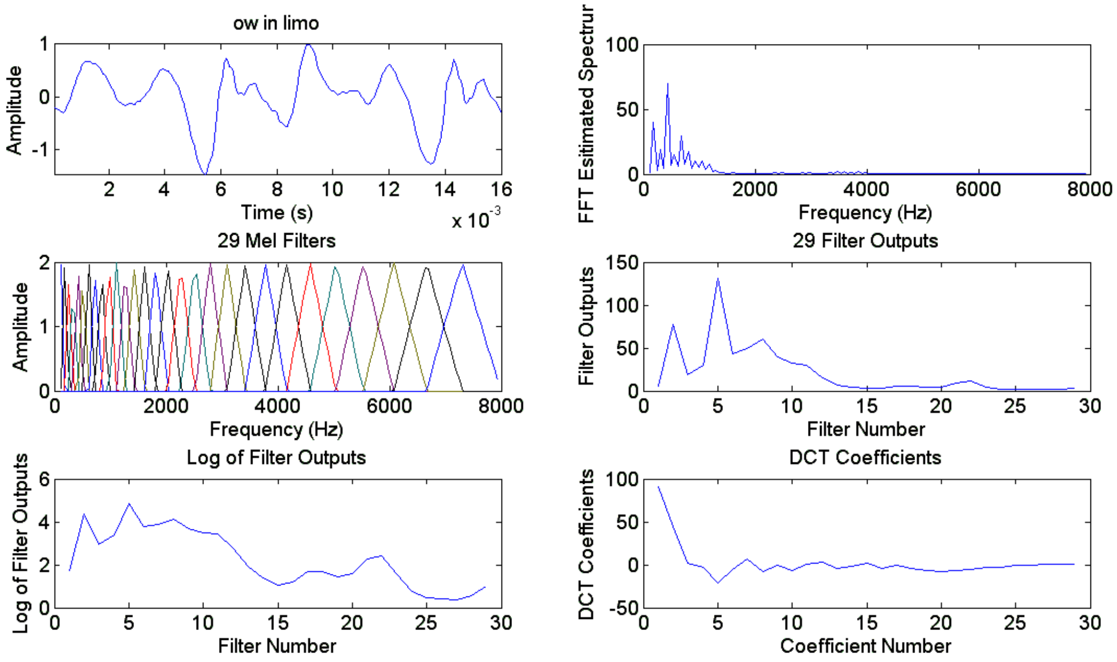


Figure D-4 Demonstration of Steps in MFCC Calculation

In Figure D-4 all the steps described above are depicted. A frame consisting of utterance of vowel ow [≅Y] in the word 'Limo' is shown in upper-left plot. The spectrum of the signal is shown on the upper right plot. The filter-bank (with 29 Mel-filters) and the outputs of filterbank (for 29 filters) are shown in the middle subplots. Note how the outputs of the filterbank are similar to the spectrum of the signal with more emphasis on the lower frequencies. The bottom subplots display the logarithms of filter-output and the DCT of these values.

D.3 Hidden Markov Models

Real-world processes including speech production generate observable outputs that are captured in form of signals (Rabiner, 1989). The signals are semi-raw representations of the process. Based on the importance of different aspects of the signal, we devise different feature extraction techniques. The term ‘observations’ could be, then, used for the sequence of ‘feature vectors’ extracted from raw signal.

Hidden Markov Modeling is a probabilistic pattern matching technique in which the observations are considered to be generated by a stochastic model that consists of an underlying Markov chain. It is suitable for modeling the processes whose probabilistic characteristics are variable over the time (labeled as non-stationary processes). Speech is one of those non-stationary processes. Two important concepts in HMM modeling are states and transition. States correspond to ‘events’ in speech. Vector quantization (see Ramachandran and Mammone, 1995) or GMMs (introduced later) are normally used for modeling these events. Transition modeling serves the purpose of finding the place (in time) that one unit of speech comes to an end and another unit starts.

A hidden Markov model is characterised by the following elements (A detailed description of hidden Markov modeling could be found in Abdulla and Kasabov (1999), Rabiner (1989), Durbin et al. (1998) and Jurafsky & Martin (2000)):

- Number of States (*N*): based on the modeling decisions, states may correspond to units of speech for example a phoneme or smaller units which are parts of the phoneme. States are shown by S_j where *j* is from 1 to *N*.

- Probability of observing a feature vector or observation at time *t* (o_t) assuming that we are in one ‘state’ (*j*)[1]:

Equation D-6

$$P\left(o_t \mid S_j\right), j = 1,...,N$$

- Probability of starting from each state (π_i, *i*=1 to *N*).

- Transition probability from one state to another at any time which is represented by a matrix (*A*) whose elements are $a_{i\,j}$ where *i* is the current state and *j* is the next state. If $Q^t = j$ represents being in state *j* at time *t*, and o_{t+1} denotes event of seeing observation o_{t+1} at time *t,* the transition probability is bound to satisfy this equation:

Equation D-7

$$P\left(Q^{t+1} = j \mid Q^t = i\,, o_{t+1}\right) = P\left(o_{t+1} \mid S_j\right) \times a_{i\,j}$$

[1] This could be also phrased as “probability of emission/generation of feature vector o_t by state *j*”.

A left to right HMM is an HMM in which the transition from one state is only allowed to itself or to the next state. This type of HMM is suitable for modeling characteristics of phonemes, when the start of phoneme corresponds to the first state, the middle portion of phoneme corresponds to the second state and the final part of phoneme is handled by the third state. We can not obviously pronounce the middle part of a phone after the final part.

Assuming a sequence of observations of length *T* (representing an entity such as a word or phone) and an HMM (λ) we should be able to deal with the following problems (Abdulla and Kasabov, 1999):

1. Evaluation/Recognition: Having a known HMM model (λ) we need to calculate the probability of generation of a series of observations (o_t, *t*=1 to *T*) by the model:

Equation D-8

$$P(O \mid \lambda) = P(o_1,.., o_T \mid \lambda)$$

If for example several phoneme models are in the candidate list, this probability helps us find out the best model which represents the observation sequence.

2. Decoding: Finding the most probable state sequence (Q^t, *t*=1 to *T*) for a given observation sequence. Viterbi algorithm is used for this task.

3. Training: Optimizing the model parameters to obtain the best model that represent the set of observations. Baum-Welch algorithm is used for this task.

D.4 Gaussian Mixture Models

Multivariate Gaussian mixture models are used to model distribution of a set of observations. If model parameters are denoted by Π, the observation at time *t* by o_t, number of Gaussians in the mixture by *M*, and the *i*-th Gaussians's weight in the mixture by w_i, the probability of a sequence of observations (from 1 to *T*) given the model will be:

Equation D-9

$$P(O \mid \Pi) = P(o_1,...,o_T \mid \Pi) = \prod_{t=1}^{T} P(o_t \mid \Pi)$$

in which:

Equation D-10

$$P(o_t \mid \Pi) = \sum_{i=1}^{M} w_i . N(o_t, \mu_j, \Sigma_j)$$

where:

Equation D-11

$$N(o_t, \mu_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_j|}} \exp(-0.5(o_t - \mu_j)' \Sigma_j^{-1}(o_t - \mu_j))$$

and μ_j, Σ_j are mean and covariance matrices of the *j*-th Gaussian in the mixture. Training of GMM models involves adjustment of μ_j, Σ_j, w_j for all Gaussians which is performed through expectation maximization (EM) algorithm. The details could be found in Reynolds (1995).

D.5 Score Normalization in Speaker Verification

GMM models produce a probability or a score which is $P(O \mid \Pi)$. For the verification task we need to compare a score with an acceptance threshold. This score is normally the likelihood ratio:

Equation D-12

$$LR(O) = \frac{P(O \mid H_1)}{P(O \mid H_2)}$$

H_1 is the hypothesis that the suspect has produced the observation or speech (*O*) and H_2 the hypothesis that someone else in the population has made the speech, then log-likelihood could be computed thorough following equation:

Equation D-13

$$s(O) = \log LR(O) = \log P(O \mid \Pi) - \log P(O \mid \Pi^M)$$

where Π is the GMM model for the speaker and Π^M is a universal/background/world model which is trained on speech from an impostors population and *s* is the score assigned to the observation *O*. *s* could be compared with a set threshold.

The above type of score normalization is usually called world model normalization. Instead of a world model a number of cohort models could be used for normalization. This type of normalization and Zero and Test Normalization (Z-Norm and T-Norm) are described in

Auckenthaler et al. (2000) in greater detail that could be covered here. In short, Through Z-Norm the score (*s*) is normalised by a mean (μ_z) and variance (δ_z):

Equation D-14

$$s_z = \frac{s - \mu_z}{\delta_z}$$

μ_z and δ_z are mean and variance of the scores obtained by the speaker model, through scoring a number of impostors' utterances at the training time.

In T-Norm, at the time of verification the speech (which is going to be verified) is scored by a number of impostor models, and the mean and variance of the scores obtained are used for normalization.

Appendix E: Voice-print and Forensic Speaker Recognition

E.1 Life of Voice-print

According to Eriksson (2005) the earliest use of the term ‘voice-print’ goes back to 1944 in Grey and Copp’s Bell Lab report (1944) where the patterns in spectrogram were analysed. Subsequent to this historical event and naming, which implied finding unique characteristics of speakers’ voice (in analogy to fingerprint), in 1960s, a two-year research by Kresta at Bell lab[1] (Lindh, 2004) drew attention to use of voice for identification again.

In 1962 Kresta reported a subjective matching accuracy of more than 97% based on the analysis of spectrograms (Kresta, 1962). Kresta recorded voice (isolated words and in context) from 25 speakers (15 male and 10 female). He showed that when 7 experts and when a panel of female students under 18 years of age (after five days of voiceprint reading experience) decided on the cases successful decisions exceeded 97% of the cases.

Later on, Young and Campbell (1967) conducted similar experiments with the words collected in different contexts. They reported two error rates one for the training task (78.4% success) where the words were uttered in isolation and the other for the experimental task (37.3% success) for different contexts (Eriksson, 2005). They concluded that in different contexts identification results were different due to reasons which outweigh “intra-talker consistency”.

Stevens et al. (1968) carried out speaker identification based on aural and spectrographic methods and reported a 6 percent error for aural presentation and about 21% for visual presentation for a closed set consisting of 8 speakers. They mentioned that the results “depend upon the talker, the subject, and the phonetic content and duration of the speech material.” (p. 1596, abstract)

Bolt et al. (1970) analysed the problems associated with spectrographic speaker identification in greater detail. They mentioned the ‘intermixing’ of characteristics of phrases and speakers and that the human observer has to decide subjectively on similarities. The similarity might indicate that ‘similar sounds’ have been spoken but not that the ‘same person’ has produced both. They

[1] when the New York Police started to receive reports of bomb threats to different airlines

also compared fingerprint to voiceprint and pointed out that the fingerprint is the direct representation of anatomical attributes of the person but voice-print is not a representative of vocal anatomy. They mentioned different identification error rates in different conditions and need for establishing the reliability of voice verification under practical conditions.

Three years later, Bolt et al. (1973) analysed the new experimental results especially those carried out by Tosi et al. during 1971 and 1972 at Michigan state university and mentioned many reasons for variation in the errors which required more caution interpreting the results and further investigation. Among the factors they mentioned were whether closed or open sets were used for identification, the age of the voices (how contemporary they were), population size, difference in words' context, emotions, noises, room acoustics, recording conditions and possibility of disguise. In the response to Bolt's letter published in 1973, Black et al. (1973) criticised some of Bolt's conclusions and interpretations and defended the methods used for voice identification.

It is worth mentioning that the opinions on voice-print identification by spectrographic methods could be as harsh as Eriksson's who expressed that "voice-printing is still done by private detectives and other non-academic experts but nobody in the speech science community believes in its usefulness for forensic purposes any more." (2005, p. 5).

E.2 Contemporary Forensic Speaker Recognition

Formants play a key role in forensic speaker recognition.

Nolan and Grigoras (2007) proposed that the formants are "acoustic signatures" of the speaker, therefore the formants frequencies and dynamics are central to speaker identity and must be used for speaker identification. They also argued that the formants are less susceptible to contextually induced variation and to distortion in transmission than parameters arising from laryngeal activity (such as f0 and global spectral slope).They presented the results of two case-studies based on two different acoustic analyses, one focusing on short-term, segmental events and the other capturing long-term trends.

McDougall (2007) tried to draw attention to the dynamic of formants instead of their static characteristics. With this aim she used regression to parameterise formant frequency contours. Her remark in the start of the paper is also interesting and worth mentioning here that "A phonetician is able to measure and quantify features of speech whose values may exhibit differences from one speaker to the next, but there is no known set of features or criteria which can be used to characterise the speech of an individual as exclusive to that person" (p. 89).

Some studies may be very specific and focused on a particular part of speech for example Morrison (2008) analysed the possibility of providing a likelihood ratio based on the coefficients of polynomial curves fitted to the formant trajectories of Australian English 'ai'.

There are also published standards for spectrographic speaker recognition.

As an example American board of recorded evidence (ABRE) has published a standard for comparison of recordings on the aural and spectrographic basis (1999)[1]. The standard explains that aural/spectrographic examination can produce one of seven decisions: Identification, Probable Identification, Possible Identification, Inconclusive, Possible Elimination, Probable Elimination, or Elimination. According to the standard the samples should meet some requirements e.g. there must be at least 10 comparable words between two 2 voice samples. The standard also explains requirements for channel, noise, quality, disguise and other factors two pieces of speech should have for comparison. The descriptions are qualitative and it is hard to determine whether samples satisfy the requirements. The characteristics the standard specifies for spectrographic analysis includes formants, pitch, energy distribution and word length.

E.3 Reliability of Extracted Parameters for Forensic Analysis

Several research results show that while spectrograms are affected by the context of the phrases there are other factors such as age, channel and disguise which may have influence on the parameters used for forensic speaker recognition.

Endres et al. (1971) based on the spectrogram of seven speakers over a period of 29 years demonstrated that the formant frequencies and pitch shift to lower values as age increases. They also showed that disguised voice exhibits different formant structure from the normal voice.

As mentioned in chapter 2 Kunzel (2001) studied the effect of mobile and telephony transmission on the measurement of formants. Kunzel used two methods for calculating formants, one by visual inspection of spectrogram and the other by use of auto-correlation function. They observed that the F1 was higher in the telephone-transmitted data compared to the data recorded directly. For F2 their results showed that F2 centres could be either higher or lower and did not follow a consistent pattern therefore despite the distortions a regular pattern could

[1] American Board of Recorded Evidence -- Voice Comparison Standards, 1999 , Last Retrieved 2010-05-09, <http://www.forensictapeanalysisinc.com/Articles/voice_comp.htm>

not be established. The maximum amount of average distortion (among 10 speakers of male or female) for F1 was 13.6% (for vowel ‘i:’) and for F2 2.2% (for vowel ‘o:’).

Byrne and Foulkes (2004) carried out similar experiments on English vowels uttered on mobile phones which showed comparable trends but in their experiments F1 increased by an average of 29%. Apart from filter characteristics they ascribed this effect to the fact that mobile phones are smaller than landline receivers and their distance to the mouth is greater. They reported that the majority of F2 measures were lower in the mobile recordings and that in general F2 measurements were less affected by transmission (because they fell within the transmission bandwidth of the phone). As for the third formant they observed that F3 values had a small downward shift for mobile phone and significant shifts were only existent for the highest values (which again could be explained by the pass-band of the phone).

Among recent explorations Guillemin and Watson studied the effect of mobile codecs (GSM-AMR) alone on the measurement of formants and observed that in general formant frequencies were decreased by the codec, particularly in the case of high-frequency formants (2006).

A few studies reviewed above show that reliability of spectrographic recognition by formants is under question when channel effect is significant.

Appendix F: Biometric Cards and Security Concepts

F.1 A Quick Review of Security Concepts

In this section three main concepts used in cryptography are explained using a simplified version of notation adopted by Katz and Lindell (2008):

1. Encryption

In asymmetric encryption we need three functions (*Gen, Enc* and *Dec).* The key generation algorithm (*Gen*) outputs a pair of public key and private key (*pk, sk*), the encryption algorithm (*Enc*) for a message *m* takes public key and returns *c=Enc(pk,m)* and the decryption algorithm (*Dec*) returns *m=Dec(sk,c).*

For symmetric algorithms *pk, sk* are the same.

2. Signature

In signature generation we need three functions (*Gen, Sign* and *Vrfy).* The key generation algorithm (*Gen*) outputs a pair of public key and private key (*pk, sk*), the signature algorithm (*Sign*) for a message *m* takes private key and returns signature *s=Sign(sk,m)* and the verification algorithm (*Vrfy*) takes public key and the signature and returns *Vrfy(pk,m,s)* which is 1 if *s=Sign(sk,m) or 0* if the signature is not correct.

3. Certificate

Digital certificate is a signature that is generated with the aim of binding some entity to some public key.

Suppose that entity *C* want to give a certificate to entity *B*. *C* can sign a message such as "I have signed *B*'s key which is (*pkB*)" using its own private key (*skC*) and hand it to *B*. (*pkB)* is the value of the public key of entity *B*. Now when entity *B* signs a message using its private key (*skB)* and sends it to a party along with its certificate, the party (having the public key of *C* and trusting it), can verify that the certificate is made by *C*. Then it can find the public key of B from the certificate message (*pkB*) and using it verify the signature made by *B*.

Using the above concepts the security of communication could be guaranteed. Several network protocols such as secure socket layer (SSL) (see Stallings, 2007) are proposed and used for maintaining the integrity of messages and for authentication of two parties in need of communications.

F.2 Use of Biometrics on Smart Cards

This section aims at showing that biometric data could easily be placed on the card with chips (processors) known as smart cards with little or no security risks (extra to those related to biometric verification itself). Security concepts already developed for cards and presented in card-standards allow realization of secure biometric authentication.

In a very simple model, the interaction during a transaction by card consists of two sides: the 'card' and a machine called 'device' here[1].

The de-facto card standard Europay, Master and Visa (EMV) which is published by EMVCo[2] describes requirements of a secure communication between card and device and can facilitate use of biometric authentication similar to PIN checking on the card.

In the process outlined in EMV documents (EMV, 2000) following two processes in which the card authenticates the device using the same security concepts and certificates described before, and the device authenticates the card and makes sure that card information is not changed after being issued, the device should check that the genuine card-holder is in possession of the card. Normally this is done by checking a PIN number.

According to EMV standards, the PIN does not exit the card but it is checked by card[3]. In card-holder verification, user enters PIN and this entered number is sent to the card in a standard command defined for smart cards. The application on the card checks the PIN and returns a response indicating whether the PIN is correct or not. If the user fails to enter the correct PIN for a number of times, the application blocks itself.

The same process could be recommended for biometric data. The biometric models must never leave the card, and instead of PIN the machine should deliver the features of biometric sample read by the device to the card. The card can use a biometric algorithm to check if the biometric sample matches the biometric template/model stored on the card or not.

[1] The device can be online and be connected to an authority at the transaction time or authorize the transaction in offline mode for a limited number of times.

[2] EMV Co. was formed in February 1999. EMV standards aim for interoperability and acceptance of payment system integrated circuit cards on a worldwide basis (EMVCo.com).

[3] See 'Integrated Circuit Card, Specification for Payment Systems, Book 3, Application Specification' (EMV, 2000) for cardholder verification methods.

Appendix G: Speaker Discriminant Spectral Areas

G.1 Rationale

The Mel-filter-bank and its position and width of filters are optimised for speech recognition and may or may not be optimum for speaker recognition.

There are two pieces of work which have independently examined the possibility of this optimization. In the first one, Kinnunen (2004) used two datasets (Helsinki and part of TIMIT, DR7) with the sample rate of 11025 Hz, and the F-Ratio method to search for such spectral areas. His results showed that apart from the low frequencies a speaker discerning spectral area exists between 2 and 3 kHz. Total bandwidth in his work however due to Nyquist theory was limited to around 5 kHz. In a more recent work, and without any reference to Kinnnen’s research, Lu and Dang (2007) employed theoretical discussions based on modeling of vocal tract and formants, as well as statistical approaches based on F-Ratio and Mutual Information to show that speaker specific spectral areas are in the ranges of 4-5 kHz, 7-8 kHz and low frequencies. The idea was further developed by design of cepstral features with higher number of filter-banks in those areas. The results reported on NTT-VR speaker recognition database for identification by GMM models showed some improvement over the traditional MFCC and uniformly distributed filter-banks.

It will be shown that F-Ratio itself may be a misleading marker of discrimination power when the data is not concentrated around one point (as in speech with different phonemes). A new metric which could be named discrimination power is suggested here with the same functionality as mutual information but with some advantages which could be used not only for this problem but for similar ones. Finally the most spectral areas are identified based on two corpora CHAIN and IVIE and the eight different sets of filter-banks are designed and tested on these corpora (on the normal sentences and retold passages).

G.2 Markers Used for Evaluation of Discrimination Power

Three types of marker are introduced in this section, two of which are borrowed from the previous works: F-Ratio, Mutual Information and Discrimination Power.

1. F-Ratio and Fisher's ANOVA (Analysis of Variance)

F-Ratio is the ratio of between-class to within-class variances among two or more classes of data:

Equation G-1

$$F - Ratio = \frac{\sum_{k=1}^{K} \frac{N_k}{N} (u - u_k)^2}{\sum_{k=1}^{K} \frac{N_k}{N} \sum_{i=1}^{N_k} (x_{ik} - u_k)^2 / N_k}$$

K denotes the number of classes. N is the number of points ($1/K$ is kept in the numerator and denominator of the fraction for illustration purposes). N_k represents the number of points in class k which weights each class according to the number of its points. F-Ratio is based on the mean of data, and when data comprises a number of subclasses F-Ratio does not portray the real discrimination power. F-Ratio is an indicator of discrimination possibility if only one Gaussian has to be used for classification per class therefore it is not the best pointer for our study. The example-in Figure G-1 shows that though the distributions of two classes are quite separate (and they could be discriminated for example by a mixture models with two Gaussians) the F-Ratio is zero since the means of two classes are the same.

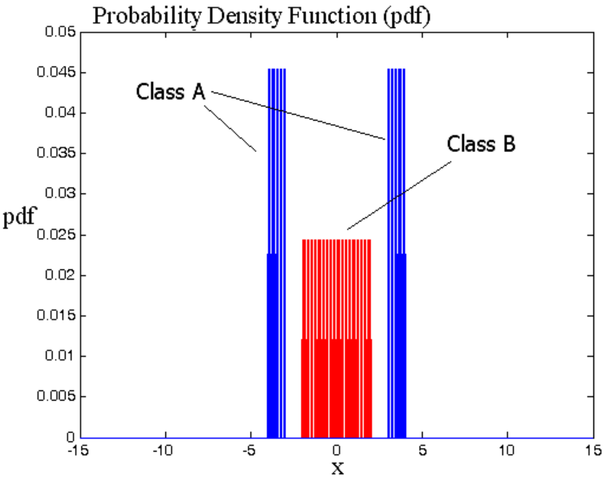


Figure G-1 PDF of two classes with the same mean

2. Mutual Information

The mutual information between two random variables is a quantity that measures the mutual dependency between those two variables. In other words it shows that, knowing one of the variables, how much allows us to guess the other one.

Mutual information is defined as:

Equation G-2

$$I(D;Y) = H(D) + H(Y) - H(D,Y) = H(D) - H(D \mid Y) = H(Y) - H(Y \mid D)$$

Equation G-3

$$H(Y) = -\sum_{x \in X} p(y)\log_2 p(y)$$

where $H(D)$ and $H(Y)$ are the information entropies (Shannon entropy) and $H(D,Y)$ is the joint entropies of D and Y. $H(D)$ specifies the uncertainty of D (data) as a random variable or the minimum number of bits needed to code it by the most ideal coding algorithm. If Y indicates the class number, $H(D|Y)$ will specify the uncertainty if the class number is known. If there is huge overlap among the data from various classes, knowing D does not give so much information about Y therefore the uncertainty remains high ($H(Y|D)$ will be high) and the mutual information will be low. In contrast, if the data of the classes are entirely separated by class number knowing D is enough to be able to tell Y, therefore $H(Y|D)$ will be low and the mutual information will be high[1].

Lu and Dang employed mutual information as an indicator of how much knowing about the data in each of the spectral areas can help us identify the speaker the data belongs to. One problem with the calculation of mutual information is that it requires working out two dimensional probabilities based on the equations[2]. The other, is that the unit and the absolute value of mutual information in bits are not very expressive. F-Ratio was used by Liu and Dang (2007), and Kinnuen (2004) for discrimination power analysis of spectral components.

3. Discrimination Power

A new discrimination score is proposed here which, for a two-class-discrimination task, measures the ratio of the number of correct classification decisions over the total number of possible decisions, where each decision is defined as assigning a sample to a class (based on the probability density function of the classes). For more than two classes the method uses an

[1] For example if the distributions of two classes are uniform with no overlap, seeing one observation, we can certainly tell that which class the observation belongs to. If there is an overlap between these two uniform distributions, for the data in common it is not possible to tell the class knowing the data, so there will be some amount of uncertainty in that region which increases $H(Y|D)$ and lowers mutual information between data and class number.

[2] However if one dimension is the class number and discreet in nature care could be taken for reduction of computational complexity.

average of the probability density function of all the rival classes. Assume that *$f_j(i,x)$* indicates the probability density function (PDF) of feature *j*, for class *i* at point *x*. In text-independent speaker recognition, *i* represents speaker index and values of *$f_j(i,x)$* could be estimated from long recordings of speech. Feature *j* could be *j*-th coefficient of feature vector or output of *j*-th spectral filter in the filterbank.

For *K* classes (*K* speakers) we can define *discrimination power* of feature *j* as:

Equation G-4

$$Disc-Power_j = \frac{1}{K}\int_{-\infty}^{\infty}(-\frac{1}{K-1}\sum_{k=1}^{K} f_j(k,x) + (1+\frac{1}{K-1}).\max_k f_j(k,x))dx =$$

$$\frac{1}{K(K-1)}\int_{-\infty}^{\infty}(-\sum_{k=1}^{K} f_j(k,x) + K.\max_k f_j(k,x))dx$$

(Since the sigma on the left side of the first integral has the max PDF in itself with negative sign, $\frac{1}{K-1}\max_k f_j(k,x)$ is explicitly added in the second expression.)

It could be noted from the definition that if all *$f_j(i,x)$* values for all speakers *i*, are the same the disc-power for feature *j* will be 0. In contrast if *$f_j(i,x)$* values do not have any overlap, the integral will add up to 1. Since the integrated expression in the bracket is positive over all values of x, the value of disc-power always lies between these two extremes (0 and 1).

While this metric best suites the problem at hand, in any problem which involves deciding the class of an observation based on the previously estimated distribution functions the equation can be helpful.

The integrated expression represents the discrimination score associated with the most rational choice made at interval *dx*. The discrimination score is, in fact, the difference between the number of correct and incorrect choices in the interval in discrete mode. If we think of it as a selection scenario between circles and triangles the scenario can be illustrated as in Figure G-2. The Intervals displayed here are similar to *dx* intervals in the suggested expression. The discrimination score (or power) within interval-1 is 3, which reflects the fact that we will incur 1 error and will make 4 correct decisions if we label the interval most rationally ('belonging to triangles!'). The integral sums these values over all the tiny intervals across *x*-axis. For more than two classes the average of all rival PDFs is used as the representative of the rival class.

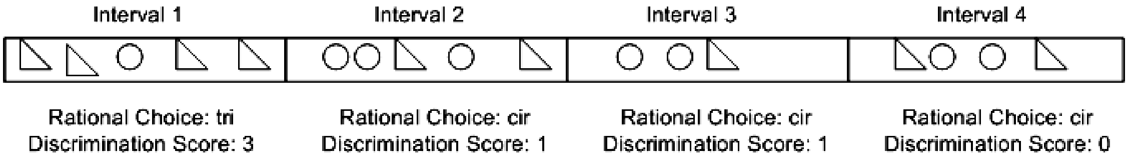


Figure G-2 Illustration of discrimination scores for three hypothetical intervals

G.3 Experimental Results

The results of spectral analysis on several subsets of CHAIN and IVIE are reported in this section. 61 filter-banks uniformly distributed across the entire spectrum (0- 8 kHz) were used for calculation of features. The normalised outputs of the filter-banks are considered as the features which their discrimination power is to be evaluated. Both mutual information and discrimination power measures were calculated. The probability distribution functions were estimated based on histograms with the width of bins adjusted based on the variance of data for each feature. No smoothing was applied on the histograms for estimation of distribution functions (equivalent to rectangular smoothing).

Three sets of results are reported here: experiments on Subset C of IVIE, Solo Subset of CHAIN, Combination of Solo and Fast recordings of CHAIN (Figure G-3).

The results show that:

1. Two measures are consistent. They offer the same spectral areas as highly discriminative. These areas are low frequencies, around 4 kHz, 5 kHz and between 6-7.5 kHz.

2. Despite the differences in the databases such as differences in accents and recording devices the results from three sets reported here and all other sets examined were close.

3. Combining Fast and Solo sentences in a set enables us to bring in some variation to the experiments and figure out whether the discerning spectral components remain the same when intersession variability and different rates of speaking are taken into account. The results are shown on the bottom subplots of Figure G-3.

4. The amount of real impact of choosing higher number of filters (packed filters) in the spectral areas with higher discrimination power is unknown considering that the variation in the discrimination power from maximum to minimum is less that one percent (use of a tangible unit is one of the advantages of the proposed measure for analysis of discrimination power).

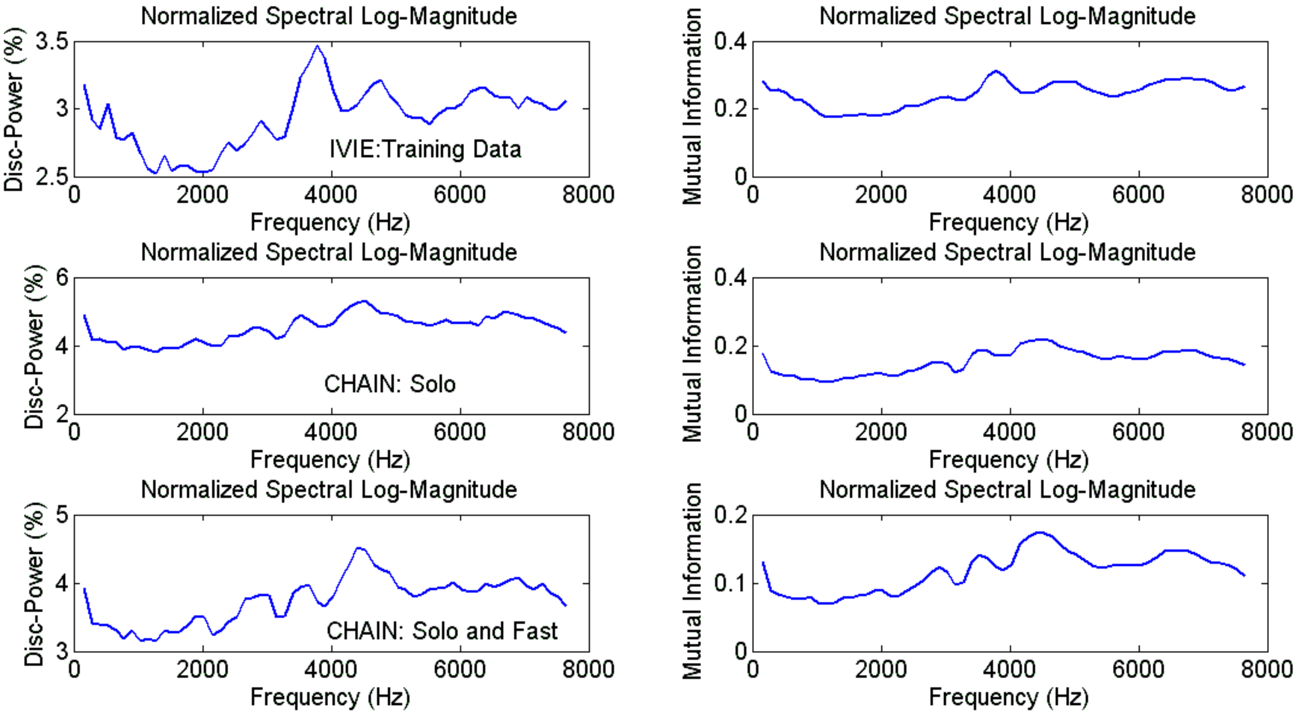


Figure G-3 Discrimination power and mutual information for three sets from CHAIN and IVIE corpora

G.4 Feature Design with Respect to Spectral Discrimination Power

To investigate whether higher values of the proposed metrics amount to higher verification accuracy when there is more focus on their associated frequency components eight filterbank are designed which focus on different frequencies in the range of 0-8 KHz. Cepstral coefficients were calculated on the output of these 8 banks (after log and DCT calculation) instead of Mel-banks.

Figure G-4 depicts the filter-banks along with a curve representing discrimination power values obtained from IVIE corpus[1]. The expectation is that filter-bank 4 and 5 should slightly outperform other filter-banks since they focus on more discriminant spectral areas.

[1] The curve is plotted to give some hints about peaks of discrimination power in the same figure as the filterbank and its unit is not the same as y-axis of the filterbank. The curve shows the cubed values of the discrimination power normalized so that its maximum becomes one (divided by the maximum).

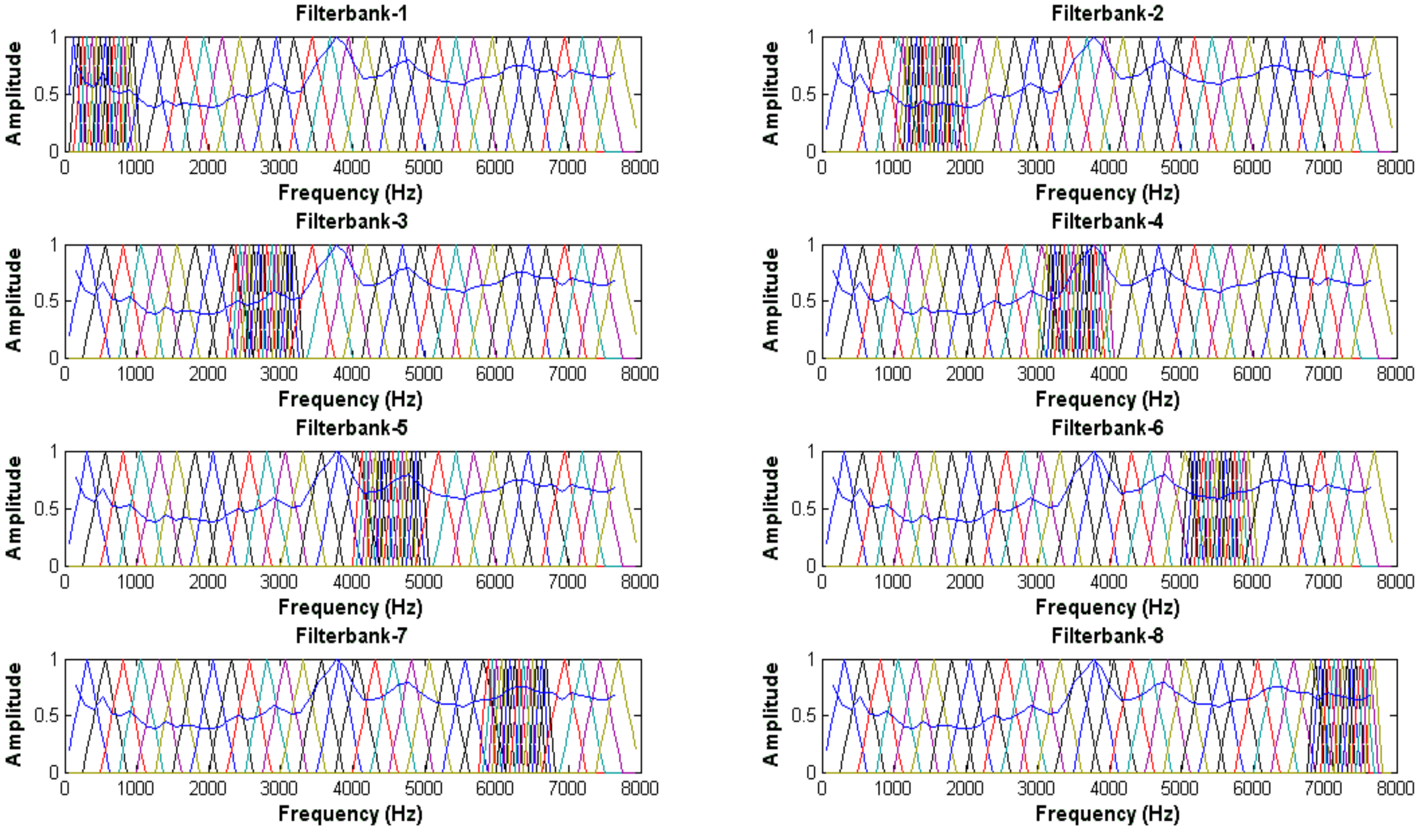


Figure G-4 Position and density of filters in 8 designed filter-banks across frequencies

G.5 Verification Results based on Use of Multiple Features

Eight sets of user and global models with different number of coefficients and mixture sizes were trained over the IVIE training subset. Figure G-5 displays the results for two sets of experiments: when 16 cepstral coefficients are used with 32 Gaussians in the components and when 24 coefficients are used with the models consisting of 64 Gaussians.

The first noticeable fact is that the best filterbank is the one which focuses on the low-frequency area (Filterbank-1). Comparing the error rates of FB-1 with those reported in chapter 5, reveals that this filter-bank works as well as Mel-bank. Verification errors for adjacent banks are consistent with the outcome of discrimination analysis when 24 coefficients are used for verification. For 16 coefficients however there are some discrepancies.

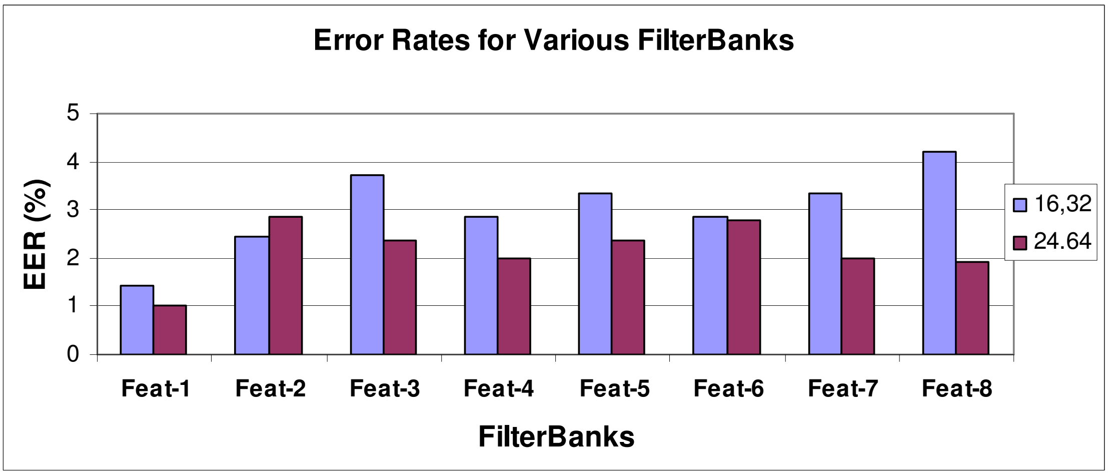


Figure G-5 Verification results in two set-ups for 8 filter-banks

The study suggests that even though the discriminative power of spectral components is not uniform across the entire frequency domain, the gain obtained by focusing on various spectral areas is little for verification purposes at least for clean speech. Nevertheless these features are used in the rest of this thesis for multi-feature analysis to reduce the effect of channel distortion and noise contamination.

Appendix H: Experimental Results Related to Speaking Style

H.1 Design and Purpose of Experiments

The aim of the experiments reported here is to determine the effect of incompatibility between training and test conditions in terms of styles of speaking on the verification results. Here the design of experiments is presented by description of corpora, silence removal module and 3 solutions tested to reduce the effect of style:

1. Corpora

The experiments are carried out on both corpora of CHAIN and IVIE. Three types of speaking styles are analysed (read, spontaneous and fast) in the following subsets:

- Read sentences (subset B of IVIE and Solo Test subset of Chain) which are the sentences normally used for speaker verification (denoted by 'Sentences' here)
- Read passages which include a passage of Cinderella story (from both Chain and IVIE corpora). One paragraph is chosen for each speaker from read passages.
- Retold passages in which the speakers retell the Cinderella story in their own narration (available in both corpora)
- Fast read passage with the same content of read passages (only in Chain)
- Fast spoken sentences (the same sentences available in Solo subset, only in Chain)

2. Silence Removal Module and Hand-trimming Data

To confine the signals' variation to the style of speaking it was made sure that for IVIE corpus all the data for one speaker was collected in one session and with the same set of equipment. The data collection setup for CHAIN experiments varied for different sets as explained in chapter 5. A possible source of error especially in the scenario tests (practical applications) is that due to the pauses and diversity in the amplitude of the signal the segments under investigation do not include enough useful data from speakers. In spontaneous and read passages especially in IVIE corpus there are periods of silence and hesitation. To make sure that this has not caused the unreliability of verification, this parameter is compensated for by use of voice activity detection module described in chapter 5, as well as manually hand-trimming data. For the read passages of

IVIE, the first sentence for each speaker was manually singled out in another subset[1] (Read, Hand-Trimmed). For retold passages, silence periods were manually omitted for speakers to construct a new subset (Retold, Hand-Trimmed).

3. Solution 1: expanding training data

Previous research reported in chapter 3 showed that each speaker acts differently in terms of varying phonological and acoustic characteristics when adopting a new style. While currently there is no technique to translate features from one style to another a possible solution would be enhancing the training data by inclusion of recordings from other styles. The success rate of this approach is analysed through experiments.

4. Solution 2: style detection

If training models on various speaking styles fails, another solution will be training style-dependent models which can be used after detection of speaking style.

5. Solution 3: use of style-independent features.

Through use of multiple features focusing on various spectral areas which were elaborated on in Appendix G we can figure out whether some parts of spectrum are less affected by change in the speaking style or not. As discussed in the next sections we can also determine whether it is possible to treat changes in the style as feature transformations similar to what we will do for channel compensation or we should accept those variations as non-linear and complicated change in the characteristics of the source signal.

H.2 Experimental Results

Two types of models were used in the experiments. The first set of models (abbreviated as MA) used 16 cepstral coefficients in components consisting of 32 Gaussians in user and global models. For the second set of models features were 24 cepstral coefficients and mixture size of 64 was used.

Figure H-1, shows the results of experiments on retold passages on IVIE corpus with VAD parameters of 25 and 0.05 as described in chapter 5. The length of test segments was limited to 3, 5 and 8 seconds after silence removal in different experiments to determine whether adding more

[1] The first sentence 'Once upon a time there was a girl called Cinderella' did not exist for one speaker, which was replaced with another sentence from the same paragraph.

data can improve the results[1]. From the figure one can find out that firstly the length has not played a significant role and secondly in all the cases the effect of change in the style of speaking (from read sentences to read passages) is substantial[2].

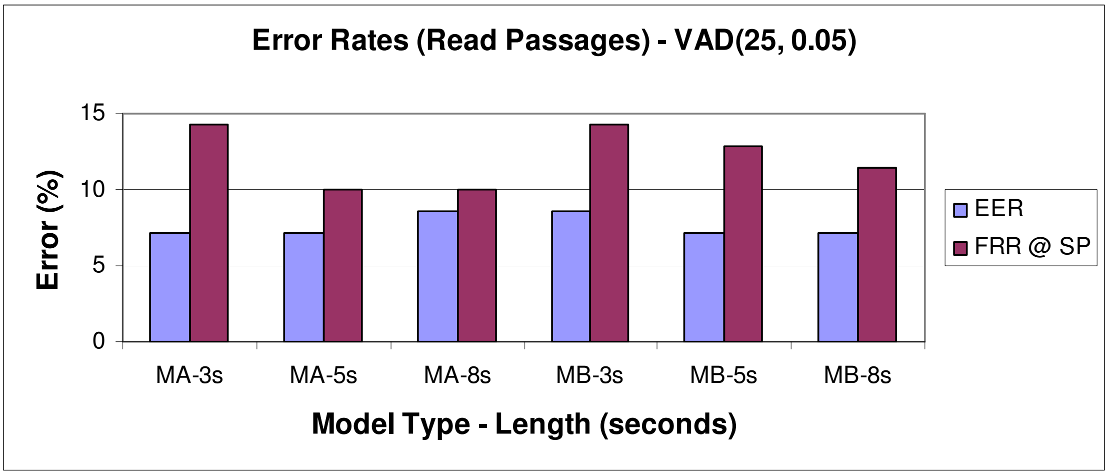


Figure H-1 Verification Results for Read Passages (IVIE Corpus)

For IVIE corpus errors for retold passages were slightly higher than read passages but the difference is not significant. For IVIE corpus the errors varied between 7 to 10% for different models, lengths and VAD parameters. Table H-1 summarises the results for IVIE corpus. In the reported experiments verification errors for hand-trimmed (HT) samples are also specified in the third and fifth column of the table. The maximum length of recordings in each category is mentioned in the last row. To emphasise on the fact that five seconds of data can provide enough information for verification, three test sentences were concatenated to make speech segments with the average length of 4.62 seconds. The EERs for those segments for both models were zero (last column of the table). It is noteworthy that hand-trimming has not constantly and consistently reduced the EERs. This fact confirms that the error is caused by factors other than pauses, artifacts and disorders in the long and spontaneous recordings.

[1] The algorithm initially took two times of the desired length and then removed silences. For this reason in some cases the length of test data when not enough speech was available in two times of the target length, was slightly less that desired length.

[2] The EERs for MA and MB models with same VAD parameters were 1.43% and 1.90% for test sentences. EER and FFR@SP (FRR when FRR=3*FAR) errors when not specified otherwise are calculated on the subset under investigation and are not the errors at the baseline EER thresholds.

Table H-1 Summary of Style-Related Verification Errors for Various Subsets of IVIE

	VAD	VAD	VAD	VAD	NO-VAD
EER (%)	**Read**	**Read-HT**	**Retold**	**Retold-HT**	**Test-Appended**
MA-3s	7.14	5.71	10.00	N/A	N/A
MA-5s	7.14	N/A	8.57	8.57	0.00
MA-8s	8.57	N/A	7.38	N/A	N/A
MB-3s	8.57	8.57	8.57	N/A	N/A
MB-5s	7.14	N/A	8.57	8.57	0.00
MB-8s	7.14	N/A	8.57	N/A	N/A
Total Length (s)	52.67	2.57	28.38	10.29	4.62

For CHAIN corpus error rates are summarised in Figure H-2[1]. Error rates with VAD module were similar and followed the same pattern for different subsets. The errors for retold passages are comparable to those of IVIE[2]. Fast sentences and fast read passages have similar errors which are significantly high. This figure also suggests that the effect of VAD is minimal (in other words the error is not inflicted due to silences and artifacts in signal).

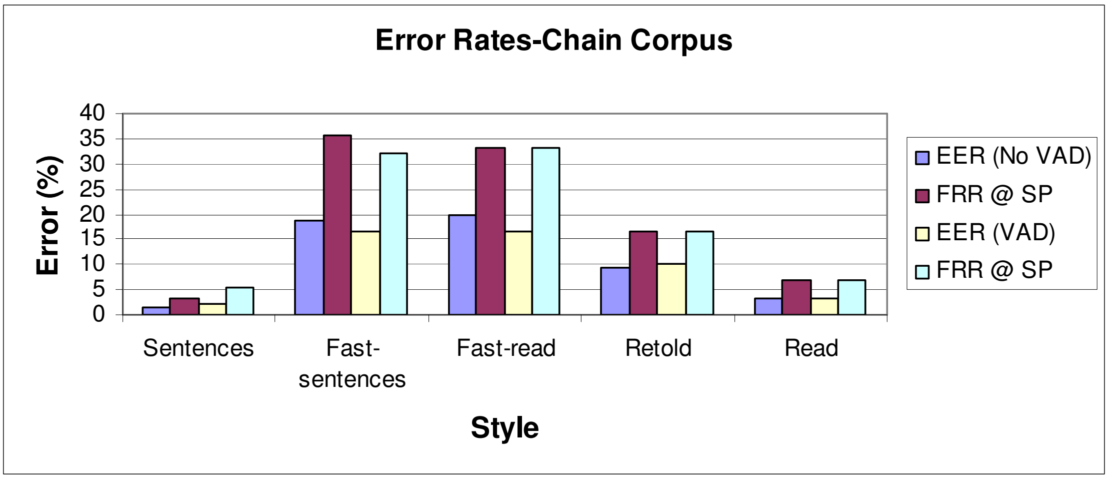


Figure H-2 Verification Results for Various Style-Related Subsets (Chain Corpus)

[1] These results are achieved using models employing 24 cepstral coefficients and mixture size of 64 with and without VAD module. VAD parameters used here are 50 and 0.1 as expanded on in chapter 5.

[2] One remarkable difference between results of IVIE and CHAIN corpora is the dissimilarity of errors for read passages. It could show that the read passages of CHAIN corpus (synchronous recording) are collected in a more controlled environment with the outcome similar to the normal sentences (demonstrating that the change of style by speakers is spontaneous and optional)

The result of testing three solutions proposed before is reported here:

Solution 1: expanding training data

For the next set of experiments (on IVIE corpus) the user models were trained on three different sets[1]. In the first experiment the ordinary models (trained on Subset-A of IVIE) were used (the same MA models). In the second experiment the user models were trained on the third part of retold recordings (which was not used for testing). In the last experiment the user models were trained on both subsets. The result is summarised in Table H-2. For the models trained on retold recordings the error rates of verification for isolated sentences is 10.87% which is close to the error of verification of retold recordings by regular models. It shows that the mismatch is the source of error. Another interesting finding is that the verification error for read passages is high in the first and second experiments. It shows that recording of read passages exhibit a separate sets of characteristics (which is not the same as retold or isolated sentences). Finally the third experiment shows that by inclusion of both types of styles in training data we can achieve lower error rates for both types of styles and there is no need for style detection and use of specific models. It is worth noting that by using data from both styles for training models the verification results for the read passages is also improved (but not as significantly as for those two styles). This suggests that expanding training data to include various styles enables reaching more reliable decisions even in new conditions and for new styles.

Table H-2 Summary of Style-Related Verification Errors with User Models Trained on Various Subsets

	User GMM Models Trained on			
EER (%)	**Sentences**	**Retold (8s)**	**Retold and Sentences**	**Length (s)**
Retold	**10.00**	**1.43**	**2.86**	8.00
Retold-HT	**8.57**	**1.43**	**1.43**	8.00
Read	**7.14**	**9.76**	**5.71**	8.00
Read-HT	**7.14**	**10.00**	**5.71**	2.57
Test	**1.43**	**10.87**	**1.43**	1.54

Solution 2: style detection

An attempt was made to detect the speaking style by training global models on the retold recordings (from the same set of speakers whose recordings were used for training regular global models). The style with higher score of global model was chosen as correct style and both user

[1] The global models were trained on the isolated sentences like before.

and global models for that particular style were used for verification. A similar idea (as explained in Appendix I) was previously used by Reynold (2003) for channel detection. The EER for verification of Retold (hand-trimmed) subset of IVIE using this method was 7.14% which determined that this was an unsuccessful attempt. In 67.14% of the cases the global model *was wrongly chosen*. The experiment suggests that unlike channel transformation, style related change is not a predictable and similar-for-all-users transformation.

Solution 3: use of style dependent features

In the last set of experiments 8 set of features calculated on filter banks 1-8 introduced in Appendix G were employed for verification. As Table H-3 demonstrates, there is no invariable spectral area, which focusing on it produces significantly better verification results.

Table H-3 Verification Errors for 8 Features Focusing on Various Parts of Spectrum

	Hand Trimmed	**Hand Trimmed**
EER (%)	**Retold-maximum 8sec**	**Read-2.57s**
Feature 1	**9.76**	**7.14**
Feature 2	**8.81**	**8.57**
Feature 3	**8.57**	**7.14**
Feature 4	**10.00**	**8.33**
Feature 5	**10.00**	**11.43**
Feature 6	**8.57**	**10.00**
Feature 7	**8.10**	**10.24**
Feature 8	**8.57**	**11.67**

Appendix I: Noise and Channel Effect

I.1 Theoretic Study of Channel Distortion and Noises

Assuming that channel effect could be modeled as a linear time-invariant transfer function and that noise effect is additive, Figure I-1 displays a model for distortion caused by channel and noise during a re-recording process.

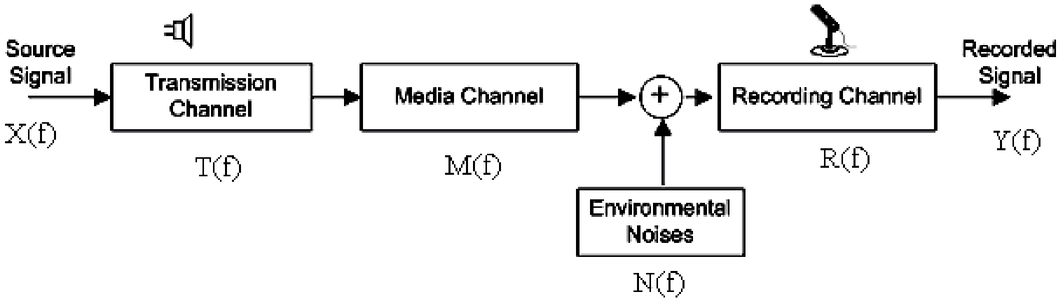


Figure I-1 Model of Channel Distortion and Noises

If $X(f)$ denotes the Fourier transform of the source signal, $T(f)$ the linear transfer function of the transmission channel (loud-speaker), $M(f)$ the linear transfer function of the media, $N(f)$ the Fourier transform of noise (assuming stationary noise) and $R(f)$ the linear transfer function of the recording device, the FFT of recorded speech $Y(f)$ will be:

Equation I-1

$$Y(f) = X(f).T(f).e^{-\alpha(f).d}.R(f) + N(f)$$

The attenuation in the media can be modeled as follows[1]:

Equation I-2

$$M(f) = e^{-\alpha(f).d}$$

[1] *Classical absorption coefficient* determines the amount of attenuation of signal as a function of distance: $S(\omega) = S_0(\omega)e^{-\alpha(\omega).d}$ in which, S_0 is the source signal, S is the received signal at distance d, and $\alpha(\omega)$ is the classical absorption coefficient as function of frequency. α could be expressed in *Np/m* (Neper per meter).If α is one Neper per meter, the sound wave is *e* (base of natural logarithm) times attenuated in each meter (See Kinsler et al. 2000).

where α is a frequency specific attenuation factor which yields:

Equation I-3

$$Y(f) = X(f).T(f).e^{-\alpha(f).d}.R(f) + N(f)$$

The above assumption about the noises may be somehow simplistic since there are transmission noises, medium noises and recording noises, but we could represent all these factors (channel distortion and noises) as a multiplying and additive elements in the distortion equation:

Equation I-4

$$Y(f) = C_A(f).(X(f) + N_A(f))$$

or:

Equation I-5

$$\log Y(f) = \log[X(f) + N_A(f)] + \log C(f)$$

If we can make three assumptions of linearity of the transform functions, proper width (long enough) of analysis window (frame window) and invariability of channel distortion for each filter (the channel gain could be approximated by a single value for all the spectrum covered by filter) the channel distortion could be modeled as a constant factor in the log-output of filters (as specified by Jin et al. (2007) the length of the channel impulse response should be shorter than the spectral analysis window). These facts have been discussed in many works dealing with channel effect e.g. Pelecanos and Sridharan (2001). Since the discrete cosine transform is linear the channel effect (if noise is ignored) is additive for cepstral coefficients.

The noise effect on cepstral coefficients is apparently more complicated. Openshaw and Mason (1994) observed three distortions in cepstral coefficients as a result of adding noise: a shift in mean and a change in variance, steeper edges for distribution of cepstral coefficients and tendency towards non Gaussian and bimodal distributions. All the three effects were stronger for the lower order cepstral coefficients.

The main purpose of channel compensation for speech and speaker recognition is to reverse these effects. The next section elaborates on the previously proposed solutions.

I.2 Available Solutions to Reduce the Effect of Channel Distortion

Cepstral mean subtraction (CMS) or cepstral mean normalization (CMN) is one of the simplest yet very effective solutions to channel distortion. Since the channel effect is additive, for cepstral coefficients, deducing the mean of cepstral values, which causes loss of mean information of the cepstrals, removes those added values (DC shift). In one of the many studies on this topic Ortega et al. (1999) tested effect of cepstral mean normalization with or without score normalization in different mismatched conditions on telephony and microphone corpora with text-dependent and independent modeling. They concluded that score normalization[1] is essential and that, the combination of CMN and score normalization decreased EER significantly. Extra to mean subtraction, mean and variance normalization (MVN) is another approach in which both the mean and variance of cepstral coefficients are adjusted to zero and one respectively.

Heck et al. used neural networks to transfer cepstral, log spectrum and prosodic features to obtain robust features for speaker recognition over the telephone (2000).

Pelecanos and Sridharan (2001) showed how noise and channel parameters affect MFCC coefficients and how normal distribution warping can reduce the effect of noise and channel. While they did not give specific figures for improvement, the DET curves showed slight but consistent improvement over other techniques such as CMS and mean and variance normalization (MVN).

Xiang et al. (2002) showed that short term Gaussianisation which involves a global transformation of features followed by CDF[2] matching as a way to compensate channel and handset variability gives 20% improvement on baseline system with cepstral mean subtraction (CMS).

Reynolds (2003) proposed a feature mapping technique in which channel specific models were adapted through maximum a posteriori (MAP) estimation from a channel independent root GMM. During verification the most likely channel dependent world (background) model was chosen and each feature vector in the utterance was mapped to the channel independent space based on its top decoded Gaussian in the channel dependent GMM. In their experiments the method showed to improve the EER from 9.1% to 8.7% if used along with T-normalization.

Alexander et al. (2004) analysed the effect of mismatched recording on human and automatic speaker verification in *forensic* application for telephone network, GSM and background noise. For automatic recognition however they used the same acoustic features and models as ordinary systems (GMM modeling and RASTA-PLP features). One of their important findings was that

[1] They used a type of normalization called nearest reference speaker normalization based on the closest model.

[2] Cumulative Distribution Function

automatic speaker recognition outperforms aural speaker recognition in matched conditions however in mismatched recording conditions automatic systems showed accuracies comparable to aural recognition when the channel conditions were changed.

In addition to applying cepstral subtraction Wu and Cao (2005) suggested median filtering on the frequency components over the time to remove highly varying components. They also argued that since the log transformation is sensitive to noise because of the steep slope in low energies the mismatches between the clean speech and the noisy speech are very large for low-energy banks. Since the low-energy banks tend to be affected by noise they replaced the log function by a new power function for smaller values.

Teunen et al. suggested a speaker independent model transformation technique which showed 20% relative improvement over cepstral mean normalization (2000). Pelecanos et al. suggested a feature transformation technique between channels based on maximization of joint posterior probabilities (2006). For NIST (2000) database, they reached 8.6% and 14% relative improvement in EER for two systems over baseline system for both electret and carbon microphone dataset.

Colibro et al. (2006) proposed a feature compensation technique based on moving the observation vector by a weighted sum of the channel compensation offset values (instead of just one Gaussian) and argued for the superiority of feature compensation over model compensation in which the features can be used by any model later. Jin et al. (2007) applied reverberation compensation, channel warping and multiple channel compensation for reducing effect of acoustic mismatch.

The methods offered can be divided into two groups:

- Methods for which applying them does not require any information about the channel in form of previously collected data or type of channel such as cepstral mean normalization, median filtering, RASTA and Normal distribution warping.
- Methods which require previous data collected from the channel, such as MAP adaptation and feature mapping. They may need specific information about the channel or determine the channel based on the available data.

The gain offered by group one methods is very close to that of cepstral mean normalization with a slight relative improvement. On the other hand the success of methods in group 2 largely depends on the amount of data available from channel and similarity of test and channel-specific information.

I.3 Types of Noise

Noises with various characteristics can be present in the recording environment. They may be stationary noises (stochastic characteristics of the noise remains constant over the time) or non-stationary (e.g. made in an unpredictable point of time: a horn or driller sound).

Table I-1 Types of Noises to be Included in Evaluation Tests

Noise Type	Subcategory / Description
Gaussian Noise (Stationary)	White noises (equal energy across entire spectrum) at several SNRs.
	Color noises (e.g. Pink, Brown) at several SNRs
	Low-frequency noises
	Narrow-band noises
Non-Stationary Noises	Spike like noises such as gun-shot noises
	Special case noises in specific environments (e.g passing of the car)
	Noises with changing profiles (switching between different types of stationary noises at different powers)
Speech Like Noises/Inference	One dominant voice over speech from another speaker on one channel (linearly mixed or linearly filtered and mixed)
	One dominant voice over speech from several other speakers on one channel (linearly mixed or linearly filtered and mixed)
	One dominant voice over speech from one/several other speaker(s) on stereo/multiple channels, linear mixture
	One dominant voice over speech from one/several other speaker(s) on stereo/multiple channels, with different channel characteristics
	Identification of one (non-dominant) speaker in the speech from several speakers

The frequency characteristics of noises may vary e.g. they could be low frequency, narrow-band, white band (having equal energy across spectrum) or color noises (with various power profiles across spectrum, e.g. pink, brown or gray). Noises could be speech like (known also as inference) for example in cases where two or more speakers talk simultaneously, cocktail party noises or office noises. The ratio of power of signal to noise (signal to noise ratio, SNR) is usually expressed in dB. Table I-1 presents the types of noises and suggested evaluation profiles for noises.

In the next section previous solutions to reduction of effect of noise in speech applications is summarised in two categories of speech like noise (inference) and non-speech noises.

I.4 Previous Research and Possible Solutions to Noise Problem

1. Non-speech Noises

Several noise reduction techniques in time and frequency domain and on the basis of various assumptions about availability of data about noise type, noise models and noise power have been proposed in the realm of speech processing.

Similar to channel compensation the simplest method for reduction of noise effect is cepstral mean subtraction or relative spectra (RASTA) filtering. These two methods require making no assumption about the noise type or data from similarly contaminated signal.

Some of the noise compensation methods are based on constructing pre-made profiles (and codebooks) for transformation of features in presence of the noise. These methods may require stereo channels, simultaneously recorded signal (in train and test environments) or knowing about the SNR. In a comparative study of the noise compensation methods for speech recognition, Liu et al. (1993) described and compared the performance of a series of cepstrum-based procedures for noise reduction in speech recognition. Those procedures included SNR-dependent cepstral normalization (SDCN), codeword-dependent cepstral normalization (CDCN), fixed codeword-dependent cepstral normalization (FCDCN), Multiple fixed codeword-dependent cepstral normalization (MFCDCN), cepstral mean normalization (CMN) and RASTA. In their results CMN showed higher (consistent) improvement in comparison with RASTA. The other methods were based on transformation of the cepstral vector extracted to a new vector based on the previous information about the environment using codebooks and in some cases required stereo recordings and simultaneously recorded data in test and train environment. Unlike MFCDCN and FCDCN, CDCN did not require stereo or simultaneous recording.

A class of noise reduction techniques consists of methods which are based on the fact that log-amplitude of two additive signals (used in cepstral analyses) can be approximated by a maximum function:

Equation I-6

$$\log(|X+N|)=\log(\max(|X|,|N|)$$

where *X* and *N* are the spectrum of signal and noise.

Deoras and Hasegawa-Johnson (2004) adopted this assumption and suggested an HMM model for speech recognition named Factorial HMM (FHMM) based on the works of Ghahramani and Jordan (1996). They tested the algorithm on simultaneous word recognition which showed some improvement in recognition rates especially when the digits from simultaneous speaking speakers where not the same. The max-approximation and masking had been proposed before e.g. Varga and Moore (1990) used the Klatt masking algorithm in which if either the model mean or the observation was below the noise mask then it was replaced with the noise mask. They tested the algorithms on pink noises and machine gun noise for word recognition.

De-Wet et al. (2005) applied several mismatch reduction techniques including time domain noise reduction (by estimating SNR of spectrums using voice activity detection modules and applying Wiener filters), as well as histogram normalization and mean variance normalization in presence of additive noises. In many combinations they achieved little or no improvement for speech recognition when these techniques were applied to MFCC features.

Some other methods have targeted changing in cepstral feature extraction especially reducing the effect of log operator. Ravindran et al. (2006) suggested three improvements in MFCC extraction techniques for reducing noise effect: first, use of root compression instead of log, second, smoothing the filter banks, low pass filtering and down-sampling of outputs and finally calculation of 'spatial derivatives' which were based on difference in adjacent channels (equivalent to filters in the bank). They used these improvements in a speech recognition system which showed relatively small improvements in recognition rates for SNR around 10-20 dB despite that the improvements were higher for very low signal to noise ratios (usually unimportant from a practical perspective).

A promising approach towards reducing the effect of noises on speaker recognition seems to be sub-band filtering. In sub-band filtering, each feature set is extracted after applying a band-selective filter to the signal (bands are frequency regions). Damper and Higgins (2003) used sub-band filtering and score fusion by product rule (summation of log probabilities). They observed a drastic improvement as a result of sub-band filtering and combination of scores, in case of narrowband noises. The results showed higher improvements when number of sub-bands was increased. The drastic improvement in identification rates (from 25% for two sub-bands to 100% for 16) can be explained in this way: by sub-band filtering fewer bands are affected by the narrow-band noise and the overall results are shaped by the features extracted from unaffected frequency areas. Chen et al. (2004) used wavelet decomposition for sub-band filtering and then extracted LPCC features on the results of each sub-band as well as full band signal and combined the scores from those multiple sources. They tested two types of score combination, different SNRs and variable number of sub-bands (2-4). The suggested techniques outperformed the traditional GMM with LPCC features especially in low SNRs (yet neither consistently nor substantially as reported by Damper and Higgins). Chen's experiments on white noise could be read as that the former improvement has been due to narrow-band nature of noise, which affects only limited number of sub-bands.

2. Inference: Speech-Like Noises

The success of signal separation in speech-like noises is largely dependent upon availability of single or multiple channels and the type of mixing.

Two general approaches for speech separation are adopted in the literature one based on blind signal separation and another based on computational auditory scene analysis.

Blind signal separation (BSS) or blind source separation techniques try to recover unobserved signals or 'sources' from observed 'mixtures' (Cardoso, 1998). It is assumed that each sensor in the environment receives a combination of source signals (*X*). These methods are labeled 'blind' since the source signals (*S* or S') is not observed and the mixture matrix (*A*) is unknown. The constraints and information about the source signals may vary for example we may know the distribution, some features about distribution or signal, its parametric family or we may have no information about the source signal (Cardoso, 1998).

In the realm of blind source separation Laheld and Cardoso (1994) proposed a class of algorithms for separation of a set of independent signals from some linear mixtures of them. The algorithms they named 'parameter free separators' (PFS) were based on 'serial updating' of the mixture matrix over the time. One constraint of the analysed problem was that the mixture matrix should have been full column rank and that the number of sensors should be equal or greater than number of independent sources. Jang and Lee (2003) proposed a blind source separation technique based on gradient ascent, MLE and independent component analysis (ICA) for separation of linearly mixed signals in a single channel.

With the aim of recognition of simultaneous speech from multiple speakers on a single channel Raj et al. (2005) employed a method based on non-negative matrix factorization (NMF) to separate power spectrums (with strictly positive values) of two mingled signal along with the spectral max estimation. Their recognition results were poor and little or no improvement was achieved due to these compensation methods in many cases.

A series of algorithms have been developed for the problems involving use of two or multiple channels in which the position of the sources of signal should be determined with applications in robot speech recognition such as Nakadai et al. by an active direction pass filter (ADPF) on stereo signals (2003, 2004) and by a micro-array and use of geometric source separation (Valin et al., 2007).

In a series of works Koutras et al. applied the ideas of BSS to the problems of phoneme recognition with different constraints and set-ups (1999, 2000, 2001). In (Koutras, 1999) they developed a frequency based signal separation based on the minimization of the cross correlation of the separated speech signals and applied the algorithm to the problem of phoneme recognition on TIMIT database and three artificial phoneme mixing scenarios. The key points about their works in 2000 and 2001 was that they established the separation technique on maximum likelihood estimation and made the assumption that the probability density of the speech signal is Laplacian based on the findings of Charkani and Deville (1997).

Trivedi et al. (2005) used independent component analysis along with MFCC coefficients and vector quantization on a very small database consisting of data from 4 channels but with only 31 training instance and 11 test instance. Using vector quantization and very small size of dataset, hinders generalization about the results.

Walsh et al. (2007) used a Kalman filter in which the model is the room acoustic and the noise is the speech signal from several speakers and uses expectation propagation (EP) framework to iteratively identify speakers and separate features. Their application was slightly different, in which several speakers present in a recording had to be identified.

In contrast with blind signal separation the second approach tries to exploit information about the nature of speech. Computational auditory scene analysis (CASA) is a field of science which aims to use lessons learnt from auditory scene analysis (ASA) for simulating the same abilities in machine. “ASA is the ability of listeners to form perceptual representations of the constituent sources in an acoustic mixture” (Brown and Wang, 2005, p. 371). Based on the explanations of Bregman (1990) Brown and Wang describe that in ASA process in the first stage, the acoustic mixture is divided into significant acoustic events and in the second phase a grouping process combines elements coming from the same acoustic source by forming a ‘stream’. Han et al. (2006) designed a CASA based speech separation system consisting of auditory peripheral model, pitch tracking, separation of signal into segments, combining segments into streams, assigning streams to speaker based on speaker recognition. The results of recognition on separated speech showed deterioration compared to the unprocessed speech. They inferred that the method destroyed the spectrum of speech especially for the unvoiced segments.

In a series of work culminating in their paper in 2008 Shao and Wang (2008) proposed a new feature set similar to cepstral coefficients based on Gammatone filters and DCT (without logarithm function) which they named GFCC. They used a CASA based approach with pitch-based speech separation (Shao et al., 2007) and time-frequency binary masks (Srinivasan and Wang, 2007). They reported a considerable improvement in speaker identification rates in face of speech-shaped noise[1], and several non-stationary noises including speech babble, destroyer operation, factory and cockpit noises.

[1] Having long term spectrum of speech.

Appendix J: Re-recording/Channel Experiments

J.1 Design of Re-recording Experiments

Re-recorded voices are the test sentences in Subset B (I1, S7, and S8) played by a loud-speaker and recorded by two different microphones[1] in a quiet large room. All the sentences read out by 70 speakers were concatenated and saved as a file, and then were played and recorded at the target distance (between the speaker and microphone).

The recording distances were 30cm, 60 cm, 90 cm and 150 cm.

After re-recording, based on the start of file which was marked by a pure sinusoidal signal, the sentences were split. Therefore it was possible to find frame by frame correspondence between original signal and re-recorded files. The 8 resulted sets of recordings were labeled G30, G60, G90 and G150 and B30, B60, B90 and B150 based on the microphone and distance of re-recording.

The experiment reported here consists of the following subcategories:

1. For all the subsets, verification by use of normal cepstral coefficients is carried out and reported.
2. The suitability of uniformly distributed filters compared to Mel-bank is examined[2].
3. Mean and variance normalisation (MVN) is exercised and the results are reported as a benchmark for further comparisons.
4. Based on the frame-by-frame comparison of spectral characteristics the channel transfer functions is determined. A discussion will follow on how much the assumption of linearity holds and how much a linear transfer function applied to all signals can improve the verification results (evaluating the possibility of using a generic transfer function based on the recording device).
5. The error rates for multiple feature fusion with and without cepstral MVN will be reported.

[1] The namely specification of microphones will not used in this study. The first microphone labeled B here was a Lanyxe l-604 microphone, polar pattern, omni direction, electrect condenser, 50-16khz, -58+-3dB 0dB=1v/micro Bar at 1khz, 1v-10v operation (standard 3v). The second microphone labeled G was a Philips SHM1000/97 microphone, sensitive directional, 10-10KHZ, -40+-3dB , with impedance of 2.2 K-ohm at 1khz

[2] When channel effect is significant distribution of Mel-cepstral filters across spectrum may not be optimal since the broad filters in higher frequencies refute the assumption concerning the width of filters presented in Appendix I.

6. Score fusion and combination of features as a solution for channel distortion is examined.

7. The error rates (FAR and FRR) at the thresholds set based on the original test signals (baseline system) is reported under acoustic mismatch conditions.

J.2 Multi-feature Fusion for Channel Compensation

Several score fusion techniques have been proposed which were briefly introduced in the literature review. A common and simple practice for score fusion is product rule when the scores are probabilities and sum rule when the scores are log-probabilities.

In addition to these methods, a few new fusion techniques are proposed and examined here.

1. Product fusion score (PFS) or fusion based on biometric gain against impostors (BGI)[1]

Fusion based on biometric gain against impostors (BGI) is exercised as a method which requires knowing PDF of the genuine and impostor speakers. BGI specifies how many times it is likely that the claimant is an impostor after observing the biometric piece of evidence than it was beforehand (Sedgwick ,2003).

If *S* ($S=(s_1,....,s_K)$) consists of scores from *K* devices or algorithms and $P(I)$ is the a-priori probability of being an impostor we can define BGI as:

Equation J-1

$$BGI(O)=\frac{P(I\mid O)}{P(I)}$$

It was shown by Sedgwick that a good approximation of BGI under the assumptions of independence and knowing that normally *P(I)*, the probability of a claimant being an impostor, is small, is modified BGI:

Equation J-2

$$\text{mod}-BGI(S)=\frac{\prod_i P(s_k\mid I)}{\prod_i P(s_k\mid G)}$$

when $P(s_k\mid I)$ is probability of receiving s_k from *k*-th device or source for impostors and $P(s_k\mid G)$ is probability of receiving s_k for genuine users.

[1] The final equations used here are the same as those used under the name of product fusion score (PFS) by Dass et al. (2005) and Nandakumar et al. (2006) as described in Appendix A.

For calculation of modified BGI we need to estimate the probability distribution function of scores for each device for impostors and true users.

When this method of score fusion is employed in the experiments, genuine and impostor probabilities are estimated by histograms for each feature-set based on one third of test data. The other two third of data is used for test purposes. The score for each observation (frame of speech) is the difference between the log probabilities of user and global model. Logarithm of modified BGI is used to assign one final number to each observation *O* (which is a feature sequence extracted on one recording/sentence):

Equation J-3

$$\log(\text{mod}-BGI(O)) = \frac{1}{N.K}\sum_{k=1}^{K}\sum_{j=1}^{N}[f_k^I(s_{k,j}) - f_k^G(s_{k,j})]$$

Equation J-4

$$s_{k,j} = \log P(o_{k,j} \mid \Pi_k^i) - \log P(o_{k,j} \mid \Pi_k{}^M)$$

$o_{k,j}$ is the *j*-th feature vector based on *k*-th algorithm. Π_k^i is the user model trained based on features from *k*-th algorithm for user *i*. $\Pi_k{}^M$ is similarly the global model for k-th algorithm. $f_k^I(.)$ is the PDF of impostors' score and $f_k^G(.)$ is the PDF for the genuine speakers. $s_{k,j}$ is the score associated with $o_{k,j}$. *N* and *K* are the number of observations (frames) in *O* and the number of algorithms respectively.

Seven other fusion rules which do not need any previous information about the impostor or genuine distributions were tested. The idea behind these methods is that, features with different filter resolutions in various parts of the spectrum, are affected by the adverse conditions to different degrees. If for each frame of speech we can employ the score obtained from less affected features we will be able to make more accurate overall decisions.

To demonstrate this, the change in the log-probabilities of a sentence and its linearly filtered version given a genuine user model, global model, and the average probability given the impostor models (for one speaker) is calculated and depicted in Figure J-1. The diagram shows the change in the log-probability before and after applying the filter for all the 8 features. The transfer function of the filter is also shown in the bottom sub-plot (The filter is combination of a low-pass Butterworth filter a high frequency band-pass Butterworth filter and an all pass filter).

As expected on the basis of channel transfer function, the probabilities assigned to the features 1 and 8 (the filter-banks focusing on the low and high frequencies) are affected to a higher degree. Despite that the amount of change is different for the global model, user model and impostor models[1].

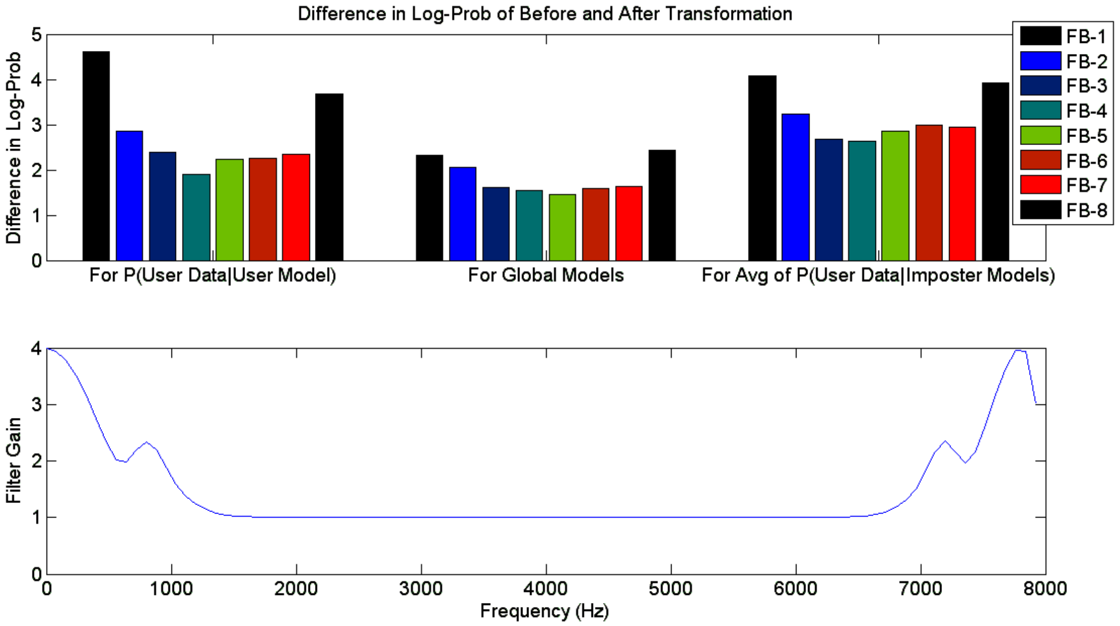


Figure J-1 A simulation showing how the scores are effected by a linear channel (top sub-plot: difference in scores obtained from features of 8 filterbanks for user/global/impostor models, bottom: frequency response of the filter)

2. Seven Rules of Fusion

For the *j*-th feature vector calculated using *k*-th algorithm ($o_{k,j}$), the log probabilities of that observation given the user model and global model are labeled as s_{kj} and g_{kj}:

Equation J-5

$$s_{kj} = \log P(o_{k,j} | \Pi_k), g_{kj} = \log P(o_{k,j} | \Pi_k{}^M)$$

since the decision at each point is based on the values for that point (sequence number, *j*) we can drop this index for the sake of brevity and at each point refer to the user and global model log probabilities as s_k and g_k. r will be the final score assigned at the time *j* to the frame as a result of fusion.

[1] The simulation is revisited after introducing other fusion rules.

Seven fusion rules were tested in the experiments which are presented below:

1. Rule one only engages the algorithm which has produced the highest g_k:

Equation J-6

$$r = s_u - g_u, u = \arg\max_k(g_k)$$

as the figure J-1 showed the features which have been largely affected by distortions receive lowest scores by both global models and user models.

2. In contrast to rule one, rule two engages the algorithm with the lowest g_k:

Equation J-7

$$r = s_u - g_u, u = \arg\min_k(g_k)$$

this rule is just kept to draw contrast with rule 1.

3. The justification for this rule is that, a higher distance between user and global model can be interpreted as a more ‘certain’ decision whether it is negative (for rejection) or positive (for acceptance):

Equation J-8

$$r = s_K - g_K, K = \arg\max_k(|s_k - g_k|)$$

4. Rule four engages the highest three values for $|s_k - g_k|$ and *u1, u2, u3* denote the index of these highest values:

Equation J-9

$$r = \sum_{k=u1,u2,u3}(s_k - g_k)$$

Instead of one value 3 pairs of scores are engaged by this rule.

5. The rationale behind this rule is that the score from user/impostor models are more affected compared to the global models (as demonstrated in Figure J-1) causing a higher difference for the less affected features:

Equation J-10

$$r = s_K - g_K, K = \arg\max_k(s_k - g_k)$$

6. The sixth rule is simply the sum rule:

Equation J-11

$$r = \sum_{k=1}^{K} (s_k - g_k)$$

7. The seventh rule uses scores from specific features (*u1,u2,u3*) which we might know would be more accurate based on previous information about the algorithms and channel.

Equation J-12

$$r = \sum_{k=u1,u2,u3} (s_k - g_k)$$

The simulation (illustrated in Figure J-1) lets us predict how successful the fusion techniques would be: the sum-rule for fusion simply adds the scores from all the sources. On the other hand the max-rules (rules 3 and 5) for fusion cause only the score with the highest values to participate in the final score and discard the others. If before the filtering all the user models produced the same difference of score (with the global model) the figure shows that, the scores obtained from more affected features would be reduced after transformation and they would be less assertive therefore are neglected in the fusion.

J.3 Experimental Results and Discussion

Experimental results are reported in the categories outlined in J.1.

1. Baseline experiments with re-recorded (distorted) signals

The sentences in the re-recorded subsets were verified by the system in various set-ups. Table J-1 specifies the models, setups and a short system description of the algorithm used in each experiment along with the EER[1].

Table J-1 Results of baseline-experiments for with various models

Abbreviation	System Details	EER (%)
MF41-MVN	MFCC-41 Filters-MVN	**4.52**
UN41-NO	Uniform-41 Filters-No MVN	**2.38**
UN41-MVN	Uniform-41 Filters-MVN	**4.29**
MF29-MVN	MFCC-29 Filters-MVN	**2.94**
MF29-NO	MFCC-29 Filters-No MVN	**1.43**

[1] Since the number of uniformly distributed filters across the spectrum was 41 for UN41 features, in addition to regular 29-filter filterbank the results for 41-filter Mel-banks were reported for comparison. 16 coefficients with 32 component GMMs were used as models. The table presents the equal error rate for verification of original test sentences (Subset B). MVN denotes mean and variance normalization.

The results show that discarding the information about mean and variance of features by MVN had inflicted around 2% increase in EER.

2. Improved Results by Mean-Variance Normalization and Linear Compensation

Table J-2 displays the error rates for all re-recorded subsets, in different setups[1].

The pre-processing with amplification (UN41-NO-AMP)[2] specified below needs some explanation. The spectral amplification functions for two microphones are displayed in Figure J-2 and J-3. The functions are estimated using outputs of 61 uniformly distributed filters across spectrum. The mean values of the spectral ratios (solid line) plus and minus one standard deviation (dashed lines) of the values are plotted. The top-left subplot of the figures for example, shows the ratio of filterbank outputs for recordings made at 30cm to those made at 60cm on a frame by frame basis and averaged over all the frames. This type of data can not practically be collected during verification session in which neither the type of recording device nor the distance of recording is known especially for remote recognition attempts. The methods however aid in analysis of the linearity assumptions and assessing the improvement made by the methods relying on the hypothesis of the linear transformation of frequency components through channel.

Table J-2 EERs for re-recorded signals verified through use of various features

EERs (%)	MF29-NO	MF29-MVN	MF41-MVN	UN41-NO	UN41-MVN	UN41-NO-AMP
G30	**19.92**	**10.56**	**11.03**	**8.57**	**8.10**	**4.37**
G60	**29.52**	**16.51**	**16.19**	**14.37**	**10.56**	**6.19**
G90	**31.51**	**18.02**	**17.22**	**14.76**	**9.92**	**6.11**
G150	**34.76**	**20.24**	**21.90**	**18.10**	**10.56**	**10.08**
B30	**16.19**	**8.02**	**8.57**	**16.59**	**9.52**	**5.16**
B60	**22.46**	**13.33**	**13.65**	**19.05**	**11.35**	**6.59**
B90	**23.97**	**16.67**	**16.11**	**17.62**	**12.54**	**7.62**
B150	**25.24**	**20.56**	**21.98**	**18.02**	**19.05**	**12.46**

An interesting observation regarding the transfer functions of two microphones in combination with various distances is that it is the joint effect of microphone and distance that determines overall effect. Each microphone has a certain range of linear behaviour in terms of frequency

[1] MFCC features with 29 filters with MVN (MF29-MVN), without any cepstral normalization (MF29-NO), with MFCC features with 41 filters and with MVN (MF41-MVN), with features based on uniformly distributed filters along with MVN (UN41-MVN) and finally with the same features without MVN but after amplifying frequency components based on the approximated spectral transfer function of the data for the entire subset at any particular distance and for each microphone

[2] NO refers to not using the MVN and not the amplification (amplification is performed).

and amplitude and when the speech signal is attenuated in the media, the spectral components may fall outside the linear range and undergo unpredictable changes. For microphone B the attenuation in 30cm, 60 cm and 150cm, is quite uniform and happens mainly between 2500 and 4000 Hz. For the other microphone (G) the attenuation around 2 kHz is persistent but higher frequencies have rapidly disappeared after 30cm, causing attenuation ratio from 30cm to 150cm for higher frequencies to be inconsistent with that for 30cm to 60cm.

The EERs in Table J-2 demonstrate that using a linear approximation of transfer function in frequency domain for small distances can drastically improve the results. For higher distances due to the fact explained the amplitude of signal falls below certain thresholds and the non-linearity of the microphones and domination of noises result in high error rates.

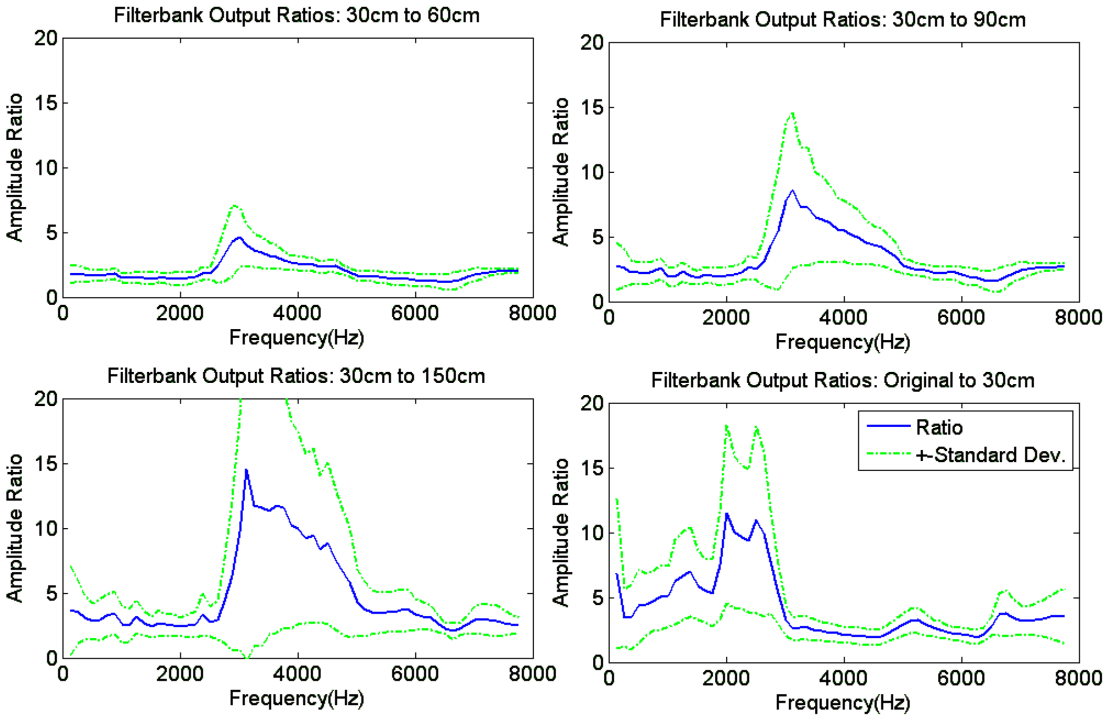


Figure J-2 Estimation of spectral distortion due to re-recording at various distances for Mic.B

Uniform filters have outperformed Mel-filters and MVN has been extremely helpful.

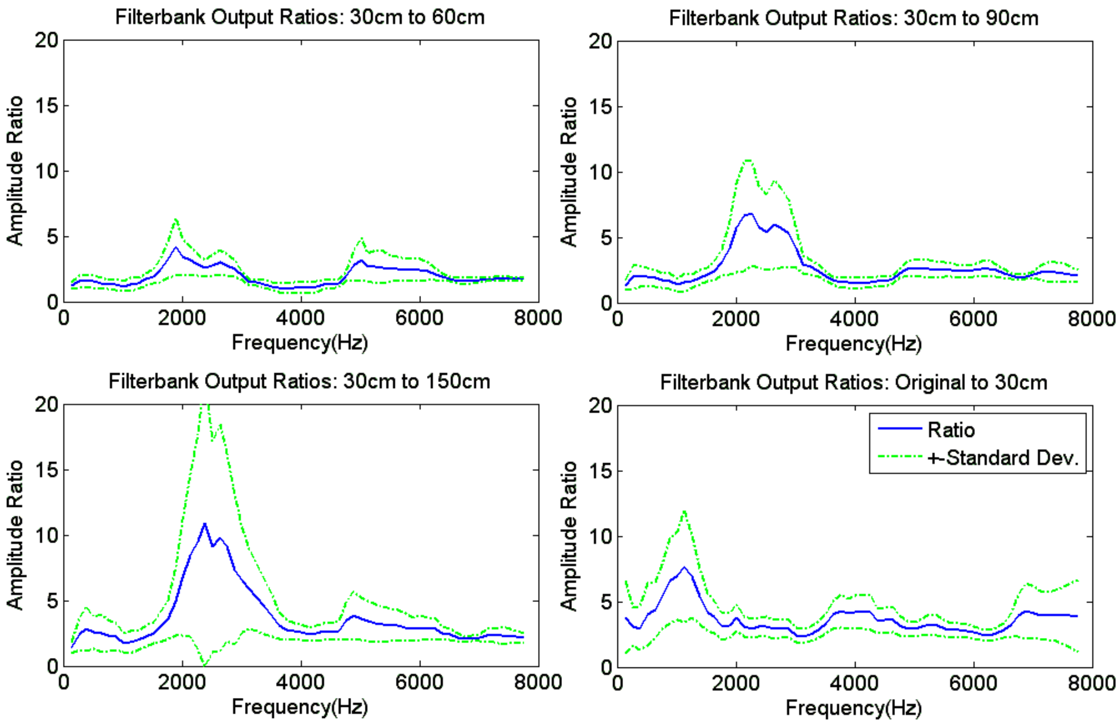


Figure J-3 Estimation of spectral distortion due to re-recording at various distances for Mic.G

Table J-3 Verification results for 8 features

EER (%)	G30	G60	G90	B30	B60	B90	Retold	Read
Feature 1	7.62	10.00	10.32	5.32	10.32	11.90	10.24	10.00
Feature 2	7.30	10.48	12.30	7.06	8.49	13.49	8.57	11.43
Feature 3	6.75	8.10	10.00	5.87	10.48	14.37	7.14	10.00
Feature 4	7.70	9.52	9.52	6.67	8.65	11.59	8.57	7.14
Feature 5	8.57	11.11	12.22	8.02	8.89	11.75	10.00	7.14
Feature 6	9.52	12.78	12.86	7.06	9.84	13.25	8.57	10.00
Feature 7	9.05	11.43	11.90	8.57	11.35	13.81	8.57	11.43
Feature 8	8.10	10.00	12.78	8.10	10.16	11.90	10.00	11.43

3. Improvements Made through Multi-algorithmic Methods and Score Fusion

Table J-3 presents the EERs(%) for each of the 8 features described and developed in Appendix G which focused on various parts of the spectrum. MVN is exercised in all the experiments. The best two results are shown in gray color. In addition to re-recorded sentences the results for hand-trimmed retold and read subsets of IVIE are included.

Table J-4 presents the EERs(%) for various fusion techniques. Score fusion in each case is applied to the scores obtained from 8 sources (8 set of features). Rule seven for fusion engages features 1, 3 and 4. The last row specifies the results based on BGI method where one third of the data in each case is used for estimation of PDFs. Rule 5 in all the cases outperformed the best individual feature. The fifth rule outdoes the sum rule (rule 6) as well. Sum rule widely used in the literature is not as effective as the best suggested rules. The fusion results in some cases are better than those for the best individual features (G90, B30, B60 and B90).

When the signal is distorted, multi-feature fusion can be helpful. It is in contrast with the conditions under which signal has completely different characteristics for example in the case of Retold recordings or Read sentences. The score fusion techniques on the retold passages have not yielded favourable results. The best feature outperforms all of the fusion techniques. It strongly supports the hypothesis that the multi-algorithmic methods, in the way I have suggested here, are helpful when there is a mismatch between test and training data caused by channel distortion but the signal characteristics of the test recordings are not inherently different.

Table J-4 Verification results for 8 fusion techniques

EER (%)	G30	G60	G90	B30	B60	B90	Retold	Read
Rule 1	8.02	10.08	9.52	6.19	8.57	11.90	8.57	9.76
Rule 2	7.14	8.65	10.95	6.19	8.97	12.78	9.76	9.76
Rule 3	7.62	8.73	9.05	4.84	8.10	11.90	10.00	8.57
Rule 4	6.27	8.10	9.44	5.24	8.02	12.22	8.57	7.38
Rule 5	5.40	8.10	9.13	3.81	5.16	8.10	8.57	10.00
Rule 6	6.27	8.10	9.44	4.76	7.06	11.43	8.57	8.81
Rule 7	5.24	7.22	7.06	4.68	6.19	10.48	8.57	7.38
BGI	7.14	9.17	8.57	4.40	7.02	10.36	10.79	11.15
Best Feat.	6.75	8.1	9.52	5.32	8.65	11.59	7.14	7.14

In Table J-5 the relative improvement made by rule 5, rule 7 and fusion techniques over the best verification results with cepstral features (UN41-MVN: features extracted through uniform filters with MVN) in 6 cases is specified. The improvement is significant especially for microphone B and for smaller distances where the recorded signal is strong.

Table J-5 Relative improvement made by rule 5 and 7 and BGI-based fusion over the best individual feature

	Relative improvement (%) over best results (uniform filters with MVN)					
	G30	G60	G90	B30	B60	B90
Rule 5	33.33	23.30	7.96	59.98	54.54	35.41
Rule 7	35.31	31.63	28.83	50.84	45.46	16.43
BGI	11.85	13.16	13.61	53.78	38.15	17.38

4. Error Rates at the System's Baseline Thresholds: Vitality of Threshold Re-adjustment

The error rates reported in the previous sections were representatives of the system's performance and were good indicators of the strength of the suggested algorithms. Nonetheless such error rates are not the ones that the system working at its baseline EER point should put up with. The true FRR obtained at the verification time is decided by the previously set thresholds and is normally higher. In accordance with the requirements of the framework and results reported in chapter 6 the error rates at the EER threshold and at the threshold of the point of high security is reported in this chapter.

Table J-6 specifies the error rates (FRR and FAR rates) at the thresholds of EER, and the FRR at the secure point threshold (threshold at which under normal conditions FRR=3*FAR).

Table J-6 Verification errors at pre-set EER and Secure point thresholds

	MF29-MVN			UNI41-MVN		
Error (%)	FRR@EER	FAR@EER	FRR@SP	FRR@EER	FAR@EER	FRR@SP
G30	38.57	1.27	49.05	15.24	4.92	20.48
G60	43.33	1.43	51.90	27.62	3.81	32.38
G90	54.76	1.11	61.43	21.43	3.97	30.00
B30	25.71	0.48	34.76	19.52	4.60	27.14
B60	35.24	1.43	46.67	22.38	5.08	27.62
B90	48.57	1.43	61.90	27.14	4.44	32.86

The results show that adjusting the thresholds based on the type of channel distortion is vital. In other words if the thresholds are not re-set the errors are intolerable.

The uniform features demonstrate a better performance and are less affected by distortion.

It is noteworthy that the figures in Table J-6 give an estimate of how much the imposture by re-playing a recording (or a manipulated recording) could be successful. Since the system under attack does not expect a channel distortion no threshold re-adjustment would be undertaken by the system. This is an example of how setting the threshold as well as applying compensation methods could offer convenience and at the same time inflicts higher security risks on the system.

In the last set of experiments reported under the same title, the errors are reported for the best fusion algorithm.

Figure J-4 shows that multi-algorithmic fusion is helpful for reducing the effect of curve shift. In the evaluation test illustrated in this figure, the system using multi-algorithmic fusion (rule 5) is exposed to the normal test subset as well as the recordings in the TG30 dataset. The FRR at the EER threshold of the baseline system for TG30 is around 11.4% which is drastically lower than the errors reported in Table J-6 (38.57% and 15.24%) suggesting that the fusion reduces the curves' shift caused by mismatched conditions.

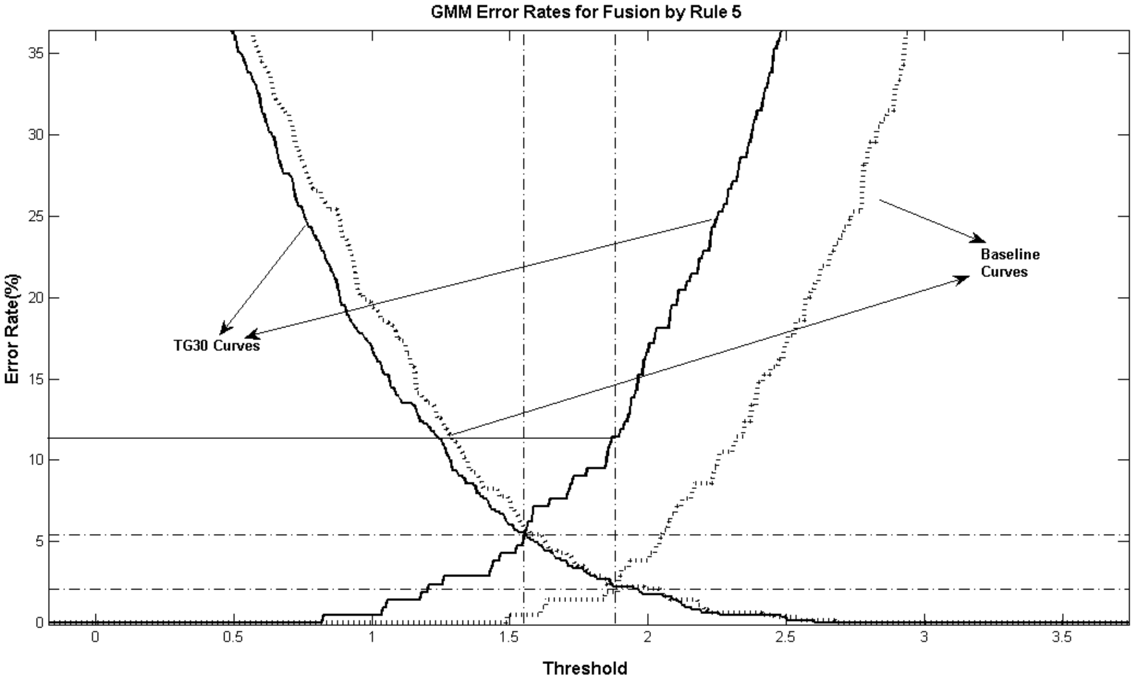


Figure J-4 Effect of fusion by rule 5 on the thresholds for normal test subset and TG30 subset

Appendix K: Noise Experiments

K.1 Purpose and Description of Experiments on Noise

While there is a rich legacy of research on the solutions to noise problem the experiments carried out here shed light on some of the unclear problems as expanded on here:

1. For comparison with other factors (style, channel, distance and coding) the prototype system is tested with various noise profiles:

Three types of uncorrelated Gaussian noises are added to our test subset (Subset-B). These subsets are labeled as *TnSmm* which *Tn* specifies the noise type and *mm* denotes SNR[1] in dB. The noise types are *T1* (narrow-band noise Butterworth filtered with band-pass of 622Hz-1352Hz following the work of Damper and Higgins (2003)), *T2* (white noise) and *T3* (low pass filtered noise by low-pass Butterworth filter of 2000Hz). Signal to noise ratios are 10, 15 and 20dB.

In addition to Gaussian noises, the interference effect is tested by adding speech from hand-trimmed Retold subset to the test sentences. The care was taken to ensure that the speech from the same speakers wasn't present in the mixture. In each test utterance speech from only two speakers was mixed. A stereo recording was made with the first channel containing the mixed signal from two speakers in the described fashion and at the desired SNR and the second channel containing the same utterances but with the signal from the other speaker (interfered) with half the channel one 's amplitude. This is necessary to make sure that two channels are not equal and the mixing matrix is invertible (necessary for independent component analysis).

Various features were tested with simple compensation methods such as CMN and median filtering to demonstrate the effect of the noises on the baseline system.

2. In order to alleviate the noise effect, sub-band filtering and score fusion with different methods is carried out on different types of noises to determine whether they can be used as solutions to the problem of additive noise. As explained before the prediction was that the drastic improvement reported by Damper and Higgins will not be attained for white noise. The score

[1] Signal to noise ratio

fusion techniques, various noise profiles and features used here enhance the previous studies. This approach didn't provide solution to white noise problem.

3. White noise effect was reduced by discarding the low energy frames of speech and adding artificial white-noise signal to the training data during model training.

4. In addition to voice verification on one channel for simultaneous speech, independent component analysis was performed on two channels. The experiments were repeated for three set-ups: when two channels are just linear mixtures of two sources, when the mixed signals pass through two substantially different channels, and when they pass through the same channels. The discussion on the implications of results will be offered after the results are reported.

It should be admitted that we will not be able to compare and analyse the effect of all the previously introduced techniques on various noises in this work. The goal here is providing a solution for each type of noise in the ways that bear some novelty.

K.2 Description of the Techniques Used for Noise Reduction

A. Subband Filtering and Score Fusion

In subband analysis the same Mel-cepstral features were extracted from the signal after applying the filter to the signal. The score fusion techniques were the same as those described before (Appendix J).

Figure K-1 displays the frequency response of the filters applied to the noise in type 1 (*T1*) and type 3 (*T3*) noises as well as three bands used in sub-band analysis. Type 3 noise is low-pass filtered noise, and type 1 noise is band-pass filtered noise.

B. Discarding Low Energy Frames

Based on the mask approximation (described in Appendix I) the value of the log-amplitude of two added signals is decided by the dominant signal. For low energy frames the distortion caused by noise is higher and the feature value is largely affected by the noise value rather than the speech. Therefore exclusion of frames which are highly contaminated by noise is one technique to reduce the noise effect.

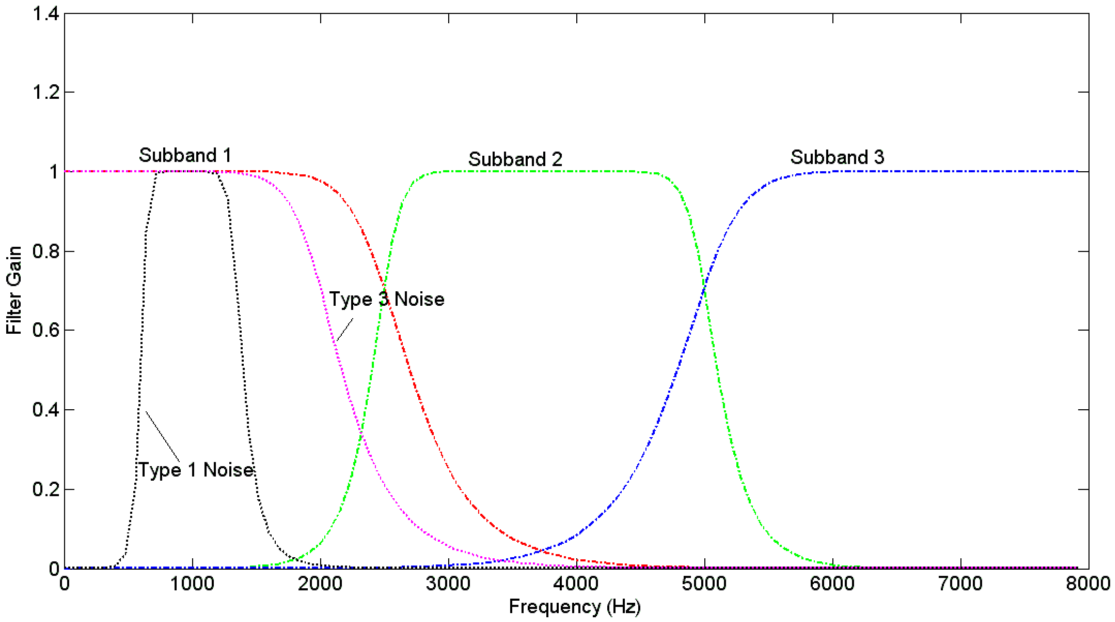


Figure K-1 Subband filters and noise profiles

There is a subtle difference between the effect of low frequency noise and white noises on speech signal. To illustrate this, three phones with low (f in feel), medium (m in mellow) and high (a: in are) energy are chosen and their frequency components are shown in the Figure K-2 and K-3 (left hand). The right hand plots display the log-amplitude of the 41 Mel-filters. In Figure K-2 the noise is white noise at 10dB (T2S10). In Figure K-3 the noise is narrow-band (T1S10). For narrow band noise it is only the low energy phone that is notably affected by the noise. On the contrary in white noise, since the higher Mel-filters are broader, the difference for clean and noised signal is considerable.

Discarding low energy frames is a solution for low-frequency noises which have similar spectral profiles to speech.

In the experiments the low energy frames (with energy under a portion of mean energy of frames of the whole utterance) are discarded. Many other criteria based on the max energy and energy of signal during silent parts (largely determined by noise) were tried having similar results.

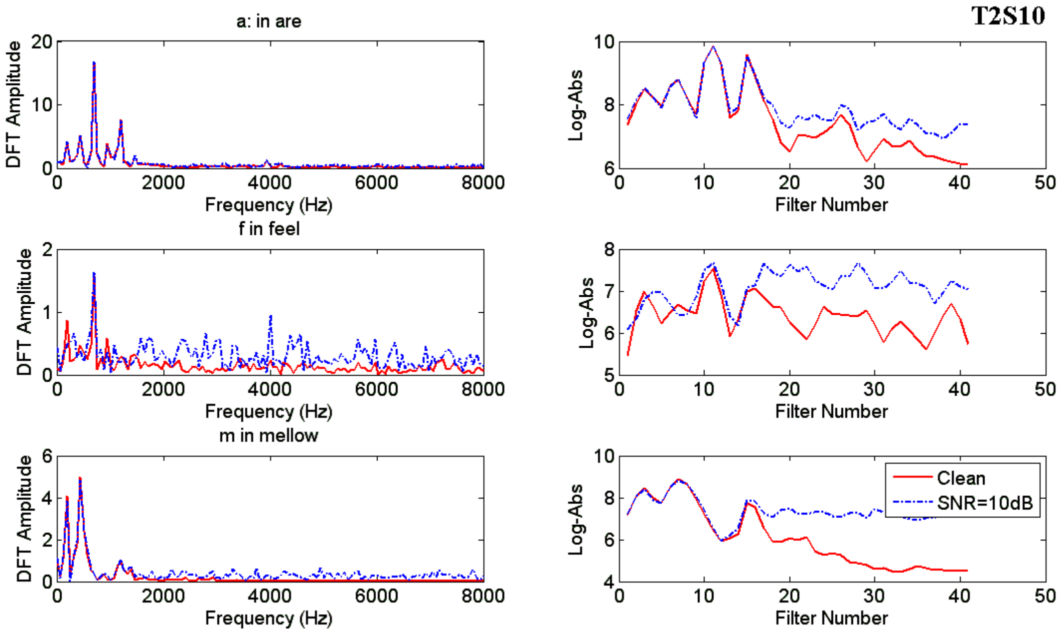


Figure K-2 Effect of contamination by white noise on three vowels (white noise)

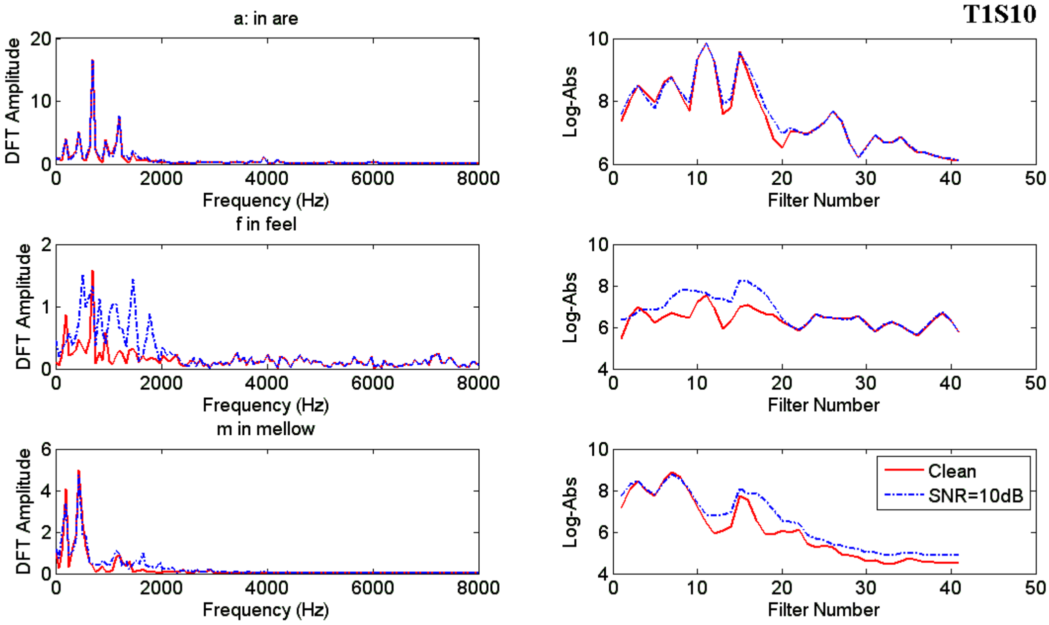


Figure K-3 Effect of contamination by narrow-band noise on three vowels (narrow-band noise)

C. Training Models on Noise-added Signal

As demonstrated in part B, based on the masking assumption the white noise is most harmful to speaker verification (compared to other types of noise) since it has high energy in the spectral areas which speech's spectral components are weak (high frequencies). White noise could be added to training data to simulate this effect at the training time. In the experiments, in addition to features extracted from clean training speech, parts of training data with the energies surpassing a threshold, was chosen for each speaker and white noise at SNR of 15 was added to the signal. Use of high energy parts of signal, prevents models to approach the pure noise profile.

D. Independent component Analysis

Independent component analysis (ICA) is used here for separation of simultaneous speech. ICA as a blind source separation technique is a powerful tool when the mixed signals do not have Gaussian distributions and are mixed in a linear fashion. While a more comprehensive description of ICA could be found in many texts such as (Hyvarinen & Oja, 2000) and (Kamsika, 2003) here a short account of the procedure is provided.

Assuming that: $X = A.S$ or $S = W.X$

where elements of *S* are s_{ij} *j*-th source signal at time *i*. The source signals represented by rows of *S* are assumed to be independent at each time. *X* is the output mixture and *A* is the mixing matrix. Each row of the output mixture (recorded by a sensor) is called mixture component and is denoted by *x* here. The key to separation in ICA approach is non-Gaussianity. Based on central limit theorem the sum of independent random variables tends toward a Gaussian distribution.

Two pre-processing steps of ICA are:

1. Centring: in which the mean of each mixture component is subtracted from its values.

2. Whitening: Through principal component analysis and use of Eigenvectors the components of mixture are uncorrelated and the variances are transformed to unit values (the covariance of the transformed points after whitening will be the identity matrix).

To quantify the amount of 'Gaussianity' many metrics such as Kurtosis, Negentropy, Mutual Information and Infomax have been proposed (Kamsika, 2003). In fast-ICA an approximation of Negentropy in calculated by:

Equation K-1

$$J(y) = [E\{G(x)\} - E\{G(v)\}]^2$$

where *G* is any non-quadratic function, *x* is any of the mixture components and *v* is the Normal Gaussian variable.

In each step of fast-ICA the elements of *W* matrix are updated. The algorithm can be summarised as follows: For each row of *W*, which is labeled *w* here and each row (component) of *X* shown as *x*:

1. Initial random values are assigned to the elements of w.
2. $w_{new} = E\{x.g(w'x)\} - E\{g'(w'x)\}.w$
3. $w_{new} = w_{new} / \|w_{new}\|$
4. If w_{new} is not in the same direction of w then: $w = w_{new}$ and go to step 2.

Where *g* is the derivative of *G* and g' is the derivative of *g*. The full demonstration of why these steps minimise Gaussianity based on Negentropy metric is given in (Hyvarinen & Oja, 2000). For independent component analysis the FastICA toolbox was used in the experiments[1].

K.3 Experimental Results

Experiment Set 1. Verification Results for Various Features

The result of speaker verification by cepstral features with 41 uniformly distributed filters and mean and variance normalization is shown in Table K-1. For three types of noises and for three SNRs (10dB, 15dB and 20dB) the EERs are specified.

Table K-1 Verification errors for various noise contamination profiles and SNRs

Model : UN41-MVN		**EER (%)**		
	Noise Type (All Gaussian Noises)	**SNR=10**	**SNR=15**	**SNR=20**
T1	Narrowband, Butterworth Filtered (622Hz-1352Hz)	**18.97**	**15.79**	**12.78**
T2	White Noise	**40.08**	**34.76**	**27.78**
T3	Butterworth Low-pass Filtered (2000 Hz)	**26.67**	**17.62**	**13.25**

In Table K-2 the ERRs after using various features with or without MVN are presented. Two instant suggestions of the results are that: first MVN is generally helpful and except in one case (for low pass filtered noise, T3 and MF29) has reduced the error and that in these cases the Mel-banks work as well or better than the uniform filter-banks. The combination of opposing factors such as filters' spectral focus, filters width, use of nonlinear log function and DCT which breaks down the effect of one channel on all the cepstral coefficients makes the prediction about the

[1] Helsinki University of Technology (2005), 'The FastICA package for MATLAB', Last Retrieved 2010-07-15, <http://www.cis.hut.fi/projects/ica/fastica/>.

performance of one feature in face of a noise profile difficult. The errors are very high and show that the features are vulnerable to noises even at low SNRs.

The experiments with features 1 to 8 (based on features focusing on different spectral areas in previous chapters) did not result in better performance that MF29-MVN in table K-2 and therefore are not reported here[1].

Table K-2 Verification errors for various noise contamination profiles and features

EERs (%)	MF29-NO	UN41-NO	MF29-MVN	MF41-MVN	UN41-MVN
T1S15	**24.92**	**25.40**	**16.27**	**19.05**	**15.79**
T2S15	**33.49**	**40.95**	**22.38**	**23.73**	**34.76**
T3S15	**15.71**	**18.57**	**23.41**	**23.89**	**17.62**

Experiment Set 2. Score Fusion Results

In this section the results of fusion of scores from multiple features is reported. Three features and corresponding models used are (Gaussian mixture size was 32):

- Feature 1: 24 Mel-cepstral features with derivatives (48 features)
- Feature 2: 24 Mel-cepstral features without derivatives
- Feature 3: 16 Mel-cepstral features with derivatives (32 features)

First the effectiveness of removing low energy frames is shows in Table K-3 and K-4. In both experiments the cepstral mean normalization is not applied to the features. For the first table the frames with lower than 0.2 of the mean energy of frames[2] in each utterance were discarded.

Table K-3 EER for various noises when frames with low energy were discarded

EERs (%)	T1S10	T1S15	T1S20	T2S10	T2S15	T2S20	T3S10	T3S15	T3S20
Feature 1	**8.65**	**5.71**	**3.41**	**28.57**	**19.05**	**10.63**	**7.22**	**4.29**	**2.30**
Feature 2	**13.33**	**8.65**	**4.76**	**26.19**	**19.92**	**12.30**	**11.90**	**7.14**	**2.86**
Feature 3	**12.38**	**7.14**	**4.29**	**34.84**	**25.79**	**13.33**	**10.48**	**5.71**	**3.33**

[1] To reduce the effect of noise three averaging techniques were also tested (some in a variety of parameters) which caused no or only a little improvement:

1. Use of median filters (non-linear) in time domain both in training time by training new models and at the test time.
2. Use of a smearing averaging filter both for training and testing data on time samples.
3. Use of median filter at the test time on frequency components (on the theoretic ground that frequency components of speech vary slower than noise) at the test time with the ordinary training models.

[2] The errors were not sensitive to change of this threshold in several cases experimented. For a range of values and based on different types of setting energy threshold (mean, max and start of utterance) the results were similar to those reported in the table.

Table K-4 EER for various noises when no frames were discarded

EERs (%)	T1S10	T1S15	T1S20	T2S10	T2S15	T2S20	T3S10	T3S15	T3S20
Feature 1	20.48	16.19	11.90	31.03	23.41	17.62	16.19	10.95	6.67
Feature 2	25.40	18.57	12.86	30.95	25.71	17.14	21.59	13.33	8.57
Feature 3	25.71	19.52	14.84	39.60	31.03	21.43	23.41	14.84	9.44

Discarding low energy frames has shown helpful in all the cases but has offered better improvement in T1 and T3 cases which is inline with the discussion presented in K.2 (B).

Further improvement could be made for T1 and T2 noises by applying mean and variance normalization as presented in Table K-5. MVN is more effective for white band noise and when the error is high otherwise (and conforming to the masking theory) if discarding noisy signal gives good results, the regular models outperform MVN normalised models.

Table K-5 EERs when low energy frames were discarded and MVN was applied

EERs (%)	T1S10	T1S15	T1S20	T2S10	T2S15	T2S20	T3S10	T3S15	T3S20
Feature 1	12.38	9.13	8.10	18.89	12.38	9.52	20.00	15.24	10.00
Feature 2	16.19	12.86	11.11	22.62	17.70	14.37	26.19	18.25	12.86
Feature 3	14.84	13.25	10.48	22.46	18.49	11.43	25.79	17.22	13.25

Finally Table K-6 presents the errors of score fusion by multiple methods[1]. The following results also confirm that Rule 5 for score fusion yields the best results in the majority of cases.

Table K-6 EER for Score Fusion when low energy frames were discarded, MVN applied

EER (%)	T1S10	T1S15	T1S20	T2S10	T2S15	T2S20	T3S10	T3S15	T3S20
Rule 1	17.06	13.81	11.43	22.70	17.22	13.33	26.67	18.97	14.29
Rule 2	14.76	10.48	9.05	20.48	12.86	9.05	23.41	17.62	11.90
Rule 3	15.71	11.51	10.08	20.95	14.76	9.05	25.79	16.59	12.38
Rule 4	15.71	11.98	10.00	21.43	14.21	9.68	25.63	18.10	11.83
Rule 5	11.43	8.57	7.22	18.73	12.94	8.65	22.38	15.16	10.00
Rule 6	15.71	11.98	10.00	21.43	14.21	9.68	25.63	18.10	11.90
Rule 7	14.76	11.43	9.52	20.00	14.84	10.48	24.84	17.14	12.78
Fusion	12.98	10.12	9.76	18.93	14.29	10.00	21.67	15.83	10.60

[1] In all the experiments if there are more than 3 features rule 7 engages 1st, 3rd and 4th feature, otherwise it engages first two features.

Experiment Set 3. Results of Subband Analysis

Results of subband analysis are reported in this part. In addition to sub-band features one full band feature (24 coefficients plus derivatives) is used. In these experiments the low energy frames are removed. To assess the effect of this removal on clean speech two other columns are included in the result table. 'Clean' column displays the performance of the features for clean test data. RMV specifies this performance when the low energy frames are removed for clean speech. **The results demonstrate that removing low-energy frames-even for clean speech-does not have a significant negative effect.**

The sub-bands which are less affected by noise have exhibited a closer performance to their performance for clean speech (compare Figure K-1 with Table K-7).

Table K-7 EER for full-band and sub-band features in clean speech and noises

EERs (%)	T1S10	T1S15	T2S10	T2S15	T3S10	T3S15	RMV	Clean
24+DRV	8.65	5.71	28.57	19.05	7.22	4.29	0.95	0.95
Subband 1	22.38	16.59	19.52	11.83	21.35	13.81	1.90	1.83
Subband 2	10.48	6.67	46.75	41.98	20.48	12.30	6.11	6.75
Subband 3	10.00	10.40	46.19	44.76	15.71	11.83	10.48	11.35

Table K-8 Results of score fusion based on sub-band and full-band feature scores

EER (%)	T1S10	T1S15	T2S10	T2S15	T3S10	T3S15	RMV	Clean
Rule 1	8.57	8.10	46.59	44.37	18.97	12.86	7.14	8.97
Rule 2	15.79	10.95	30.16	22.54	11.51	6.19	0.95	0.95
Rule 3	15.71	10.56	31.43	22.94	11.43	6.59	2.38	2.38
Rule 4	14.68	10.00	30.95	22.46	11.98	6.19	1.90	1.83
Rule 5	8.57	6.19	34.37	26.19	11.35	6.27	1.03	1.43
Rule 6	15.79	9.60	30.95	21.90	12.38	6.11	1.90	1.51
Rule 7	8.65	5.24	38.10	32.86	12.86	7.14	2.38	2.38
BGI	9.40	7.14	27.38	19.76	7.86	5.00	2.14	1.43

For T1 and T3 not only the individual features have better performance, but the combination of features has shown a promising performance. In the majority of the cases however the single best feature (full-band) has had better results than combination. This experiment is another example that demonstrates how fusion with a weak (unreliable) source of information could degrade the overall performance.

Experiment Set 4. Adding White Noise to the Training Data

To reduce the mismatch between the training and test conditions white noise is added at SNR of 15dB to the training signal. The tests are performed for SNR of 10 and 20dB which conforms to the assumption that we do not have a reliable estimate of the noise power in real applications. Table K-9 presents the EER error rates for various noise profiles. The results shows that subsequent to detection of noise profile (which could be done based on the silent parts of speech) the models compatible with this noise contamination profile (here white noise) could be employed for verification rather than normal models trained on clean speech. Nevertheless the specified errors also demonstrate that if noise profile is detected wrongly this initiative could degrade the verification accuracy (negative relative improvement percentages).

Table K-9 Effect of training models on noise-added signals on error rates

EER(%)	Noise Added Models	Normal Models	Relative Improvement (%)
Clean	**1.83**	**1.51**	-21.06
T1S10	**18.10**	**18.97**	4.60
T2S10	**7.22**	**34.37**	**78.98**
T3S10	**14.68**	**13.25**	-10.78
T1S20	**6.27**	**6.19**	-1.28
T2S20	**5.79**	**15.87**	**63.50**
T3S20	**3.81**	**3.25**	-17.07

Experiment Set 5. Simultaneous Speech Separation

Table K-10 displays the results of six sets of experiments on simultaneous speech (interference) separation. The noisy set is labelled *T4* here. The 6 sets of experiments include:

1. Experiments on simultaneous speech at the target SNR (or signal to interference ratio, SIR) on the first channel without separation by independent component analysis (Ch. 1). These are in fact the baseline verification errors for interference (without any compensation).

2. Experiments with simultaneous speech on two channels when two signals are mixed linearly. ICA is applied to the stereo recordings when each channel is a linear mixture of two signals. For the first channel the SIR is the target SIR but for the seconds channel the interfered signal amplitude is approximately half its value for the channel one.

3. Experiments with simultaneous speech on channel one when the signals on two channels have passed through two different linear filters each representing the channel affect of their corresponding channel (Ch. 1 Tr.).

4. Same as 3 but ICA is applied to two channels for separation[1] (Ch. Tr., ICA).

The transform function of two linear channels is displayed in Figure K-4.

5. Experiments with channel one when low energy frames are removed (under 0.2 of mean energy of signal) (Ch.1, R0.2)

6. Experiments with channel one when the same signal has undergone the channel distortion with removal of low energy frames (Ch.1, Tr. R0.2).

16 cepstral coefficients and mixture sizes of 32 was used in the experiments. The non-linearity function used for ICA was: $g(x) = x.e^{-x^2/2}$

Table K-10 Error rates for speaker recognition in simultaneous speech

EER (%)	Ch.1	ICA	Ch.1, Tr.	Ch.1 Tr., ICA	Ch.1, R0.2	Ch.1, Tr. R0.2
T4S10	10.48	1.43	13.81	11.51	7.14	10.00
T4S20	4.68	1.43	7.14	8.97	1.43	3.89

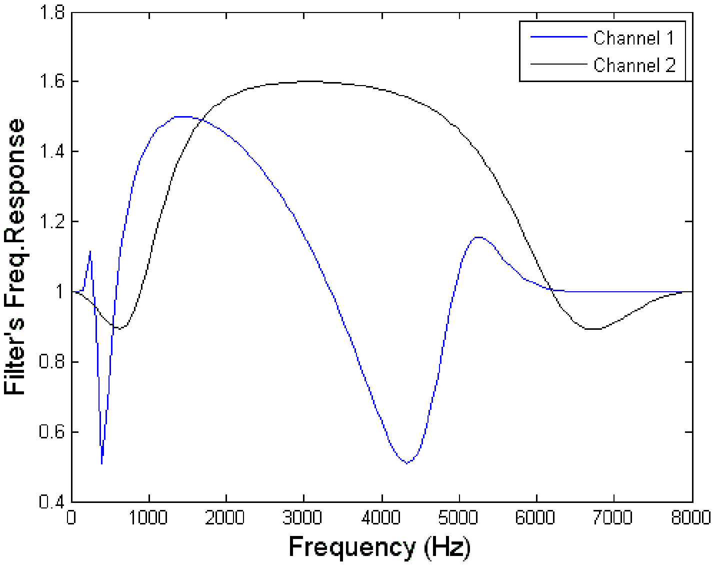


Figure K-4 Channels' spectral profile for stereo recordings

The error rates in column one indicate that the interference effect is less significant than noise effect[1] especially in the case of white noise.

[1] In ICA the order of signals can not be determined after separation and the overall power of the separated signals are similar therefore a challenge is to determine which of the separated signals is the target speech itself. Since we know that the dominant signal has been the speech from true speaker in these experiments the signal with higher mixture elements in the matrix is selected as the true speaker's speech. If there is no dominant speech finding the true speaker's signal will be an issue.

ICA proves to be extremely successful in separating signals when they are mixed linearly through a mixture matrix.

When the mixed signals have passed through different channels however the results (column 4) show that the ICA is not successful. In practical applications if two sensors are not very close due to the significance of ‘distance distortion’ even if the microphones are the same, the mixed signals pass through different channels and the practical results are similar to the ones specified in column 4. Therefore other algorithms should be used for solving this problem. The pursuing of this objective however goes beyond remit of this research.

The inference error rates are low compared to other types of noise and eliminating low energy frames improves the verification results especially for high SIRs.

Appendix L: Voice Signature

L.1 Electronic Signature

A voice signature scheme (which satisfies the requirements of electronic signature) is proposed and analysed in this section. As a prerequisite, the function and requirements of electronic signature is reviewed here with a look over Uniform Electronic Transaction Act (UETA) (US, 1999) and Woodward remarks on biometric signature (2003).

Woodward (2003) states that signature serves several functions including evidentiary, cautionary, approval and efficiency. With regard to biometric signature he explains that to satisfy the electronic signature requirements the parties should produce a document explaining the transaction and that their biometrics (for example their fingerprint templates) are appended to the document. Woodward enumerates the requirements of biometric signature and states that such a signature should be unique to the person, capable of being verified, under the sole control of the person using it and attached or logically associated with the document created for.

From a legal perspective according to Uniform Electronic Transaction Act (UETA) "Electronic signature means an electronic sound, symbol, or process attached to or logically associated with a record and executed or adopted by a person with the intent to sign the record.”. UETA states that “Another important aspect of this definition lies in the necessity that the electronic signature be linked or logically associated with the record”.

A design for voice signatures is presented here which shows the advantage of voice over other biometrics in satisfying two requirements of being in sole control of the person and being logically associated with the document.

L.2 Design of the Proposed Voice Signature

A BCA with a pair of asymmetric keys generates a text-independent voice model for *Person_A* which we name *Model_A*. BCA makes a certificate (using its private key) for *Person_A* with this content: “*Person_A* with public key of *Pub_A* has the voice model : *Model_A*” (This certificate is called *Cert_A* hereafter). Everyone can use BCA’s public key and decrypt *Cert_A* and verify that *Model_A* is associated with *Person_A* by BCA and *Person_A* has public key of *Pub_A*.

When *Person_A* needs to sign a document, he reads the whole or important parts of the document, along with the document's time and date. The electronic signature attached to the document is this message:

"I *Person_A* with *Cert_A*, accept the contents of this document (*D*) at time and date *X*."[1]

The above message is encrypted by private key of *Person_A*.

The verifier (any party in need of verification of the signature) decrypts the above electronic signature using *Person_A*'s public key, then extracts *Cert_A*. Using *Cert_A* which is signed by the BCA, the verifier extracts *Model_A* and verifies that the read message matches the model parameters (e.g. using the GMM based text-independent voice verification methods).

Use of the voice signature according to this scheme, has a great advantage over other types of biometric signature. This advantage relates to the fact that the signature allows a perfect linkage with the contents of the document. Except for those identifiers which could be used as a modality of communication-such as handwriting-other biometric records and thereby the signatures built upon them are normally independent of the document for which they are created.

It also shows a great advantage over other biometrics in terms of being 'in sole possession of the owner' since unlike fingerprint, after being revealed to a party, can not be re-used and attached to any other documents.

[1] The person does not need to read out his certificate. He only needs to read *D* the contents of the document and *X* (time / date). The implication of the suggested notation is that the final message that is going to be encrypted contains: parts of the document (*D*), time and Date (*X*), *Person_A* information and *Cert_A* information.