Automatic Recognition of Ageing Speakers

A dissertation submitted to the University of Dublin
for the degree of Doctor of Philosophy

Finnian Kelly
Trinity College Dublin, April 2014

Signal Processing and Media Applications
Department of Electronic and Electrical Engineering
Trinity College Dublin

For Grandad

Abstract

The process of ageing causes changes to the voice over time. There have been significant research efforts in the automatic speaker recognition community towards improving performance in the presence of everyday variability. The influence of long-term variability, due to *vocal ageing*, has received only marginal attention however. In this Thesis, the impact of vocal ageing on speaker verification and forensic speaker recognition is assessed, and novel methods are proposed to counteract its effect.

The Trinity College Dublin Speaker Ageing (TCDSA) database, compiled for this study, is first introduced. Containing 26 speakers, with recordings spanning an age difference of between 28 and 58 years per speaker, it is the largest longitudinal speech database in the public domain. A Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification experiment demonstrates a progressive decline in the scores of genuine-speakers as the age difference between training and testing increases. The scores of imposters, over the same period, are relatively stable. Consequently, verification error increases with age difference.

A novel *stacked classifier* approach, exploiting an ageing-dependent decision threshold is introduced, significantly reducing verification error rates at large age differences. A new model-based quality measure, Wnorm, is incorporated into the stacked classifier framework alongside ageing information, resulting in a further reduction in the baseline error.

A second novel approach, *eigenageing compensation*, operates by determining the dominant directions of change in the models of ageing speakers, and using this information to compensate for an age difference between training and testing samples. Eigenageing compensation results in a relative reduction in baseline error at large age differences that compares favourably to the stacked classifier approach. A by-product of the eigenageing compensation method is shown to enable a promising new approach to automatic age estimation.

Vocal ageing is of particular relevance to the forensic domain. An evaluation of five ageing Irish males is presented in a forensic automatic speaker recognition framework. Vocal ageing is shown to significantly weaken strength-of-evidence estimates, leading to cases of erroneous support for the different-speaker hypothesis within 10 years. Eigenageing compensation is shown to be suitable for the forensic domain, and is effective at reducing the impact of ageing. A listener test demonstrates that vocal ageing is detected with increasing accuracy as age difference increases, and is significantly more detectable in the voices of females than males.

The effect of ageing on the performance of an i-vector speaker verification framework is evaluated. Compared to a GMM-UBM approach, lower absolute error rates are achieved. As ageing progresses however, the performance of both systems degrades at the same rate, demonstrating that current inter-session compensation approaches are not sufficient for dealing with ageing variability. This demands that specific strategies, such as the proposed stacked classifier and eigenageing compensation methods, be adopted.

Declaration

I hereby declare that this thesis has not been submitted as an exercise for a degree at this or any other University and that it is entirely my own work.

I agree that the Library may lend or copy this thesis upon request.

Signed,

Finnian Kelly

April 7, 2014.

Acknowledgments

Firstly, a huge thank you to Dr. Naomi Harte for her endless encouragement, guidance and good humour over the last four years.

Thanks to all the characters of Sigmedia, who make it friendly and light-hearted place to be. Special thanks to Róisín for helping me to stay sane, laughing, and full of coffee during the final weeks, and to Andrew for his nuggets of wisdom over the years. Sincere thanks to Dr. Andrzej Drygajlo for guiding me through six months of research at EPFL. Thanks also to Dr. Rahim Saeidi and Prof. David van Leeuwen for hosting me at Radboud University Nijmegen for an enjoyable few weeks. Thanks to Dr. Niko Brümmer for his helpful advice. To the Irish Research Council, your financial support made this work possible, and was greatly appreciated.

Lastly, to Mum and Dad, Olwyn, Ailbhe, Diarmaid and Cormac - thanks for everything.

Contents

List of Acronyms

BS	Between-Source
BBC	British Broadcasting Corporation
CMVN	Cepstral Mean and Variance Normalization
DCF	Decision Cost Function
DET	Detection Error Trade-off
DFT	Discrete Fourier Transform
EER	Equal Error Rate
EM	Expectation Maximisation
FA	Factor Analysis
FAR	False Acceptance Rate
FASR	Forensic Automatic Speaker Recognition
FRR	False Rejection Rate
FFT	Fast Fourier transform
GMM	Gaussian Mixture Model
GMM-UBM	Gaussian Mixture Model-Universal Background Model
GMR	Gaussian Mixture Regression
GSV	Gaussian SuperVector
HTER	Half Total Error Rate
HR	Hit Rate
JFA	Joint Factor Analysis
KDE	Kernel Density Estimation

LDA Linear Discriminant Analysis

LLR Log-Likelihood Ratio

LR Likelihood Ratio

LPCC Linear Predictive Cepstral Coefficient

MAE Mean Absolute Error

MAP Maximum A Posteriori

MFCC Mel-Frequency Cepstral Coefficient

ML Maximum Likelihood

NAP Nuisance Attribute Projection

NHR Noise-to-Harmonic Ratio

NIST National Institute for Standards in Technology

PCA Principal Component Analysis

PDF Probability Density Function

PLDA Probabilistic Linear Discriminant Analysis

PLP Perceptual Linear Prediction

RASTA RelAtive SpecTrAl

RBF Radial Basis Function

ROC Receiver Operating Characteristic

RP Received Pronunciation

RTÉ Raidió Teilifís Éireann

SII Speech Intelligibility Index

SNR Signal-to-Noise Ratio

SRE Speaker Recognition Evaluation

SV SuperVector

SVM Support Vector Machine

SVR	Support Vector Regression
TCDSA	Trinity College Dublin Speaker Ageing
TCDSA-FD	Trinity College Dublin Speaker Ageing - Forensic Development
UBM	Universal Background Model
UBML	Universal Background Model Log-likelihood
VAD	Voice activity detection
VQ	Vector Quantisation
VOT	Voice Onset Time
WCCN	Within-Class Covariance Normalization
WDP	Within-source Degradation Prediction
WMVL	Within-source Minimum Variance Limiting
WS	Within-Source

1 Introduction

Human speech is a rich signal. Embedded in its linguistic message is information as to the gender, age, health, emotional state and *identity* of the speaker. In recent decades, there has been rapid development in technology to automatically recognise individuals from their voices. Speaking is a natural, non-invasive activity, that produces an easily-transmittable signal. These attributes have led *automatic speaker recognition* to be increasingly deployed in real-world applications such as authentication over the telephone[1] and smart human-computer interfaces[2].

Forensic speaker recognition has a much longer history; an early case of the voice being used to identify a suspect heard (but not seen) committing a crime was the case of William Hulet in 1660, accused of having executed King Charles I. A witness testified that he knew the masked executioner was Hulet, "by his speech". After being sentenced to death however, Hulet was subsequently pardoned when the true perpetrator came forward and confessed [59]. This early case of mistaken identity highlights the problem with subjective evidence in forensics. The application of automatic methods to forensic speaker recognition is gaining increasing support, for this reason [75].

The case of William Hulet also highlights the constant challenge present in speaker recognition. The same richness of speech that enables it to encode identity, also allows emotion to be conveyed, or the speaker's state of health to be revealed. The issue of ever-present variability

[1] e.g. Nuance Communications announced in May 2013 that their FreeSpeech text-independent voice biometrics solution is to be used by Barclays Wealth & Investment Management to verify customer identities over the phone: `http://www.planetbiometrics.com/article-details/i/1580`

[2] e.g. The Xbox One games console, to be launched in November 2013, incorporates automatic speaker recognition for personalisation of the user-experience: `http://www.xbox.com/en-US/xbox-one/get-the-facts#6-6`

within a speaker's voice is compounded in automatic speaker recognition, where the effects of environmental noise and the transmission channel are superimposed on the signal. Thus, much of the recent research focus in speaker recognition has been on dealing with the impact of these multiple sources of everyday variability. Alongside regular variability, the process of ageing leads to change in the voice over the long-term. The most rapid (and therefore noticeable) changes to the voice occur in childhood, adolescence and old age. However, ageing drives constant, if more subtle, vocal change throughout adulthood. The physiological causes of vocal ageing and the resulting acoustic effects have been well investigated. The effect of ageing on automatic speaker recognition however, has received only marginal attention.

Biometric technology is becoming more pervasive in everyday life. Given an increase in the average human lifespan globally, and the increasingly frequent contact between individuals and biometric systems, the impact of ageing is inevitable [123, 159]. In addition to the context of biometrics, vocal ageing is of particular relevance to forensic speaker recognition. The nature of the forensic scenario dictates that the speech samples under comparison are always non-contemporary. The duration of the time-lapse between samples can, and does, stretch into years [63]. It is therefore important to establish the consequences of ageing for forensic comparison, especially in the case of forensic automatic speaker recognition.

In this Thesis, the impact of vocal ageing on automatic speaker recognition, from biometric and forensic perspectives, is assessed. The resulting observations prompt the proposal of new strategies for dealing with this challenging variability.

1.1 Thesis outline

The remainder of this Thesis is organised as follows.

Chapter 2: An overview of automatic speaker recognition

This Chapter provides background for the automatic speaker recognition systems used in subsequent Chapters. The main components of a speaker recognition system, comprised of feature extraction, pre- and post-processing, speaker modelling and decision making, with a focus on topics of relevance to this Thesis, are introduced. Some of the recent advances in speaker recognition are highlighted. The domain of forensic speaker recognition is introduced, and the use of an automatic speaker recognition system to provide evidence suitable for the court is outlined.

Chapter 3: Vocal Ageing

This Chapter sets the context for the experimental studies in this Thesis by introducing the subject of vocal ageing, i.e. the changes in the voice that occur due to ageing. The physiological changes and associated acoustic effects accompanying vocal ageing are reviewed. The compilation and content of the Trinity College Dublin Speaker Ageing (TCDSA) database is

described. An experimental evaluation of the acoustic correlates of ageing is presented using the TCDSA subjects. A Gaussian Mixture Model-Universal Background Model (GMM-UBM) system is utilised to explore the effect of an increasing age difference between training and testing samples on the verification scores of the TCDSA subjects, and on the overall classification error. The insight gained prompts the development of strategies to compensate for ageing variability in speaker verification.

Chapter 4: Stacked Classification for ageing speaker verification

This Chapter presents the first contribution towards ageing variability compensation in speaker verification. Observing the behaviour of genuine-speaker and imposter GMM-UBM verification scores, an ageing-dependent decision threshold is proposed. The proposal is implemented in a stacked classifier framework, and its performance is evaluated. Owing to the non-ageing-related variability present in the TCDSA database, the concept of quality in speaker verification is introduced. A new quality measure, Wnorm, along with several established quality measures, is evaluated for its ability to predict the utility of scores in a speaker verification evaluation. A decision threshold incorporating measures of quality is subsequently implemented in the stacked classifier framework, and its performance is evaluated on the TCDSA database. The framework is then expanded to consider both ageing and quality information in determining a decision threshold. The relative influences of both ageing and quality factors in the TCDSA database are consequently observed.

Chapter 5: Eigenageing compensation for ageing speaker verification

This Chapter presents eigenageing compensation, the second contribution towards ageing variability compensation in speaker verification. Observations from the stacked classifier, which operates at the score-level, suggested that an approach to ageing variability compensation operating at the model-level may be advantageous. Eigenageing compensation uses information learned from the changes in the models of ageing speakers to compensate for ageing variability. Its performance is evaluated with respect to the stacked classifier. A new approach to age estimation, drawing from the eigenageing compensation technique, is then presented.

Chapter 6: Forensic speaker recognition and ageing

Vocal ageing is of particular relevance in the forensic domain. In this Chapter, the consequences of vocal ageing for forensic speaker recognition are investigated. The effect of vocal ageing on speech 'evidence' is established with a forensic automatic speaker recognition (FASR) system evaluation of the TCDSA males. A subsequent in-depth FASR evaluation of five Irish-accented ageing males provides a deeper insight. As forensic speaker recognition includes approaches based on subjective listening, the human perception of ageing change is of interest. The outcome

of a listener test is presented, showing the degree to which humans can detect the presence of ageing in the voice of the same speaker at different ages.

Chapter 7: Effect of ageing on i-vector speaker verification

Much of the recent research efforts in speaker recognition have focused on counteracting the negative effects that everyday 'inter-session' variability has on recognition accuracy. In this Chapter, the effect of ageing variability on the performance of a speaker verification system with a state-of-the-art inter-session variability compensation framework is evaluated. The performance of an i-vector system with probabilistic linear discriminant analysis (PLDA) modelling is compared to that of the GMM-UBM approach, given an increasing age range between training and testing samples.

Chapter 8: Conclusions

The final Chapter draws together the main conclusions of this Thesis, and suggests possibilities for future study.

1.2 Contributions of this thesis

This Thesis explores the effect of ageing on speaker recognition, and presents two new solutions to compensate for its negative effect on recognition accuracy. The contributions in each Chapter can be summarised as follows:

Chapter 3

Introduced the Trinity College Dublin Speaker Ageing (TCDSA) database - a new longitudinal speech database compiled for this study.

Presented the largest longitudinal analysis of the acoustic correlates of ageing to date.

Demonstrated the significant degradation in the performance of a Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system due to the effects of long-term ageing.

Chapter 4

Introduced a stacked classifier approach for ageing speaker verification that exploits an ageing-dependent decision threshold. Demonstrated this proposal to significantly reduce long-term classification error.

Proposed a model-based quality measure, Wnorm, for speaker verification, and demonstrated that its ability to predict the verification scores of genuine-speakers could reduce classification error.

Demonstrated that combining quality and ageing information in the stacked classifier framework reduces long-term classification error to a greater extent than either factor alone.

Chapter 5

Proposed eigenageing compensation for speaker verification, which learns information about the changes in the models of ageing speakers to compensate for ageing variability. Demonstrated that eigenageing compensation significantly reduces long-term classification error, and compares favourably to the stacked classifier approach.

Proposed a new method of automatic age estimation using an ageing model learned as part of the eigenageing compensation procedure.

Chapter 6

Demonstrated that vocal ageing significantly undermines forensic automatic speaker recognition (FASR) by progressively weakening strength-of-evidence estimates.

Applied eigenageing compensation to FASR, and demonstrated both its suitability for the domain and its ability to reduce the negative impact of ageing.

Presented the outcomes of a listener test, and showed that vocal ageing becomes progressively more detectable with age difference and is significantly more detectable in female voices.

Chapter 7

Compared the performance of i-vector and GMM-UBM speaker verification systems in the presence of ageing. Demonstrated that the performance of both systems degrades at the same rate as ageing progresses. Concluded that dealing with ageing variability demands specific compensation strategies in addition to standard inter-session compensation approaches.

1.3 Publications

Portions of the work described in this Thesis have appeared in the following publications:

Journal article

F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space". *Computer Speech & Language*, 27(5):10681084, 2013.

Book Chapter

F. Kelly and N. Harte, "The impact of ageing on speech-based biometric systems", In *Age Factors in Biometric Processing*, pp 171-184, IET, 2013.

Conference papers

F. Kelly and N. Harte, "Effects of Long-Term Ageing on Speaker Verification". In *Biometrics and ID Management*, vol. 6583, Lecture Notes in Computer Science, pp 113124. Springer Berlin/Heidelberg, 2011.

F. Kelly, A. Drygajlo, and N. Harte, "Speaker Verification with Long-Term Ageing Data". In *International Conference on Biometrics (ICB)*, New Delhi, India, 2012.

F. Kelly, A. Drygajlo, and N. Harte, "Compensating for Ageing and Quality variation in Speaker Verification". In *InterSpeech*, Portland, Oregon, 2012.

F. Kelly, N. Brümmer, and N. Harte, "Eigenageing Compensation for Speaker Verification". In *InterSpeech*, Lyon, France, 2013.

F. Kelly and N. Harte, "Auditory detectability of vocal ageing and its effect on forensic automatic speaker recognition". In *InterSpeech*, Lyon, France, 2013.

Presentations

Oral presentation at "Irish Signals and Systems Conference (ISSC)", Cork, Ireland, 2010.

Oral & poster presentations at "International Summer School on Biometrics for Secure Authentication", Alghero, Italy, 2010.

Poster presentation at "Workshop on Innovation and Applications in Speech Technology (IAST)", Dublin, Ireland, 2012.

Poster presentation at "BBfor2 Short Summer School in Forensic Evidence Evaluation and Validation", Madrid, Spain, 2012.

Oral presentation at "6th European Academy of Forensic Science Conference (EAFS)", The Hague, The Netherlands, 2012.

Oral presentation at "European Association of Biometrics Research and Industry awards", Darmstadt, Germany , 2013.

Publications under review

F. Kelly and N. Harte, "Eigenageing compensation for long-term speaker verification". *IEEE Transactions on Audio, Speech and Language Processing*, submitted for review in March '14.

F. Kelly, R. Saeidi, N. Harte and D. v. Leeuwen, "Effect of long-term ageing on i-vector speaker verification". *InterSpeech 2014*, submitted for review in March '14.

F. Kelly and N. Harte, "Forensic comparison of ageing voices from automatic and auditory perspectives". *The International Journal of Speech, Language and the Law*, to be submitted for review in April '14.

2

An overview of automatic speaker recognition

Speaker recognition is the task of identifying an individual from a sample of their speech. Within speaker recognition, a distinction can be made between speaker verification and speaker identification. Speaker verification is the task of using a sample of an individual's speech to accept or reject a claim as to their identity. Speaker identification, on the other hand, is the task of using a sample of an individual's speech to identify them from a set of possible candidates. In speaker identification, the possibility of the speaker not being in the list of candidates usually has to be entertained.

Regardless of the task, the key components are the same. The process has two distinct stages; enrolment and recognition. At enrolment, a user identifies themselves and provides a speech sample. At recognition, a user provides a speech sample and receives a decision from the system.

A schematic of a speaker recognition system is depicted in Figure 2.1 [18, 32, 118]. The first operation of both enrolment and recognition stages is 'feature extraction', in which the speech is transformed into a more compact representation that emphasises speaker-dependent properties of the signal. In the enrolment mode, a statistical model is generated given the extracted speech features and stored in a database of speaker models. Considering a speaker verification task, in the recognition mode, the speaker model corresponding to the identity claim is retrieved from the database and compared to a set of extracted speech features in a 'pattern matching' operation. Based on the outcome of this comparison, a 'decision logic' module responds with an accept or reject outcome. In a speaker identification scenario, at the pattern matching phase, the extracted speech features are compared to all speaker models in the database. The decision logic module,

will output which, if any, of enrolled speakers are the source of the extracted features. Both verification and identification modes may have a 'no-decision' output, if for example, the speech signal is of insufficient quality to reach a decision. In the application of speaker recognition to forensics, the use of a 'hard' decision is unsuitable. Thus, the output of the pattern matching operation is interpreted directly.

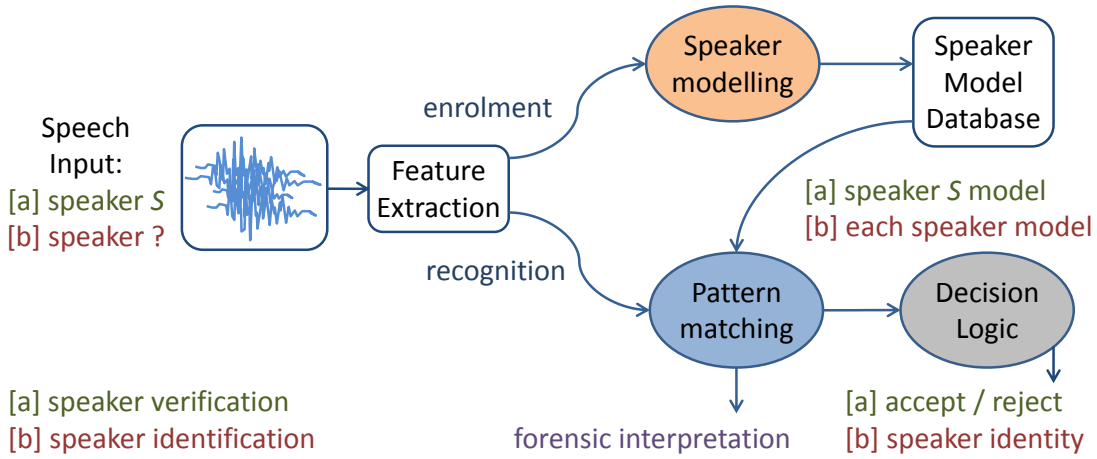

Figure 2.1: A schematic of the main components of an automatic speaker recognition system. Speaker verification and speaker identification modes are indicated by the labels [a] and [b] respectively.

A major challenge faced by all speaker recognition systems is dealing with variability between recordings at the enrolment and recognition stages [118]. There are multiple potential sources of variability, including acoustic environment changes (e.g. background noise, echo), transmission channel differences (e.g. mobile phone, landline, VOIP), within-speaker changes (e.g. health, emotion, ageing), and differences in recording duration and phonetic content. Variability between recordings due to any combination of these factors can be termed *inter-session variability* [114].

2.1 Feature Extraction

Speech is a highly complex signal, containing multiple streams of information. The aim of feature extraction is to reduce the dimensionality of the signal by suppressing or ignoring unwanted information and retaining and enhancing speaker-specific information. Unwanted information, in this instance, is any attribute of the of signal that is independent of the speaker, such as the recording environment or transmission channel. Speaker-specific information is an attribute of the signal that remains consistent in the speech of a given speaker and aids in discriminating their speech from the speech of others. As set out in [172, 199], the ideal feature would:

- have large inter-speaker variability and small intra-speaker variability, i.e. it should provide maximum discrimination between speakers while minimising variability within speakers.
- occur frequently and naturally in speech
- be easy to extract from the signal
- be difficult to mimic
- have robustness to noise and distortion
- have robustness to intra-speaker vocal change due to health or ageing
- be manageable in size, avoiding the so-called *curse of dimensionality* [102].

In practice, it is unlikely that a single feature would fulfil this full list of requirements. As noted by Kinnunen [118], there is no globally "best" feature, and a trade-off must be made between speaker discrimination, robustness and practicality. However, the complexity of the speech signal can be taken advantage of by extracting multiple complimentary features and combining them to improve the speaker discrimination of the system [161]. There have been many features proposed for speaker recognition, broadly categorizable based on the duration of speech required for their extraction and on the 'level' of the information they capture.

As speech is produced, the vocal organs are constantly in motion. Due to the physical limits of the vocal tract, it can be assumed that the speech signal is stationary over 20-30 ms segments. Short-term spectral features are based on the spectrum of speech over 20-30 ms frames, therefore acting as descriptors of the resonance properties of the vocal tract. Thus, a direct link between short-term spectral features and physical characteristics, e.g. vocal tract dimensions, can be drawn. Mel-frequency cepstral coefficients (MFCCs) [45], linear predictive cepstral coefficients(LPCCs) [96] and perceptual linear prediction (PLP) coefficients [90] are examples of frequently-used short-term spectral features.

Prosodic features, extracted over windows of speech in the range of 100s of ms, capture information like pitch, rhythm, energy and speaking rate [11,161,182]. Thus, these features relate to a mix of physical characteristics, e.g. gender and age, and learned or acquired 'behavioural' factors, like speaking style or health issues.

'High-level' features, extracted over a duration ranging from seconds to minutes, capture exclusively 'behavioural' attributes, including phonetic-level information, such as accent, and word-level information such as semantics and idiolect - an individual's personal lexicon [161,183].

Short-term features are an advantage from the perspective that are easy to extract and do not place a minimum speech duration requirement on the system. However, they are generally affected by noise and other sources of mismatch between enrolment and recognition conditions in a significant way. Longer-term features, while requiring a lot more speech, and likely being more computationally expensive to extract, are generally robust to noise and mismatch [118]. Therein lies the trade-off to be made in feature extraction. In this Thesis, short-term spectral features, MFCCs specifically, have been used exclusively.

2.1.1 Mel-fequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCCs) [45], have been used successfully as features for many speech processing applications, including speech, emotion and language recognition, along with present task of speaker recognition.

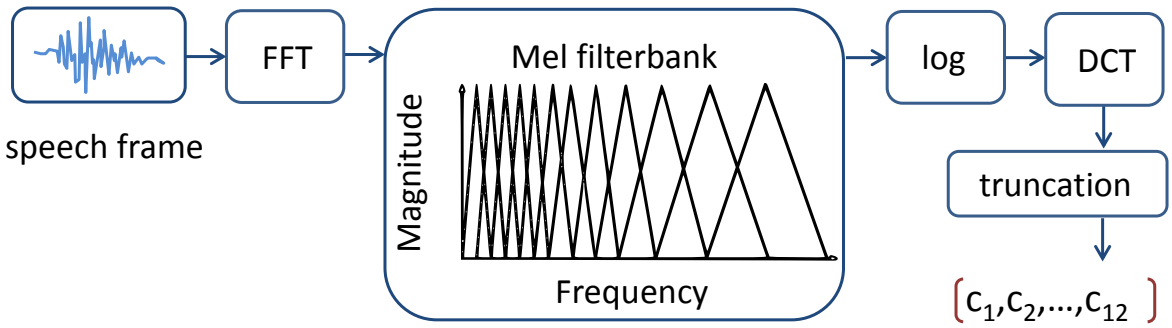

Figure 2.2: The stages of MFCC extraction

The process of MFCC extraction is illustrated in Figure 2.2. Given a speech frame, a window function, usually Hamming, is applied. The fast Fourier transform (FFT), an efficient implementation of the discrete Fourier transform (DFT), then decomposes the speech frame into its frequency components. The resulting magnitude spectrum, the *spectral envelope* of the speech signal, is a descriptor of the resonance properties of the vocal tract, which carry speaker-specific information.

A set of bandpass filters, with bandwidths and spacing determined by the perceptually-motivated *Mel* scale, are then applied. This filtering gives a stronger weighting and higher resolution to the lower frequencies in the signal.

A logarithmic compression is then applied, followed by a decorrelation with the discrete cosine transform (DCT). Defining the outputs of an M-channel Mel filterbank as $Y(m)$, where $m = 1, ..., M$, the MFCC coefficients are obtained as follows:

$$c_n = \sum_{m=1}^{M} \left[\log\left(Y(m)\right)\right] \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \tag{2.1}$$

The final MFCC feature vector is obtained by retaining the first 10-20 coefficients. Coefficient c_0 represents the energy of the frame and is often not included in the feature vector.

2.1.2 Pre-processing and Post-processing

Prior to MFCC extraction, a standard pre-processing step is *pre-emphasis* [118]. This is applied to boost higher frequencies in the speech signal that would otherwise have low intensities as a

result of the downward sloping spectrum of the glottal voice source. Pre-emphasis is applied with a high-pass filter, where α is generally taken in the interval [0.95, 0.98] [18]:

$$H(z) = 1 - \alpha z^{-1} \tag{2.2}$$

Voice activity detection (VAD) is an important step in processing a speech signal for speaker recognition. Since (MFCC) features are chosen based on their ability to extract discriminant information from speech, it is not likely that the non-speech regions of the signal will be informative. Sufficiently accurate discrimination of speech and non-speech regions can be achieved with an energy-based detector [118], which discards speech frames if their energy falls below a threshold based on the overall signal energy. Other approaches use a Gaussian model of the frame energy [139] or signal periodicity [89].

MFCCs are sensitive to environmental noise and channel effects. Several post-processing steps are typically applied to MFCC features to minimise variability due to this sensitivity. A simple approach to reducing the effect of channel influence is feature normalization [65]. In the cepstral (log spectral) domain, convolutive channel effects become additive. Therefore, subtracting from each frame the mean value over all frames in the sample normalizes the features with respect to the channel. In addition, the feature variance can be normalized by dividing each feature by its standard deviation. When these normalizations are applied to MFCCs, the operation is referred to as Cepstral Mean and Variance Normalization (CMVN),

The assumption in CMVM is that channel effects are stationary. RelAtive SpecTrAl (RASTA) filtering [91] applies a bandpass filter in the cepstral domain along the temporal trajectory of each feature. It suppresses modulation frequencies which are outside the normal rate of change of a speech signal. Thus, it can suppress, for example, the influence of slowly varying convolutive channel noise.

In addition to these typical post-processing operations, MFCC features are usually augmented with temporal information. This is achieved by calculating the first and second order derivatives, referred to as (Δ) and double-delta (Δ^2), between adjacent feature vectors and appending them to the original feature vector. Thus, given an MFCC feature with 12-coefficients, appending Δ and Δ^2 yields a 36-dimensional feature vector. The derivatives can be computed as time differences between adjacent vectors, or by fitting a regression line [156]:

$$d_t = \frac{\sum_{\theta=1}^{D} \theta \left(c_{t+\theta} - c_{t-\theta} \right)}{2 \sum_{\theta=1}^{D} \theta^2} \tag{2.3}$$

Where c_t is the original coefficient at time t and D is the window length over which to calculate the derivatives. A value for D in the range of 2-4 is typically used. The boundaries are dealt with by replicating adjacent frames or zero padding.

While there have been many other post-processing schemes proposed for speaker recognition, e.g. speech enhancement, feature warping and feature mapping [118], the procedures elaborated on in this Section have been those applied in the speaker recognition implementation in this

Thesis. Figure 2.3 indicates the order in which these processing modules are applied in this implementation.

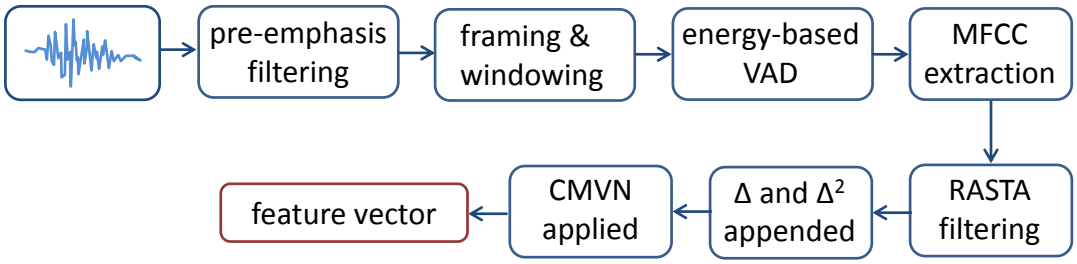

Figure 2.3: The speaker recognition system 'front-end'.

2.2 Speaker Modelling

At the enrolment stage of speaker recognition, a model is trained from a set of feature vectors. In text-independent speaker recognition, there is no correspondence between the speech content of enrolment and recognition utterances. Therefore, the model must be general enough to describe the typical 'feature space' of a speaker, but discriminative enough to distinguish between the feature spaces of different speakers. Modelling approaches for text-independent speaker recognition began with the non-parametric Vector Quantisation (VQ) approach in the 1980s [184]. VQ modelling maps a set of features to a 'codebook', by associating them with quantized, non-overlapping regions in the feature space. Comparison of reference and test speech samples is achieved via a distance measure (e.g. Euclidean) between their respective codebooks. The parametric Gaussian Mixture Model (GMM) [163], can be considered an extension of the VQ model, in which the clusters are overlapping, and for which there is a non-zero probability of a feature originating from each cluster. GMMs have been the dominant modelling approach in speaker recognition for some time: from the Gaussian Mixture Model - Universal Background Model (GMM-UBM) [163]; to factor analysis methods [29, 114]; and to the recent 'total variability' approach [48, 87].

2.2.1 Gaussian Mixture Modelling

A Gaussian Mixture Model (GMM) is a weighted sum of M multivariate Gaussian components, allowing it to model an arbitrary distribution of observations. The likelihood of an observation x given a GMM denoted by λ is:

$$p\left(x|\lambda\right) = \sum_{m=1}^{M} w_m p_m\left(x\right) \tag{2.4}$$

where x is a D-dimensional vector, w_m is the weight of the mth Gaussian component $p_m(x)$,

$$p_m(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_m|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_m)'\Sigma_m^{-1}(x-\mu_m)\right\} \tag{2.5}$$

μ_m and Σ_m are the mean vector and covariance matrix of the mth component respectively. The component weights $w_m > 0$ must satisfy $\Sigma_{m=1}^{M} w_m = 1$. In practice, for reasons of data requirement and computation, covariance matrices are usually diagonal [118].

Training a GMM involves finding the parameters $\lambda = \{w_m, \mu_m, \Sigma_m\}_{m=1}^{M}$ given a training sample $x = \{x_1, x_2, ...x_T\}$. The log-likelihood LL of x with respect to λ is given by:

$$LL = \frac{1}{T}\sum_{t=1}^{T} \log p(x_t|\lambda) \tag{2.6}$$

The higher the value of LL, the stronger the indication that x originates from the GMM λ. Maximum likelihood (ML) estimation is an approach to maximise the likelihood of the model with respect to given data, and can be achieved with the iterative Expectation Maximisation (EM) algorithm [17].

As an alternative to ML estimation, GMM parameters can be adapted from a previously trained model referred to as a Universal Background Model (UBM). The motivation for this is that the model is not estimated 'from scratch', and can take advantage of prior knowledge about speech data that is represented in the UBM. This is the principal behind the Gaussian Mixture Model - Universal Background Model (GMM-UBM) approach to speaker recognition [163].

In GMM-UBM speaker recognition, a UBM is trained with the EM algorithm using speech data from a large number of speakers. The UBM is therefore representative of speech 'in general'. When enrolling a new speaker into the system, the parameters of the UBM are adapted towards the speaker's feature vectors using *maximum a posteriori* (MAP) adaptation. Adapting only the UBM mean components has been shown to be effective in practice [163]:

$$\mu_m = \frac{n_m}{n_m + r} E_m(x) + \left(1 - \frac{n_m}{n_m + r}\right) \mu_m^{UBM} \tag{2.7}$$

where n_m is the probabilistic count of the feature vectors assigned to the mth component and r is a *relevance factor*, balancing the relative contributions of the UBM and the new data. $\tilde{x}_m$ is given by:

$$E_m(x) = \frac{1}{n_m}\sum_{t=1}^{T} P(m|x_t)\, x_t \tag{2.8}$$

where $P(m|x_t)$ is the posterior probability of the mth component given the sample x_t. The use of the UBM as prior information means that this approach can deal with limited quantities of data: the well-trained UBM parameters will fill in any 'holes' in the feature space (i.e. components with a low probabilistic count, $n_m \approx 0$) of the training data.

In the recognition mode, the 'match score' between a feature vector $X = \{x_1, x_2, ...x_T\}$ and a speaker model λ_s is given by the log-likelihood ratio (LLR):

$$LLR = \frac{1}{T}\sum_{t=1}^{T} \log p\left(x_t|\lambda_s\right) - \log p\left(x_t|\lambda_{UBM}\right) \tag{2.9}$$

Here, the UBM provides normalization by shifting the log-likelihood scores from different feature vectors into a comparable range. In addition to the normalizing effect of the UBM on the LLR score, a further level of normalization is generally applied. *Score normalization* is a procedure to reducing score variability across different speakers and sessions. In speaker verification, this improves the accuracy of the system given a common (speaker-independent) decision threshold.

A *Genuine-speaker* score is produced when a speaker model is tested with his/her own recording, and an *imposter* score is produced when a speaker model is tested a recording from a different speaker. In Zero-normalization (Z-norm) [162], a normalized score LLR_{norm} can be obtained from the 'raw' score LLR_{raw} as follows:

$$LLR_{norm} = \frac{LLR_{raw} - \mu_s}{\sigma_s} \tag{2.10}$$

where μ_s and σ_s are the mean and variance respectively of the imposter score distribution for a speaker s. The aim is thus to normalize the imposter score distribution for a given speaker to zero mean and unit variance. The Z-norm statistics μ_s and σ_s are speaker-dependent, and can be computed off-line (e.g. at the enrolment phase) by computing the LLR scores for a set of imposter speech samples given the speaker model. Another common normalization technique is Test-normalization (T-norm) [9]. T-norm is applied in the same manner as Z-norm, but with test-sample-dependent normalization statistics, which must be computed on-line (e.g. at the recognition phase) by computing the LLR scores for a test sample given a set of imposter models. Z-norm and T-norm can also be combined, to good effect [196].

A schematic of GMM-UBM speaker recognition, the system used for the majority of the experimental studies in this Thesis, is shown in Figure 2.4.

2.3 Decision Making

At the recognition stage, given a speech sample, the system outputs a decision. In the GMM-UBM framework, the decision is based on the LLR score of a test feature vector given a speaker model and the UBM. Considering the speaker verification scenario, the speaker model under comparison corresponds to the claimed identity of the speaker, and the decision is one of two possibilities: accept or reject. A predetermined threshold is compared to the LLR score to make this decision.

There are two types of error that can occur: a false acceptance (or false alarm) occurs when a genuine-speaker attempt is rejected. A false rejection (or miss) occurs when an imposter

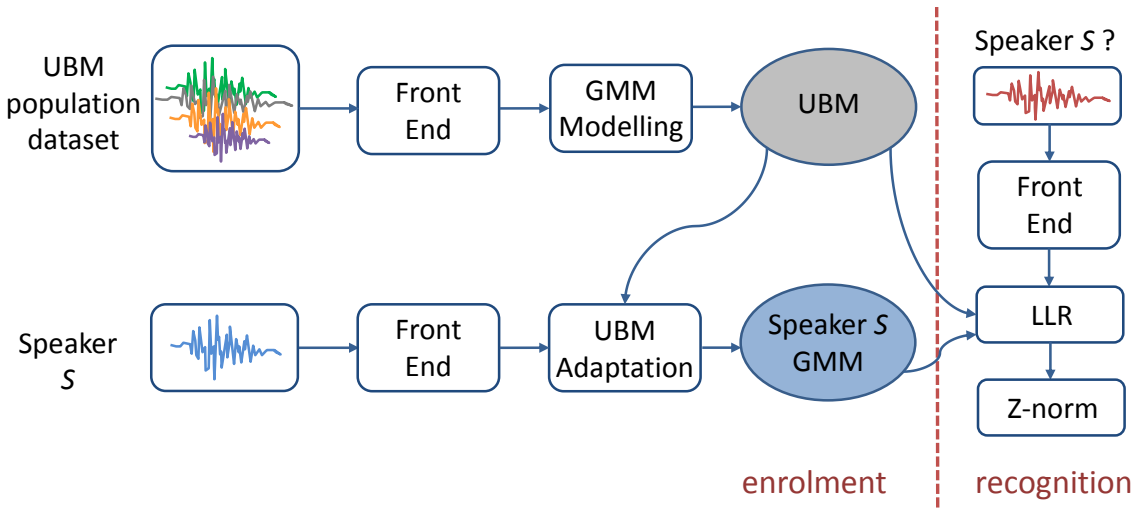

Figure 2.4: The components of a GMM-UBM speaker recognition system, in verification mode, from speech signal to normalized LLR. 'Front End' refers to the series of operations in Figure 2.3.

attempt is accepted. By counting these errors over a large number of trials, the false acceptance rate (FAR) and false rejection rate (FRR) of the system at a given decision threshold can be obtained. In a system evaluation, genuine-speaker and imposter attempts are often referred to as *target* and *non-target* trials respectively. Setting a decision threshold is a task of reaching a desired trade-off between the FAR and FRR. By sweeping the decision threshold over all possible values, the full trade-off between the FAR and the FRR can be visualised as a detection error trade-off (DET) curve (essentially a receiver operating characteristic (ROC) curve applied to speaker recognition). This provides a means of inspecting the behaviour of the system over all operating points. The performance of the system can be characterized by the equal error rate (EER), which is the point at which the FAR and the FRR are equal [18, 118]. An example of a DET curve with EER superimposed is given in Figure 2.5.

Typically, a development set of data is used to set a decision threshold such that a specific error criterion is minimised. One such error metric is the half-total error rate (HTER), which is the average of the FAR and the FRR. The regular National Institute for Standard in Technology (NIST) Speaker Recognition Evaluations (SREs) [51] specify a decision cost function (DCF) as a performance metric. The DCF applies a different weighting to each error type and also incorporates the prior probability of a genuine-speaker (or target) attempt.

When a decision threshold determined in this way is applied to test data, a measure of the actual 'operational' performance of the system is obtained. The difference between the development and testing error measurement represents the *calibration* of the system. The DCF (and HTER) metrics are application-dependent, in the sense that they represent the performance of the system at one operating point. An application-independent metric is the log-likelihood ratio cost function C_{llr} [28, 127]. This metric generalises the performance of the system over the

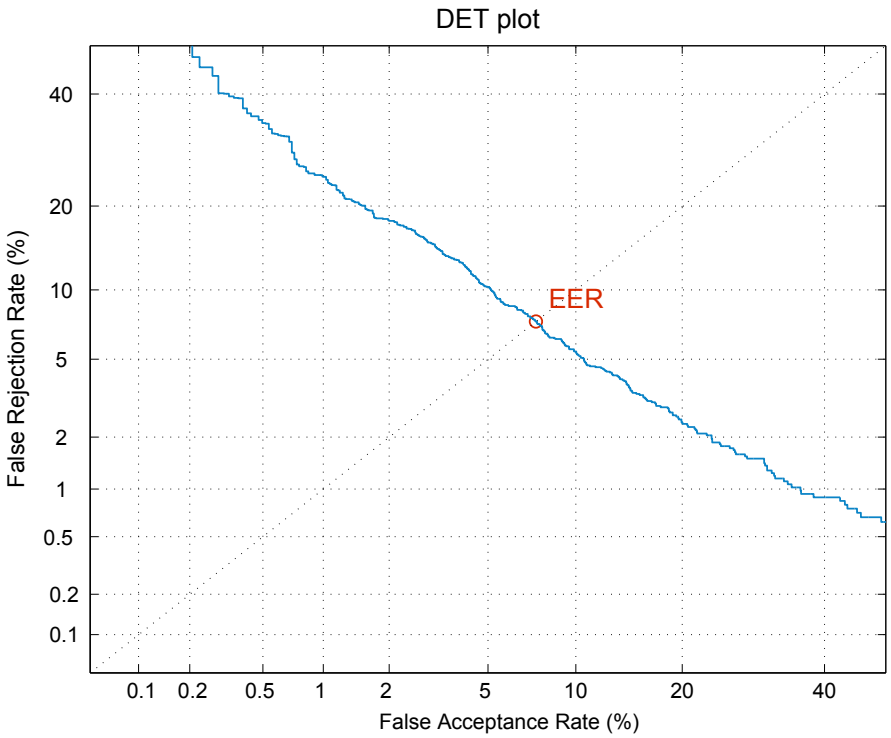

Figure 2.5: An example of a DET curve and the EER point.

whole DET curve, providing a measure of both discrimination and calibration.

2.4 A brief review of recent advances in speaker recognition

Since the introduction of the GMM-UBM speaker recognition framework by Reynolds et al. in 2000 [163], the major focus in the research community has been to build on this framework by compensating for inter-session variability, at the feature, model and score levels.

Advancement of the GMM-UBM framework began with the introduction of 'supervector' approaches [118]. In the context of GMM-UBM speaker recognition, a supervector refers to the stacking of the (D dimensional) mean components of a GMM into a single vector. Speaker verification with a discriminative support vector machine (SVM) classifier, using GMM supervectors as 'features', was introduced by Campbell et al. in 2006 [35]. Methods of inter-session compensation introduced for 'Gaussian supervector SVM' (GSV-SVM) were nuisance attribute projection (NAP) [36] and within-class covariance normalization (WCCN) [88]. NAP operates by finding an eigenchannel matrix, which is obtained by pooling the supervectors from a large number of speakers in varying conditions and subtracting the mean supervector. An eigen-analysis on the resulting supervector matrix yields the principal directions of inter-session variability in supervector space. This is then used to 'project away' the supervector dimensions responsible

for variability in a test sample. WCCN operates in a similar manner to NAP, but applies a weighting of supervector dimensions, rather than projection, to reduce inter-session variability.

A factor analysis (FA) approach to GMM-UBM speaker recognition was introduced in the form of eigenchannel compensation [25,29,116]. As in the case of NAP in GSV-SVM, eigenchannel compensation operates by finding an eigenchannel matrix. Eigenchannel compensation is applied by adapting (via either MAP or ML) a speaker GMM in the direction of a test sample, but constraining this adaptation to the subspace of supervector space specified by the eigenchannel matrix. Testing with the adapted GMM then proceeds within the GMM-UBM framework in the usual way.

Joint Factor Analysis (JFA) [112–114] extends this idea by compensating for both speaker and channel variability. JFA operates by considering a speaker utterance as an addition of speaker- and session-dependent (or channel-dependent) components, along with a speaker- and session-independent component (e.g. a UBM mean supervector). This approach is applied by first finding a session subspace (eigenchannel matrix) and speaker subspace (eigenvoice matrix) and then jointly estimating speaker and session factors given each new test utterance. Removing the channel component from the supervector representation results in a speaker-dependent supervector. A variety of methods for scoring have been proposed [70].

Recently, speaker recognition research has focused on the i-vector framework [48, 178]. As with JFA, the i-vector framework compensates for speaker and session variability. It does so however, with a single 'total variability' subspace, rather than separate subspaces for speaker and session variability. Given a total variability matrix and a speaker- and session-independent component, a total factor or i-vector is estimated for each utterance. The i-vector is a compact representation of an utterance in low-dimensional space. Inter-session compensation is typically applied in the i-vector space with WCCN [88] and linear discriminant analysis (LDA). For classification, the currently favoured approach is probabilistic linear discriminant analysis (PLDA) [154]. PLDA is a generative approach to modelling target and non-target i-vector distributions. The score computed with PLDA is in the form of a ratio between the likelihoods of the i-vector coming from the same and different speakers respectively.

A speaker verification system based on an i-vector framework with PLDA modelling is used for experiments in Chapter 7. For all previous experimental Chapters, namely 3, 4, 5 and 6, a GMM-UBM speaker recognition system, as depicted in Figure 2.4 is used. A discussion on the motivation for using a 'classic' GMM-UBM system for the experiments in this Thesis, as opposed to a 'state-of-the-art' approach, is provided in Chapter 7.

2.5 Forensic speaker recognition

Forensic science can be described as the application of scientific principles to investigate and establish evidence for presentation in a court of law [52]. Forensic speaker recognition is the task of determining whether two (or more) recordings of speech originate from the same speaker,

within a framework of the legal process [172].

In general, forensic speaker recognition involves the comparison of an 'evidential' and a 'suspect' speech sample. The evidential sample, also referred to as the questioned recording or trace, is typically obtained from a telephone recording or through police surveillance [146, 172]. The suspect sample, also referred to as the reference recording, is generally an excerpt from a police interview of the suspect [146, 172]. The role of the forensic expert, the individual tasked with comparing the speech samples, is to reach a conclusion that can be used as evidence in court [52]. In reaching this conclusion, two measures must be considered: the *similarity* of the features in the evidential and suspect recordings, and the *typicality* of such features in the wider (relevant) population.

2.5.1 Methods of forensic speaker recognition

There are varied approaches to the comparison of speech samples in forensic speaker recognition. While traditionally the task was the preserve of 'expert listeners', the use of computer-based approaches has become widespread in the domain [72].

Technical forensic speaker recognition [52, 173] can be defined as the expert application of a trained skill or technologically supported procedure to the recognition task. *Naïve* speaker recognition, on the other hand, refers to the everyday ability of people to discriminate between the voices of speakers that are not known to them. This ability becomes important in the case of an 'earwitness', an individual who heard but did not see the perpetrator of a crime, who may be called upon to identify the voice of the suspect from a voice line-up [59, 93]. Although it is technical forensic speaker recognition that is of primary interest in this Thesis, the topic of naïve speaker recognition will be revisited in Chapter 6.

Technical forensic speaker recognition can be further categorized: 'Auditory-perceptual' analysis [18, 52, 172], also referred to as 'auditory-phonetic' analysis [72] is essentially a careful listening of the speech undertaken by a trained phonetician. The perceived differences between certain speech features in the two samples are used to estimate the extent of the similarity between the two voices. The outcome can only be a subjective one, limiting its use in a courtroom setting.

'Acoustic-phonetic' analysis [72, 172, 173] is a computer-aided analysis of various speech features, including fundamental frequency, formant centre-frequencies and statistical measures of the speech signal. An objective (numerically-based) outcome of the strength of similarity between the two samples can be achieved with this approach. An auditory-perceptual procedure extracts features which are of interest in an acoustic-phonetic analysis. Thus the two approaches are complimentary and commonly used in tandem [72].

In forensic automatic speaker recognition (FASR), acoustic parameters of an unknown voice in the evidential recording are compared with a statistical model derived from the acoustic parameters of a suspect sample [52]. The application of automatic speaker recognition methods, i.e. those introduced in the preceding sections in this Chapter, to the forensic scenario, is well

established [6, 52, 53, 74, 76, 141, 143, 158, 172, 173].

2.5.2 Application of forensic speaker recognition

Technical forensic speaker recognition, FASR in particular, can be applied at different stages of the legal process. The general forensic *investigative mode* follows a process of generating a set of likely explanations, testing them with new observations, then eliminating and re-ranking the explanations [101]. In the context of forensic speaker recognition, an FASR system can be operated by police in an investigative mode by establishing a short-list of the most relevant sources amongst a group of suspected speakers known to them [140]. An automatic speaker recognition system, in identification mode, as in Figure 2.1, could be applied directly to this task, with a short-list established on the basis of the score obtained at the pattern matching stage.

The forensic *evaluative mode* is the process of arriving at an opinion of evidential weight given the case-specific hypotheses, which can be presented to the court within a framework of circumstances (i.e. taking prior information into account) [101]. A crucial issue on the application of FASR to the forensic evaluative mode is the way in which the strength of evidence should be expressed for presentation in court.

2.5.3 Representation of the strength of evidence

Accepting that any automatic speaker recognition system operates with non-zero error, then the output of a binary accept/reject decision is clearly unsuitable in the forensic domain [38]. An FASR analysis must *contribute* to the fact-finding role of the court, not usurp its position as a 'trier of fact' [174].

Advancement in other fields of forensics, such as fingerprint, fibre, DNA and toxicology analysis, where the strength of evidence estimates are delivered in a 'likelihood ratio' framework [5] has prompted similar developments in forensic speaker recognition. The LR framework is becoming increasingly accepted as the "only legally and logically valid" [174] and "logically and legally correct" [73] method for presenting speech evidence in court, which brings it in line with other modalities of forensic evidence [76]. The likelihood ratio model considers two competing hypotheses, H_0 and H_1:

- H_0 is the hypothesis that the suspect speaker is the source of the speech in the evidential recording.
- H_1 is the hypothesis that a speaker other than the suspect speaker is the source of the speech in the evidential recording.

These will be referred to as the same-speaker and different-speaker hypotheses respectively. By considering both hypotheses, the LR describes both the similarity and typicality of the evidence. The likelihood ratio (LR) can then be defined as [74]:

$$LR = \frac{P(E|H_0, I)}{P(E|H_1, I)} \tag{2.11}$$

Where $P(E|H_0, I)$ is the probability of the evidence given the same-speaker hypothesis and some background information I, e.g the gender, age and accent of the suspect, that may taken into account when creating their model. $P(E|H_1, I)$ is the probability of the evidence given the different-speaker hypothesis and background information.

In the context of a GMM-UBM system, the value of the evidence E is the score of the evidential speech sample $X = x_1, x_2, ..., x_t$ given the suspect speaker GMM λ_s and the UBM λ_{UBM}.

$$E = \log p\left(X|\lambda_s\right) - \log p\left(X|\lambda_{UBM}\right) \tag{2.12}$$

In the GMM-UBM framework, the estimation of the LR is a two-stage statistical approach [141]. The first stage is to estimate the distribution of scores in the cases where H_0 and H_1 are known to be true. The second stage is to evaluate the value of the evidence E given both of these distributions. Figure 2.6 illustrates the estimation of the LR.

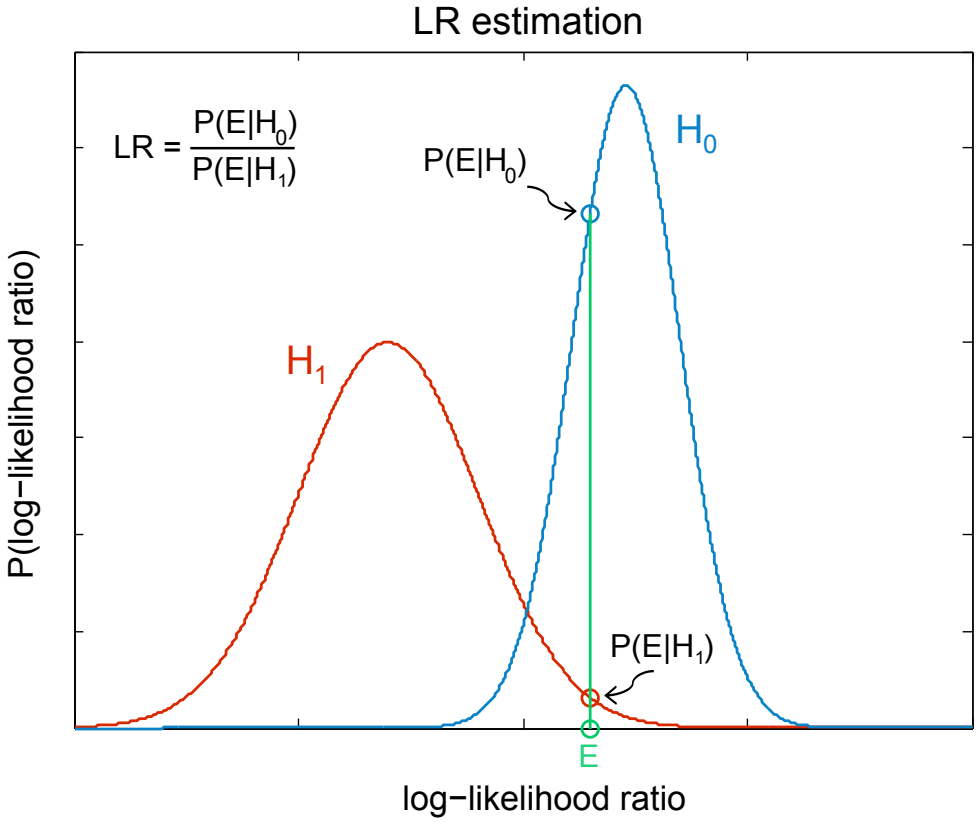

Figure 2.6: Estimation of the likelihood ratio (LR) given the value of the evidence E, the same-speaker hypothesis (H0) and the different-speaker hypothesis (H1).

LR	$\log_{10}$LR	Verbal equivalent
>10,000	5	very strong evidence for the same-speaker hypothesis
1,000-10,000	4	strong evidence for the same-speaker hypothesis
100-1,000	3	moderately strong evidence for the same-speaker hypothesis
10-100	2	moderate evidence for the same-speaker hypothesis
1-10	1	limited evidence for the same-speaker hypothesis
1-0.1	-1	limited evidence for the different-speaker hypothesis
0.1-0.01	-2	moderate evidence for the different-speaker hypothesis
0.01-0.001	-3	moderately strong evidence for the different-speaker hypothesis
0.001-0.0001	-4	strong evidence for the different-speaker hypothesis
<0.0001	-5	very strong evidence for the different-speaker hypothesis

Table 2.1: A verbal interpretation scale for the likelihood ratio (LR)

The odds form of Bayes' theorem shows how speech evidence, represented by the LR, can be combined with prior information (determined by the court) to arrive at posterior probabilities, which can be used to make judicial decisions [74, 172]:

$$\frac{P(H_0|E,I)}{P(H_1|E,I)} = \frac{P(E|H_0,I)}{P(E|H_1,I)} \cdot \frac{P(H_0,I)}{P(H_1,I)} \tag{2.13}$$

Thus, prior probabilities relating to the same speaker-hypothesis $P(H_0, I)$ and different-speaker hypothesis $P(H_1, I)$ are supplemented by new evidence provided by the forensic expert in the form of an LR, in a manner compatible with the judicial process. Note that while LR estimation was described in terms of the output of an FASR system here, the LR can be estimated from acoustic-phonetic features, such as fundamental frequency, in a similar manner [173].

2.5.4 Verbal LR scales

Since the numerical value of the LR may not be readily interpretable by the court, a verbal LR scale can be used to aid understanding. One such scale, proposed by [37] and used by the UK Forensic Science Service [172] is provided in Table 2.1.

There are issues with this proposal, around the relative interpretation of the words 'limited' and 'moderate' for example, and the 'cliff-edge' effect, where a small numerical LR difference can change the verbal translation. In addition, this scale does not take into account that the LR may not be directly presented in court, as ultimately it should be combined with prior probabilities. However, the scale provides a straightforward way to relate numerical results to a legal interpretation, and thus is used in the discussion of experimental results in Chapter 6 of this Thesis.

2.5.5 Implementing a GMM-UBM FASR system

Figure 2.7 illustrates the LR estimation procedure in the GMM-UBM framework, originally proposed by Meuwly et al. [141,142]. In addition to the suspect speaker and evidential recordings, a reference population database and suspect speaker control database are required.

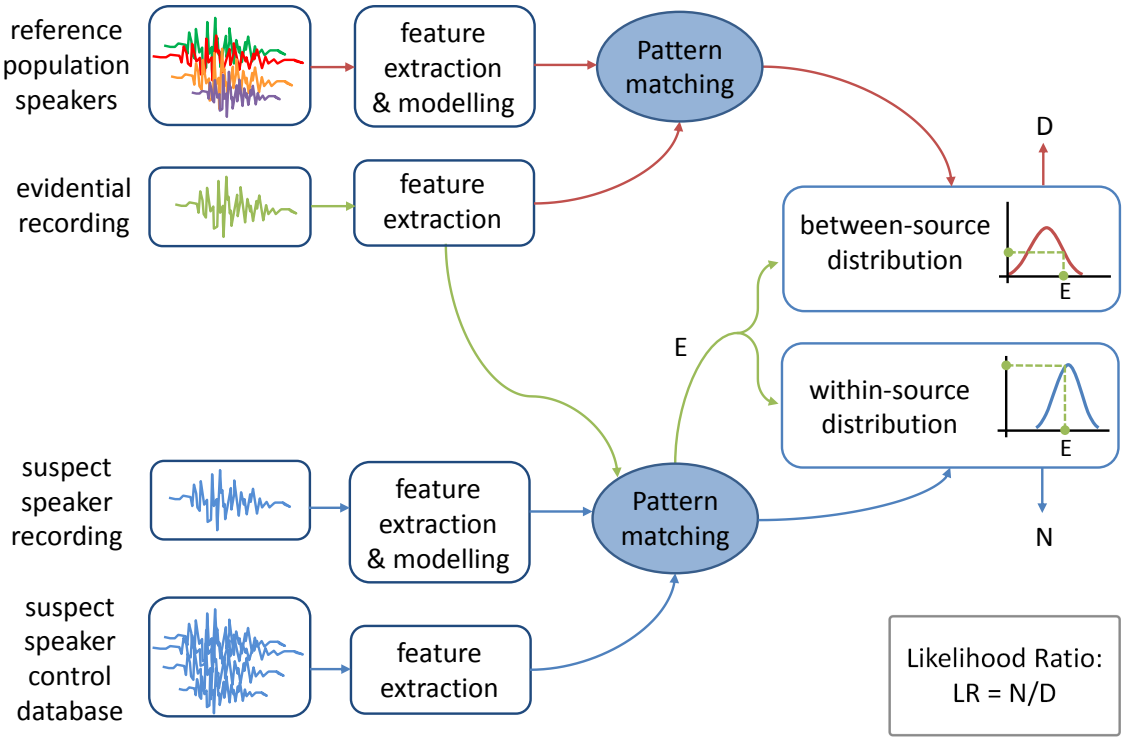

Figure 2.7: LR estimation given a suspect speaker recording, a suspect speaker control database, an evidential recording and a reference population database. The colour of the arrows correspond to the processing chain for each input.

- The H_0 distribution is estimated by modelling the within-source (WS) variability. A set of WS scores are obtained by testing a GMM trained from the suspect speaker recording with the contents of the suspect speaker control database. The WS scores can be modelled with Kernel Density Estimation (KDE) [141] or estimated as a single Gaussian [74].

- The H_1 distribution is estimated by modelling the between-source (BS) variability. BS scores are obtained by testing a set of GMMs trained from each speaker in the population database with the evidential recording. The BS scores can by modelled with KDE [141] or a mixture of Gaussians [74].

- Finally, the evidence E (Equation 2.12) is evaluated on the H_0 and H_1 distributions, enabling the LR (Equation 2.11) to be expressed.

Speakers are selected for the reference population based on an analysis of the evidential recording. Reference population speakers should be tailored to each case, to match as closely as possible the detailed speaker profile determined from the evidential recording analysis. In practice however, the diversity of speech data required to meet this aim is not readily available. At the very least, the reference population speakers should be of the same gender and accent as the suspected speaker [97, 164]. This is an open issue in the field, and is revisited in Chapter 6.

2.5.6 Assessment of forensic automatic speaker recognition systems

For scientific evidence to be admissible in court, its validity and reliability should be known and accepted by the relevant scientific community [3]. In terms of speech evidence, this demands that the accuracy and reliability of the FASR system is known [74].

A representation of the results of an FASR system offering this information is the Tippet plot. Originally proposed for forensic DNA analysis [60], Tippet plots represent the actual distributions of LR values for the same-speaker and different speaker hypotheses. Thus, the proportion of *misleading* evidence, i.e. the proportion of LR values that provide support for the wrong hypothesis, can be observed. An example of a Tippet plot is provided in Figure 2.8.

The C_{llr} metric [28] can be used to evaluate the performance of the FASR system across the full range of observed LR values, assessing them depending on their numerical value, and penalizing highly misleading LR values [76]. Since the LR is to be used in combination with a prior (set by the court), its performance can be evaluated by considering the effect that all possible priors have on the posterior probabilities. The APE-curve [76] is a plot of the total (posterior) probability of error with respect to all possible prior probabilities. The area under the curve corresponds to the C_{llr} value. Thus, APE-curves can be used as a measure of both the discrimination and calibration performance of the FASR system.

2.5.7 Problems in LR estimation

The challenge of inter-session variability faced by speaker recognition systems in general is especially prevalent in the forensic domain. Due to the fact that an evidential recording is obtained in way that is completely uncontrolled, it may be of short duration, poor quality and contain intra-speaker variability that is intentional (e.g. voice disguise) or natural (e.g. stress) [34].

This presents a challenge in estimation of the between-source score distribution, as there will likely be a mismatch between the characteristics of the evidential recording and the recordings of the reference population speakers [6, 20]. There can also be a difficulty in estimating the within-source score distribution, particularly if there is only one suspect speaker recording and one evidential recording to work with [20, 74]. A strategy for estimation of the within-source score distribution where only one suspect speaker recording is available, proposed by Gonzalez-Rodriguez et al. [74], is elaborated on in Chapter 6.

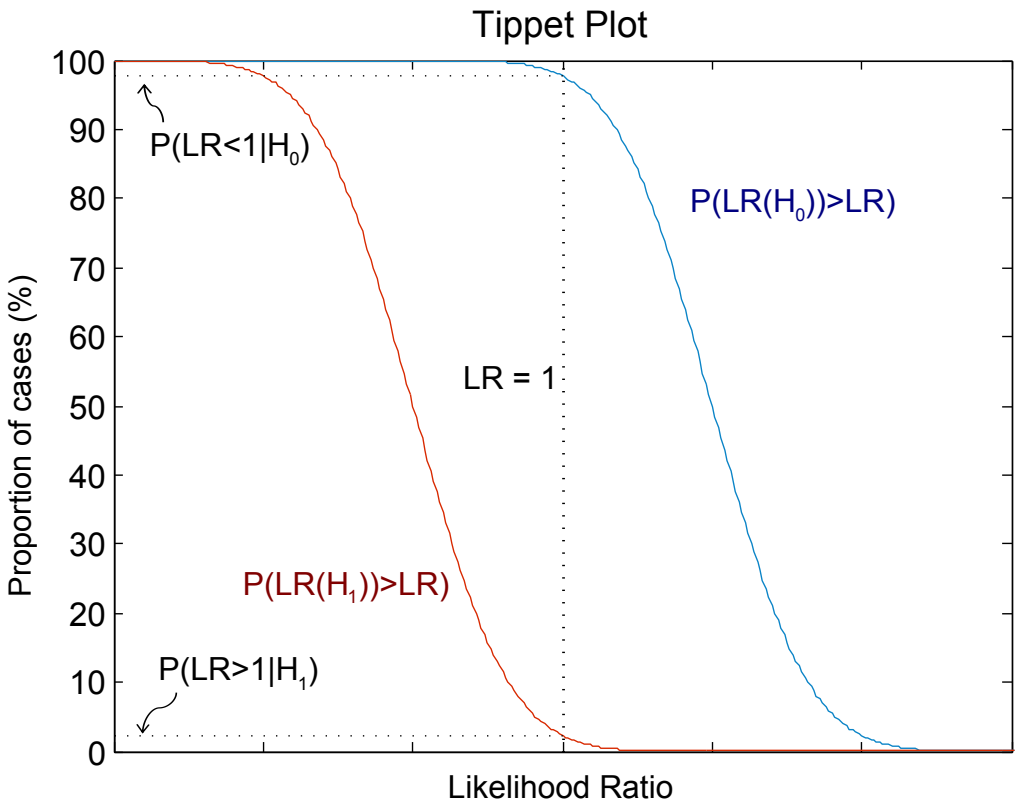

Figure 2.8: An example of a Tippet plot. The blue curve indicates the probability that the LR is greater than a given value when the same-speaker hypothesis H_0 is true. The red curve indicates the probability that the LR is greater than a given value when the different-speaker hypothesis H_1 is true. Thus, the proportion of misleading cases, where the LR supports the wrong hypothesis, is given by $P(LR < 1|H_0)$ and $P(LR > 1|H_1)$.

2.5.8 LR estimation with limited data

Given that there is likely to be a mismatch between suspect speaker and evidential recording conditions, it is desirable to have multiple recordings of the suspect speaker in different conditions to generate a representative set of scores. In the case of only having one suspect speaker recording, the first issue is obtaining a suitable number of scores, and the second is the use of these (matched condition) scores to estimate a distribution representative of mismatched conditions.

Gonzalez-Rodriguez et al. [74] propose a 'bootstrapping' procedure to generate a set of scores given one suspect recording. The recording is divided into N segments, and a model is trained for each. For each segment model in turn, the remaining N-1 segments are treated as test samples, and a set of N-1 scores are obtained. This results in a very conservative estimate of the within-source score distribution. Thus there is a 'separation' of within-source and between

source distributions, Figure 2.9. This leads to erratic LR values when the (log-likelihood ratio) score lies between the two distributions, as may happen in the case of mismatched conditions.

Within-source degradation prediction (WDP) [74] was proposed to adjust the within-source distribution to account for degradation due to mismatched conditions. The procedure assumes that the between-source distribution is estimated correctly (i.e. no imposters are expected to score higher than observed in its estimation) and that the within-source distributions is Gaussian.

Defining $f(x|H_0) = \mathcal{N}(\mu_{WS}, \sigma_{WS})$ and $f(x|H_1)$ as the within-source and between-source distributions, the aim is map the within-source distribution to $f_{WDP}(x|H_0) = \mathcal{N}(\mu_{WDP}, \sigma_{WDP})$. To estimate the mean and variance parameters of $f_{WDP}(x|H_0)$, the score S_{low} is obtained, according to:

$$\int_{S_{low}}^{\infty} f(x|H_1)\, dx = \alpha \tag{2.14}$$

where the α parameter controls the lower bound of the mapped within-source distribution, and is set to 0.01. The new within-source mean is then given by:

$$\mu_{WDP} = \frac{\mu_{WS} + S_{low}}{2} \tag{2.15}$$

Given μ_{WDP}, the within-source variance σ_{WDP} is found such that the following is satisfied:

$$\int_{-\infty}^{S_{low}} f(x|H_0)\, dx = \alpha \tag{2.16}$$

Figure 2.9 demonstrates the effect of applying WDP to the bootstrapped scores from a single suspect speaker recording. There is no longer an 'erratic' region between the score distributions. In addition to WDP, two complementary operations are applied to the WS score distribution.

In order to prevent the very low within-source variance estimates that can occur given one suspect recording, within-source minimum variance limiting (WMVL), which prevents the variance dropping below a certain value, is applied.

A second technique is outlier removal, which aims to remove the effects of spurious within-source scores (which can be significant in the case of a small quantity of scores). This is applied by discarding scores below a threshold S_{out} that is dependent on the between-source distribution:

$$\int_{-\infty}^{S_{out}} f(x|H_1)\, dx = \gamma \tag{2.17}$$

where the γ parameter controls the threshold and was set as 0.25.

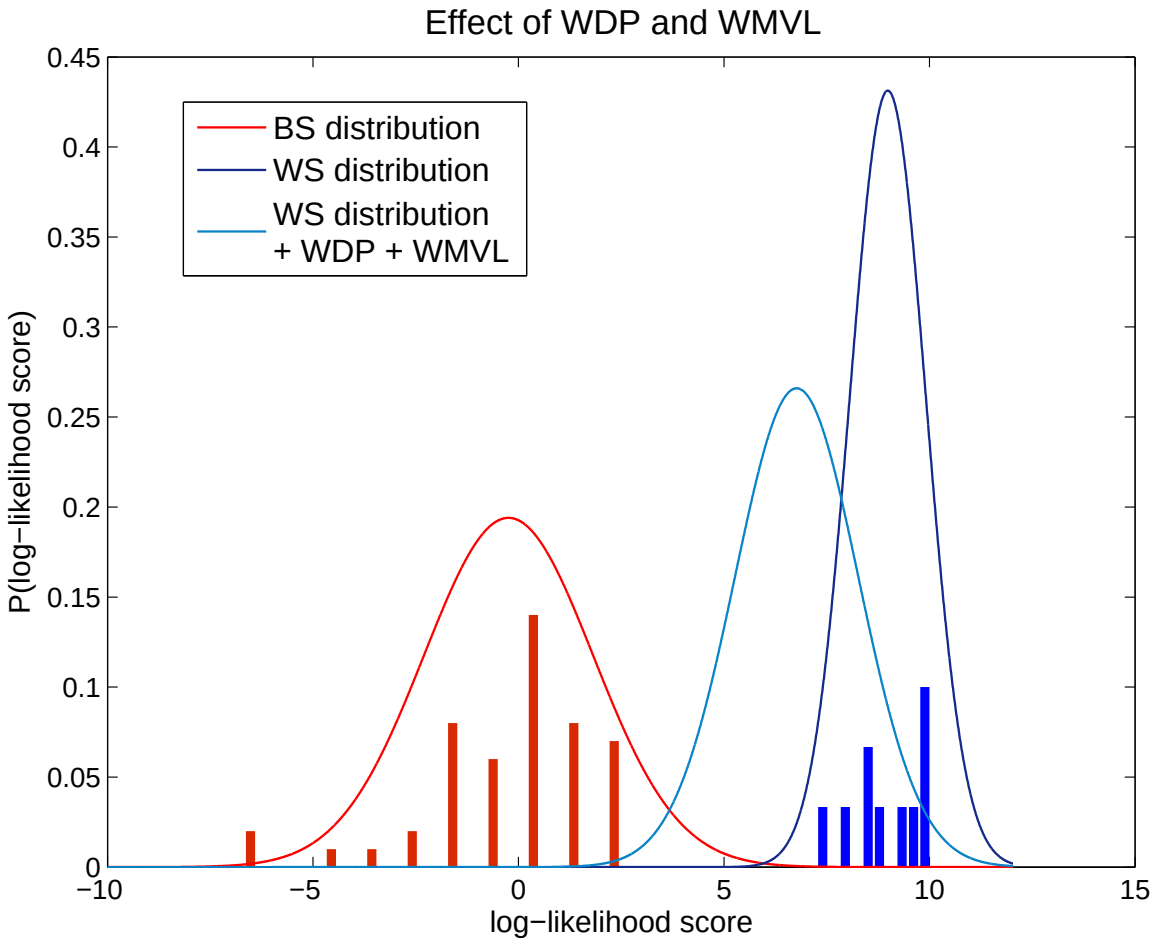

Figure 2.9: Effect of applying WDP (within-source degradation prediction) and WMVL (within-source minimum variance limiting) to the WS distribution. The distribution of log-likelihood ratio scores output from the GMM-UBM system for same-speaker and different speaker comparisons are given by the blue and red histograms and PDFs respectively. The increase in variance and shift in log-likelihood score is evident after applying WDP and WMVL to the WS distribution.

3

Vocal Ageing

The ageing process constantly drives change in all aspects of the human physiology. The voice is one of the many observable indicators of ageing - a young man and an elderly man, for example, are usually easily distinguishable based on their voices alone.

Although ageing-related vocal change is most noticeable in adolescence and old age, ageing affects the voice throughout adulthood. The primary source of change is physiological, due to degeneration and/or growth of the speech production organs. General health and lifestyle, and whether someone is geographically or socially mobile contribute to vocal change throughout adulthood. Neurological degeneration, affecting cognition and motor control, and neurological conditions, like Parkinson's disease, are common sources of vocal change in old age. The combination of these factors make the effect of the ageing on the voice, or *vocal ageing*, a complex and individual process.

This Chapter offers an experimental investigation into the effects of vocal ageing on various acoustic features of the voice and on the behaviour of an automatic speaker verification system. The resulting observations motivate and inform the development of ageing compensation methods for speaker verification, which are presented in subsequent Chapters. The Chapter begins with a brief description of the human speech production system, followed by a review of adult vocal ageing. An ageing database compiled for this Thesis is then introduced and used to evaluate proposed acoustic correlates of ageing. This is followed by an initial investigation into the effects of ageing on GMM-UBM speaker verification.

3.1 Speech Production

The production of human speech involves multiple complex systems working in synchronisation. A brief background of speech production, relevant to an understanding of ageing-related vocal change, is presented here. The descriptions are based on those found in high-level speech technology texts [95, 172].

The speech production system consists of the respiratory (airflow), laryngeal (phonation) and the supralaryngeal (articulation & resonance) systems, Figure 3.1. The respiratory system is contained within the chest cavity, or thorax. It is comprised of the lungs, respiratory musculature (diaphragm & intercostal muscles) and the trachea. Upon diaphragm and/or intercostal muscle contraction, the thoracic volume is increased and air is drawn into the lungs. When the respiratory muscles relax, air is forced from the lungs and through the trachea.

The laryngeal system generates acoustic energy from airflow. The larynx contains the key components of phonation: the vocal folds (or chords). These are small folds of tissue, stretched across the opening of the larynx, attached to cartilages at each side. The activation of muscles controlling the rear cartilages allow the folds to open and close, modulating airflow from the trachea. In an open position, the triangular passage between the folds forms the glottis. Muscles along the folds themselves allow their tension to be controlled.

When the vocal folds are closed and kept under tension, with sufficient air pressure, they begin to oscillate. The modulation of airflow by the vibrating vocal folds is referred to as phonation. The frequency of vibration is determined by both the mass and tension of the folds. Incomplete closure of the vocal folds during phonation results in a breathy voice. By changing the opening size and fold tension, other types of phonation, including creak, falsetto and whisper can be produced.

The supralaryngeal system is comprised of the oral and nasal cavities and the pharynx (the section of throat above the larynx). The function of this system is to provide resonation and articulation of the laryngeal output. The volume and shape of the supralaryngeal tract is highly flexible. By manipulating the oral cavity (i.e. the tongue, teeth, lips, hard and soft palates and the lower jaw) and the pharyngeal cavity, the production of complex speech sounds is made possible.

There is great variation in the anatomy of the speech production system between individuals. Gender is a major factor; the length of the vocal folds and vocal tract, and the size of the larynx are all greater in adult males than adult females [12]. However, even among (genetically related) same-gender members of a family for example, there will be differences in the precise size and shape of the supralaryngeal components [12].

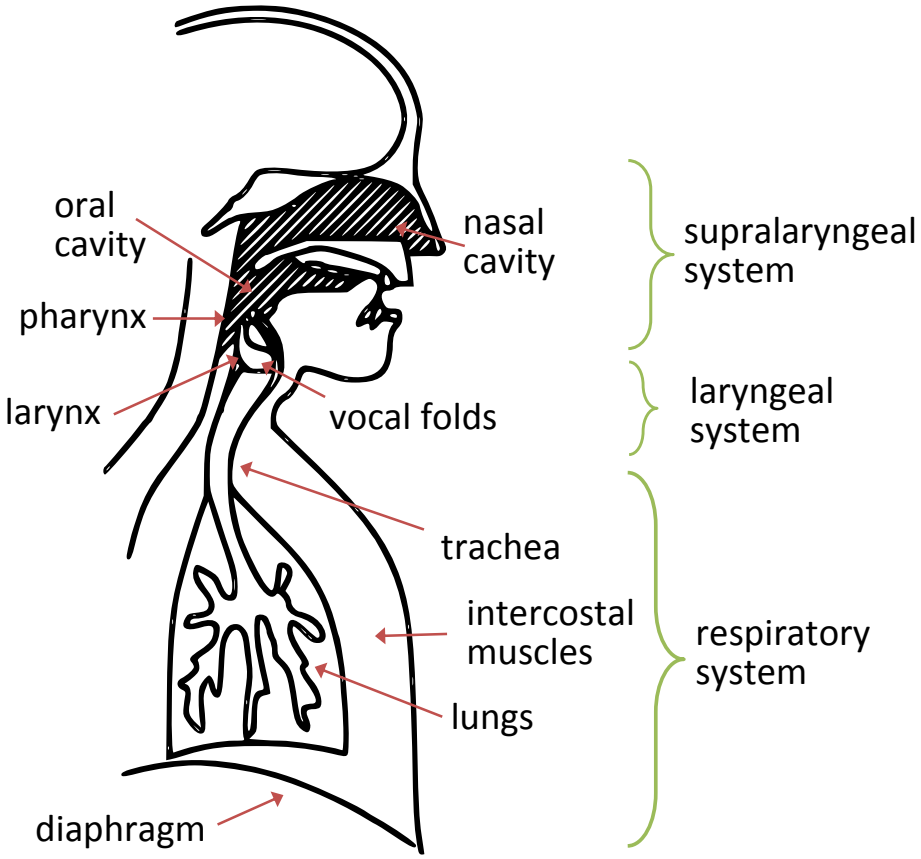

Figure 3.1: The main components of the speech production system (adapted from [176])

3.2 Ageing Process

The study of ageing has received an increasing amount of research attention in recent years. The increase in the average human lifespan globally, and the rising proportion of elderly people in developed countries has stimulated this interest [187].

It is generally agreed that there is no distinct 'cause' of ageing, but rather a complex combination of genetic and environmental factors. Weinert [198] provides a review of the main theories of ageing, falling into four categories: evolutionary, molecular, cellular and system-level. There is overlap between these theories, allowing them to be broadly split into two categories: programmed events and cumulative events. The first category considers ageing changes as resulting from genetic, cellular or hormonal triggers, i.e. the regulation of ageing is controlled in a similar manner to general development in childhood. The second category considers ageing the result of cumulative 'wear and tear' during the lifespan, i.e. the ageing process for a particular individual is a result of lifestyle, environmental factors and 'random' events like disease. The relevance of ageing theory to this Thesis is that it underlines the fact that ageing is a very complex, not-fully-understood process, that appears to be a combination of genetic and situational factors, and it can affect individuals in very different ways.

3.3 Vocal Ageing Change

The changes to the physiology of the speech mechanism that occur with 'normal' ageing have received much research attention over the last few decades. Linville has provided a number of comprehensive reviews on the subject [130, 133]. These, along with studies by Beck [12], Mueller [144] and Kahane [105] form the basis of the following review.

Ageing affects the speech mechanism in the same way as many of the body's systems: weakening of muscles and loss of tissue elasticity. The specific effects on each of the main parts of the speech mechanism are detailed in the following subsections.

3.3.1 Respiratory system

The respiratory system is central to the speech production system, acting as its power source. The efficiency of the system is dependent on the state of its main components: the lungs, thorax and respiratory muscles (diaphragm and intercostal muscles), Figure 3.1. From birth through to adolescence, respiratory function increases rapidly with growth. It peaks in young adulthood (20s) before gradually declining with age [105, 132]. Overall respiratory function (vital lung capacity and rate of forced expiration) of a man in his 70s may be less than 50% of that of a man in his 30s [105]. The primary ageing processes at work are muscle weakening and loss of tissue elasticity:

- Decreasing strength of respiratory muscles
- Stiffening of the thorax
- Loss of lung elasticity

These anatomical changes lead to decreases in vital lung capacity, forced expiratory volume and rate of respiratory (diaphragm and intercostal) muscle contraction [105, 130].

3.3.2 Laryngeal system

Changes in the laryngeal system (Figure 3.1) differ between sexes, both in development and ageing. From puberty to adulthood, male laryngeal cartilages and vocal folds grow at a greater rate than those of females [12]. With ageing, the properties of the laryngeal cartilages and vocal folds change, with different patterns in males and females [12, 105, 130, 144, 181].

- Calcification and Ossification of laryngeal cartilages: affects males earlier (30s-40s) than females (40s-50s). Affects males more extensively.
- Atrophy of laryngeal muscles: both males and females.
- Thickening (edema) of vocal fold layers (the lamina propria and/or the epithelium) in females (50s) - likely due to hormonal changes at menopause. Thinning of vocal fold layers in males (50s-60s).

- Degeneration of laryngeal mucous glands: both males and females.

These anatomical changes lead to stiffening, weakening and dehydration of the laryngeal cartilage, musculature and tissue, and a change in the density and flexibility of the vocal folds. There is some conflict in the literature as to the specifics of laryngeal changes in healthy ageing adults. The above points are generally agreed upon, and provide explanations for the change to the acoustic properties of the ageing voice (Section 3.4).

3.3.3 Supralaryngeal system

The supralaryngeal system (Figure 3.1) controls the articulation and resonance of speech. There are many changes to the supralaryngeal system with ageing, which contribute in a complex way to the acoustics of the ageing voice [105, 130, 181]:

- Continuous growth of facial skeleton throughout adulthood: 3-5% symmetrical enlargement from 30s-50s.
- Atrophy or weakening of the facial, mastication, pharyngeal muscles and the tongue.
- Thinning and loss of elasticity of oral mucosa epithelium (outer layer of tissue inside the mouth).
- Changing oral cavity size due to tooth loss, bone growth or swelling of mucosa.

These changes within the oral and nasal cavities primarily affect the speed and precision of articulation and the timbre of the voice.

3.3.4 Neurological system

General motor function is affected by the ageing-related changes to the central nervous system. This may slow the speech production mechanism and affect the coordination of the articulators [181, 192]. Diminished cognitive-linguistic function may also play a role [190], although this is likely to become apparent only in advanced old-age. In addition, there a number of neurological disorders common in old age (addressed in Section 3.3.5) that may affect the voice noticeably [133].

3.3.5 Other sources

The preceding Sections 3.3.1 to 3.3.4 detailed the main anatomical, physiological and neurological ageing-related changes to the subsystems of speech production.

These changes can be considered ‘normal’ or ‘typical’ in the sense that they occur in all ageing individuals to a certain extent. However, throughout the lifespan, the voice is influenced by many other factors. Differences in vocal anatomy and physiology, and a genetic predisposition to a particular pattern of ageing, combine with the external factors of lifestyle and health to

produce different ageing patterns for different individuals [12]. The cumulative effect of these influences result in increasing variability between speakers as ageing progresses [144].

Lifestyle factors can either postpone or accelerate the effects of normal vocal ageing. The most controllable, and perhaps the most significant, are physical fitness and smoking [16, 133, 193]. The maintenance of physical fitness with ageing promotes the preservation of respiratory system health and motor function. Smoking, on the other hand, accelerates the changes eventually seen with ageing: declining respiratory function and laryngeal deterioration. In addition, smoking is typically a factor in acute respiratory conditions like emphysema [57], which results in a weak, hoarse and breathy voice. It also contributes to laryngeal disorders such as edema of the vocal folds [12], leading to increased vocal fold mass and hence a lowering of speaking fundamental frequency.

Other environmental factors include the geographical mobility of the individual [22,179], and whether they have had professional voice training [180].

Social and cultural factors can result in intentional, or unintentional, shifts in pronunciation [165]. It has been proposed that socially-motivated adaptation of language and speaking style are used by individuals to profit from a particular social surrounding [21]. An example of cultural influence is the observation that the pronunciation of an individual can change as a result of mainstream trends: Harrington et al. [85] concluded that there was a change in Queen Elizabeth II's standard British received pronunciation (RP) over several decades due to the influence of mainstream accent shift.

Physiological and neurological conditions that affect the voice become increasingly common with ageing; voice disorders affect up to 12% of the elderly population [133]. Above the age of 60, central neurological disorders including stroke, Parkinson's disease and Alzheimer's disease are common. Physiological conditions include benign (non-cancerous) lesions, particularly in elderly females. They are characterized by edema (or swelling) along the layers of the vocal folds, and are exacerbated by smoking [133]. Inflammatory conditions include laryngitis, which can be caused by smoking and poor hydration [133]. Side effects of medication for treating the above, or other, conditions may also adversely affect speech [133]. These disorders typically occur in speakers above the age of 60, but may occur earlier, particularly in the case of smokers.

An interesting point to consider is that recent advances in medicine, such as tissue engineering utilising stem cells, open up the possibility of therapy of the severely disordered voice [8]. Since many of the symptoms of a disordered voice may also occur as part of the 'normal' ageing process, the more severe effects of vocal ageing may be preventable or even reversible in the future.

In summary, the change in the voice with ageing can be viewed as a physiologically and neurologically driven process, influenced by the genetic make-up of the individual, and the contribution of external factors.

3.4 Acoustic correlates of adult speaker ageing

The sources of change in the ageing voice, as reviewed in Section 3.3, help to inform an understanding of some of the complex acoustic changes to the voice accompanying ageing. This section will discuss the main acoustic correlates of the ageing voice. In the subsequent Section 3.6, an experimental evaluation of the main acoustic correlates will be presented.

3.4.1 Fundamental Frequency

Out of all ageing voice characteristics, fundamental frequency (F0) has received the most extensive research attention [12, 47, 130, 131, 144, 160, 164, 185, 189, 194, 201]. Despite this, the pattern of F0 change is not yet fully understood. What is certain however, is that there is a difference between male and female F0 change.

Male F0, having decreased significantly during adolescence, continues to decrease at a more gradual rate throughout adulthood (by around 10 Hz). In the 50s-70s, the trend reverses, and F0 begins to increase (by up to 35 Hz) [130]. The precise age at which this occurs appears to vary significantly between individuals. The decline throughout adulthood may be attributed to the ossification, and associated stiffening, of the laryngeal cartilages. The rise of F0 into old-age is likely due to muscle atrophy and loss of flexibility of the vocal folds [12].

Female F0, after decreasing during adolescence, remains relatively stable until the 50s, where it begins to decrease by up to 15 Hz [130]. This is presumed to be as a result of hormonal changes during menopause, which bring about swelling or edema of the vocal folds, increasing their vibratory mass.

The F0 trend is more variable in males. For example, listeners have reported that a lower male F0 is a cue to older age [132], while at the same time a 'weak', 'hollow' or 'thin' voice is also a cue to old age [12] - F0 in this case would not be described as low. This suggests that in males, interpersonal variation and external factors of lifestyle and health may exert a more significant effect on F0 than normal ageing.

Aside from these gender-specific changes, the standard deviation of F0 for both males and females increases with age [144, 189], reflecting an increase in vocal tremor or 'wobble' [131]. Increased F0 variability may be attributed to reduced breath support and loss of laryngeal control and flexibility.

3.4.2 Speaking Rate

The speaking rate of both males and females has been shown to decrease with ageing [100, 131, 181]. A decline in speaking rate in the range of 20-25% has been observed in general [181]. The rate of speaking is linked to speech segment (e.g. syllable) duration and frequency, and also pause duration and frequency. It has been proposed that the cause for an increase in syllable and pause duration and a reduction in their frequencies is due to the impairment of the

respiratory system; leading to increased breathing (and hence pause) rate [131,144,157]. Decline in neurological motor control is also a factor [117, 181, 192]. Deterioration in the flexibility of the articulators may be another contributor.

3.4.3 Jitter and Shimmer

Jitter and shimmer are measures of stability of vocal fold vibration. Jitter is the cycle-to-cycle variation in the fundamental frequency of vocal fold vibration. Shimmer is the cycle-to-cycle variation in the amplitude of vocal fold vibration [131]. Increased amounts of either jitter or shimmer result in speech perceived as having a rough or hoarse voice quality [181]. Several studies have reported jitter and shimmer to be greater in the voices of older males and females than those of younger speakers [148, 194, 201]. However, there have also been reports of jitter and shimmer remaining relatively stable with age [181]. The source of jitter and shimmer is likely a combination of muscular atrophy and a decline in neuro-muscular control [144]. As a result, these measures are strongly linked with the overall health of an individual [131]. Thus, although increased jitter and shimmer may be characteristics of an aged voice, they may not conform to general trends with ageing across the population. Mueller [144] suggests that these measures may be more indicative of the overall health of an individual than their chronological age.

3.4.4 Spectral Noise

Speech with a relatively high measure of spectral noise could be described as having a 'breathy' quality. As breathiness is a characteristic of an aged voice, spectral noise has been proposed as an acoustic correlate of ageing. Ossification and atrophy of laryngeal cartilages and muscles, and reduced breath support may all be responsible for increased spectral breathiness of speech. The spectral noise in a speech sample can be measured with the noise-to-harmonic ratio (NHR). The NHR is defined as the ratio of the noise to the energy of the speech in the periodic part of the signal. The NHR has been reported to increase with age in both men and women [201]. However it has also been reported to increase only in men [181], and conflictingly, to remain stable in men [194]. Thus, as with jitter and shimmer, spectral noise may be a symptom of health deterioration beyond the bounds of 'normal' ageing.

3.4.5 Other acoustic correlates

Other acoustic correlates of the ageing voice have also been reported:

- Spectral changes: F1 (first formant) decreases in men and women. F2 and F3 also decrease, significantly more in women than men [165, 181]
- Vowel space area estimations (VSAe): given by the area of the polygon described by plotting average F1 vs average F2 [164, 165] show a tendency to decrease above the age of

35.

- Voice onset time (VOT): has been reported both to increase [189] and decrease [47] with ageing. However, there is agreement on VOT becoming more variable with ageing [15, 47, 189].
- Intensity: The maximum intensity range decreases for men and women [181]. However, conversational speech intensity has been found to increase in men after age 70, while remaining stable in women [181].

3.5 Speaker Ageing Data

The vast majority of experimental studies on the ageing voice have used cross-sectional data, e.g. [24, 80, 185, 189, 194]. An obvious advantage with this approach is that recordings can be obtained in controlled conditions from many speakers covering the desired age-range. The downside of cross-sectional study of speaker ageing is that it assumes analysis of contemporary groups of speakers of different ages is representative of individual speaker analysis across ageing. Where the study of ageing speaker verification or forensic speaker recognition is concerned, longitudinal data is a necessity. The practicalities in obtaining longitudinal data, however, make it a difficult approach. Aside from beginning long-term data collection, the only way to proceed is to source pre-recorded material, over which control of the speech content and recording method is impossible. As a result, vocal ageing studies utilising longitudinal recordings vary widely in their methodology.

In one of the earliest longitudinal studies, Endres et al. [58] analysed formants and fundamental frequency measurements for 6 speakers from several recordings over a 13-15 year period. Harrington et al. used recordings of Queen Elizabeth II's annual Christmas broadcasts to analyse longitudinal shifts in her pronunciation over periods of 30 years [85] and 50 years [83]. In further studies by Harrington et al. [84] and Reubold et al. [160], trends in formant and fundamental frequencies from 4-5 speakers over a period of 29-56 years were presented. The subject speakers again include Queen Elizabeth II, and also the broadcaster Alistair Cooke, for which recordings spanning 56 years were analysed. Sankoff [179] and Rhodes [164–166] used excerpts from 'The Up Series' television series [2] to analyse longitudinal trends in multiple phonetic and acoustic features, for 2-8 speakers over a 28 year period. Decoster and Debruyne [47] used recordings of 20 male speakers 30 years apart in a longitudinal analysis of several acoustic measurements.

In the forensic speaker recognition domain, there have been several longitudinal studies. Hollien and Schwartz [94] used recordings from 10 males over a 20 year period in a naïve speaker recognition experiment (i.e. with human listeners, Section 2.5.1). Künzel [122] carried out a study comparing recordings of 10 males 11 years apart in both auditory and automatic speaker recognition experiments. Rhodes [165,166] presents both automatic and acoustic-phonetic (Section 2.5.1) forensic speaker recognition experiments using 6 males from 'The Up Series' [2], with

recordings over a 28 year period at 7 year intervals. A longitudinal study based on real forensic casework is the Yorkshire Ripper Hoaxer trial [63], which involved auditory-perceptual and acoustic-phonetic comparison of two recordings 26 years apart, claimed to be from the same speaker.

Aside from the publications contributing to this Thesis, there are no other published studies on long-term (greater than 3 years) speaker verification the author is aware of. There have been several studies utilising short-term longitudinal databases: CSLU [42] (Section 3.5.5.1), which contains speech over a three year period, and MARP [124], which contain speech spanning a two year period.

A speaker verification study by Lawson et al. [125], on the MARP corpus, concluded that over a period of three years, the extent of any ageing-related variability will most likely be exceeded by normal inter-session variability. A subsequent study on the MARP corpus [71] confirmed that over a three year period, any variability due to ageing was not detectable; inter-session variability over 1 month was comparable to that over 36 months. Another study, by Rose [171], found no difference between day-to-day variation in formant frequencies compared with variation over one year. The study of short-term longitudinal speaker verification is of importance for understanding day-to-day intra-speaker variation. To investigate the influence of ageing, which is the goal of this Thesis, a time-lapse of significantly greater length must be considered.

The Greybeard Corpus [23], compiled by NIST for their Speaker Recognition Evaluation (SRE) 2010, contains telephone recordings from 172 speakers over a 2-14 year period. For the majority of the speakers (154), there are recordings present for a 2-4 year period. Recordings over an 11 year period are present for 16 speakers, and over a 14 year period for 2 speakers. There are 12 recordings included for most speakers. The corpus was included in the NIST SRE 2010, and was made publicly available in June 2013. As of yet, no performance results or publications reporting on the Greybeard corpus have been released by NIST or any SRE participants.

For the purpose of investigating ageing speaker verification, an ideal database would contain regular recordings for each speaker in controlled conditions across several decades [123]. Due to the evolution of technology for recording, transmission and storage of media over the last few decades, the prospect of obtaining data in controlled, consistent conditions is unrealistic.

The Trinity College Dublin Speaker Ageing (TCDSA) database was compiled for this Thesis as a means to investigate ageing speaker verification and forensic speaker recognition. Recordings were collected for the database such that the ageing-range and quantity of recordings was maximised, while non-ageing-related variability was limited as much as possible. The database forms the core test set for all of the publications that contribute to this Thesis. The database was expanded from an original set of 13 speakers in [110], to 18 speakers in [107–109], and again to 26 speakers in [106, 111]. The most recent version is detailed in Section 3.5.1.

3.5.1 TCDSA Database

The Trinity College Dublin Speaker Ageing (TCDSA) database contains 26 speakers (15 male, 11 female) from a variety of sources. There are between 4 and 35 recordings per speaker, spanning a time-lapse of between 28 and 58 years in total per speaker. The recordings vary between approximately 1 and 30 minutes in length. The database contains 30 hours of content in total. Its content is summarised in Figure 3.2, with further details provided in Table A.1.

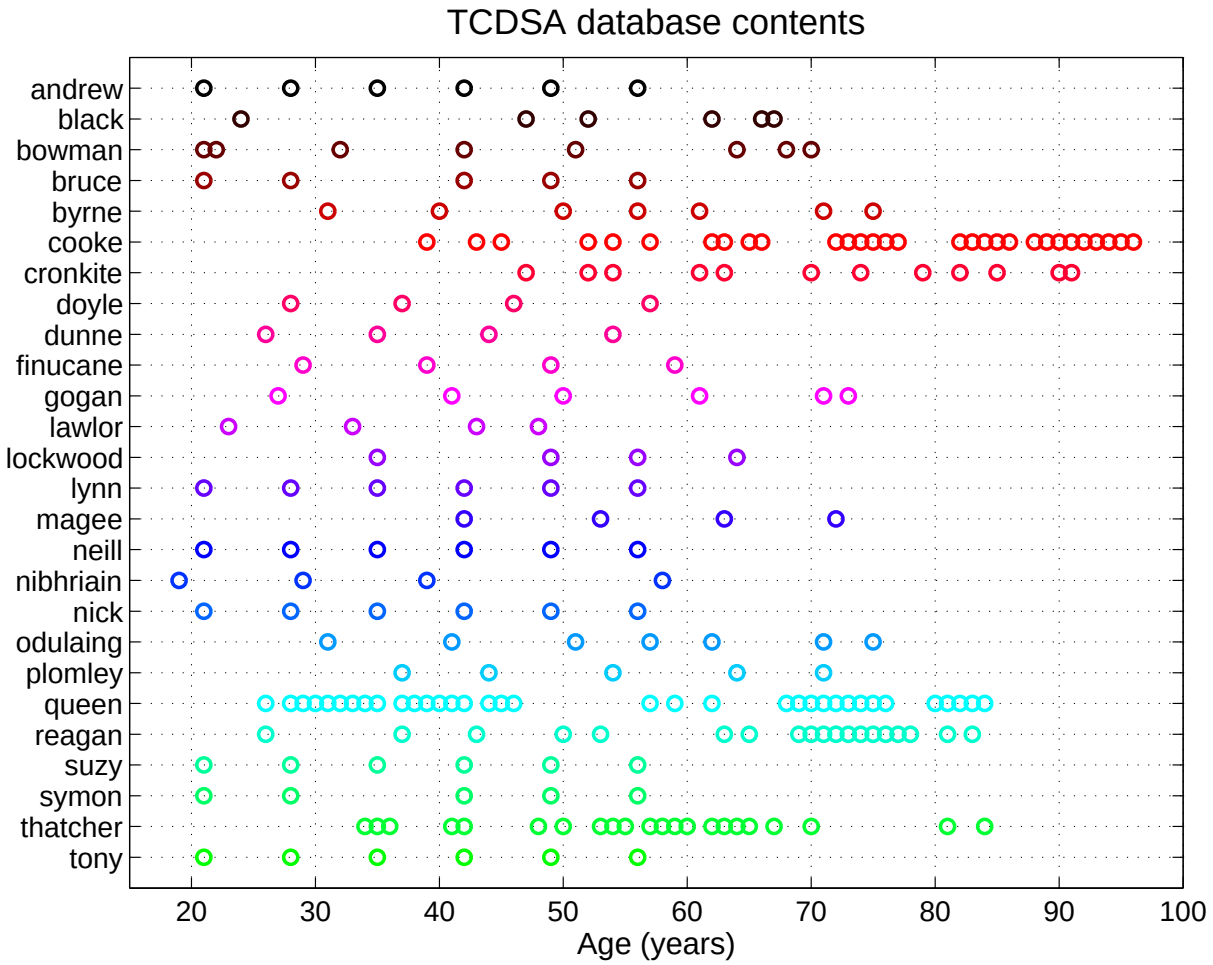

Figure 3.2: A schematic of the TCDSA database. Each circle indicates a year where a recording is available for a speaker. Speaker names are shown on the y-axis, with their age in each recording given on the x-axis

The primary sources of the database content are the national broadcasters of the U.K. and Ireland, the BBC (British Broadcasting Corporation) and RTÉ (Raidió Teilifís Éireann), from whom the recordings of 14 speakers were acquired. Recordings for 8 speakers were obtained from 'The Up Series' Granada television series [2]. The remaining speakers' recordings were sourced from YouTube and The Miller Center [153] (an American Presidential archive). Supplementary

recordings for a number of speakers were sourced from YouTube.

All recordings were downsampled to 16 kHz (most of the original sampling rates were 44.1 or 48 kHz, while the lowest rate was 16 kHz) and converted to mono using *Adobe Audition 3.0*. Long silences, music, applause and interlocutors were manually removed in *Audition*. All recordings were saved as 16 bit .wav files. The full database has been made freely available for academic research.

3.5.1.1 Sources of variability

As expected for a database of this nature, the recording conditions are variable. While all of the RTÉ and BBC broadcast data are studio recordings, the specifics of the recording equipment is unavailable. Recordings from the other sources are more variable, particularly in terms of recording environment. There are a mix of studio, small rooms, large halls and outdoor locations for The Up Series, YouTube and The Miller Center recordings.

Another source of variability in the database is accent. There are a range of Irish, English and American accented speakers in the database (detailed in Table A.1). All recordings are in the English language. A further complication is speakers who were geographically mobile. However, there is only one speaker, 'Nick', for which the resulting accent-change is perceptible over the samples in the database, and the effect is mild (this is the assessment of the author; there was no formal listening test to assess accent migration).

A stronger source of within-speaker variability is those speakers who smoked, and those with health issues (detailed in Table A.1).

The type of content varies also. The majority are interview or opinion-piece extracts, and thus are conversational-type speech. There are also a number of news/weather bulletins and prepared speeches, which are obviously less spontaneous in nature. For each speaker individually however, there is generally consistency in the type of content. Most of the speakers who are presenters (see Table A.1) remain broadcasting the same (or similar) programme for all of their samples in the database. All of The Up Series recordings, for example, are single sides of an interview with the same interviewer (Michael Apted), with similar questions in each interview. This results in a constrained speaking style and content between samples for most speakers.

3.5.1.2 Minimising variability

To limit non-ageing-related variability, the prospective database material was screened to remove highly-variable or low-quality samples. As a first step to limit non-ageing-related variability, the prospective database material was screened for quality. Any audibly noisy recordings, or segments of recordings, were discarded. The spectral content of the recordings was examined, and any with significant frequency artefacts (such as microphone interference, pops or low frequency hum) were removed. In addition, an objective model-based measure of quality was applied to screen the data prior to the speaker verification experiments in subsequent chapters. This

screening operation is detailed in Section 3.8.1.

There are benefits to using a database with widely varying content: firstly, it is a closer representation of realistic operating conditions for a speaker verification system. In the case of the TCDSA database, recording quality is much better than in a typical forensic speaker recognition comparison [165, 172]. In terms of recording environment and content, while inter-speaker variability is high, intra-speaker variability is less so - many of the speakers' recordings are in consistent locations and contain similar type dialogue across several years. Most of the individual Up Series samples are composed of short segments from different locations over several days or weeks of filming. This capture of short term intra-speaker variability is desirable for a speaker modelling [172]. For a few other BBC speakers, there are a number of blocks of consecutive recording years (see Fig. 3.2), which will also capture non-ageing-related intra-speaker variability. Thus, an analysis of speaker modelling and verification on this dataset will enable realistic observations on long-term ageing-related variability to be made.

3.5.2 TCDSA-UBM Database

The TCDSA Universal Background Model (TCDSA-UBM) database was compiled to provide UBM development data suitable for use with the TCDSA database. The University of Florida Vocal Aging Database [81], extemporaneous set (UFvadEx), was combined with data sourced from YouTube and The Miller Center [153]. The resulting dataset contains 1 hour of speech from 120 speakers (30 seconds each) evenly balanced across gender, age and accent. There was no objective measure of recording quality used for screening samples, and thus there is a range of recording quality represented. However, audibly noisy recordings were not included. All the recordings were downsampled to 16 kHz, converted to mono and all saved as 16 bit .wav files using *Adobe Audition 3.0*. Long silences, music, applause and interlocutors were manually removed in *Audition*. There are 40 speakers from each of three age groups: under 35, 36-55 and 55+. An approximately equal number of English, Irish and American accented speakers make up each of these groups. The dataset was composed in this way in an effort to ensure it was a valid representation of the population of the TCDSA database. A detailed schematic of the database is provided in Table A.2.

3.5.3 TCDSA-FD Database

The TCDSA Forensic Development (TCDSA-FD) database was compiled to provide suitable background and population modelling data for a forensic speaker recognition evaluation using the TCDSA males as subjects. It contains 5.3 hours of speech from 115 Irish-accented males (between 30 seconds and 5 minutes each) evenly spread across age. Samples were screened for quality subjectively and objectively before inclusion, in the same manner as the TCDSA database, Section 3.5.1.2. All recordings are interviews and speeches sourced from YouTube. All the recordings were downsampled to 16 kHz, converted to mono and all saved as 16 bit .wav

files using *Adobe Audition 3.0*. Long silences, music, applause and interlocutors were manually removed in *Audition*. There are approximately 20 speakers in each of five age groups: 25-35, 36-45, 46-55, 56-65 and 66+. All speakers are Irish-accented, with some variation in accent source and strength. Most speakers are public figures (politicians, presenters, musicians etc.), which enabled their exact age to be obtained. Further details for each speaker in the database are provided in Table A.3.

3.5.4 Irish-accented Females Database

An additional database of Irish-accented female speakers, similar in their profile and age distribution to the male-only TCDSA-FD database, was collected to provide development data for the experiments in Chapter 7. It contains approximately 30 seconds of speech from 25 Irish-accented females spread across age. Samples were screened for quality subjectively and objectively before inclusion, in the same manner as the TCDSA and TCDSA-FD databases. All recordings are interviews and speeches sourced from YouTube. All the recordings were downsampled to 16 kHz, converted to mono and all saved as 16 bit .wav files using *Adobe Audition 3.0*. Long silences, music, applause and interlocutors were manually removed in *Audition*. There are approximately 8 speakers in each of three age groups: 20-40, 41-55 and 56+. All speakers are Irish-accented, with some variation in accent source and strength. The speakers are a mix of public figures and members of the public. Further details for each speaker in the database are provided in Table A.4.

3.5.5 Other Databases

In addition to the ageing-specific databases compiled for this Thesis, the following existing datasets were used for experiments:

3.5.5.1 CSLU

The CSLU Speaker Recognition Corpus [42] contains telephone speech from 91 American-accented speakers (44 male, 47 female) recorded over a two year period in 12 different sessions. The database is sampled at 8 kHz 16 bit and stored as .wav files. The content of the recordings is a mix of scripted and spontaneous speech. Only the spontaneous speech was used for experiments in this Thesis, of which there is between 2 and 4 minutes in total available per speaker. A 5-10 year age-range is given for each speaker, from 16-20 years up to 61-70 years. The majority of speakers (70 out of 91) are between 21 and 60.

3.5.5.2 TIMIT

The TIMIT [69] corpus contains microphone speech from 630 American-accented speakers (428 male, 192 female), and was designed for developing automatic speech recognition systems. The

speakers are sourced from 8 different dialect regions and speak 10 sentences each. The average quantity of speech available per-speaker is 30 seconds. The content of the recordings is exclusively read speech. The ages of the speakers are heavily biased towards younger speakers; the vast majority (558 out of 630) are between 21 and 40. The database was released as 16 kHz 16 bit SPHERE format files.

3.5.5.3 YOHO

The YOHO [33] speaker verification database contains microphone speech from 138 American-accented speakers (106 males, 32 females) recorded over a three month period in 14 different sessions. All of the recordings are 'combination-lock' phrases e.g. "twenty-six, thirty-two, fifty-seven". There are 146 phrases per speaker, totalling approximately 10 minutes of speech per speaker. No age information is available, although the documentation states that a 'wide range of ages' are represented. The database was released as 8 kHz 16 bit SPHERE format files.

3.6 Analysis of acoustic correlates of adult speaker ageing

In Section 3.4, the main acoustic correlates of ageing were detailed. In this section, an evaluation of these acoustic measurements on the subjects in the TCDSA database is presented. The purpose is to provide a longitudinal analysis of those features commonly cited as ageing-dependent, but most often investigated via cross-sectional studies. This allows individual as well as general speaker trends to be observed. Results are presented on an individual speaker level, and thus can be discussed in relation to their speaker profile and recording characteristics. Building up a background profile for speakers in the TCDSA database in this manner will inform discussion on automatic experiments in subsequent Sections. An individual speaker approach is essential to a forensic analysis context [172].

All acoustic features were extracted automatically from voiced segments of speech using *Praat* [19] software. *Praat* has been used for analyses in similar studies [165, 181, 194]. The extraction of acoustic features from sustained vowel phonemes (such as 'aa'), as in [160, 185], would provide more accurate measurements. However, the manual labelling and segmentation of vowel phonemes would have been prohibitively time consuming. As a result, there is likely some variability in F0 and its related measures, due to unwanted phonemes or segments of raised pitch or creak being considered. To smooth out the effect of these variabilities, the length of recording time used for analysis is maximised. Several studies have used automatic extraction of some or all of the features considered here from spontaneous speech [24, 47, 160], so it is not without precedent. Furthermore, several of the TCDSA database speakers' F0 trends were analysed in previous studies, allowing the analysis here to be validated. F0 measurements for Queen and Cooke, estimated from both voiced frames and segmented vowel phonemes, are found in [160]. F0 measurements for all of The Up Series speakers, estimated from segmented vowel phonemes, are presented in [164, 165].

Based on the observations of the differences between male and female vocal anatomy and ageing patterns, the acoustic feature analysis is presented in a gender-dependent manner. For every recording in the TCDSA database, Figure 3.2, each acoustic feature was extracted from a segment of up to 15 minutes (dependent on data length) duration. Both the absolute value of the features at each of the available ages and a relative percentage change across different age ranges is provided.

3.6.1 Fundamental Frequency

The prediction for F0 in males is for it to fall gradually until the 50s-70s and then begin to rise. For females, it is expected to fall gradually until the 50s, where it will decrease more sharply. The mean F0 for each male and female speaker across ageing is shown in the upper plots of Figure 3.3. The percentage change in F0 with increasing age, averaged over 10 year intervals, is shown in the lower plots of Figure 3.3.

Reubold et al. [160] present F0 measurements of Cooke and Queen using the same data as was used here (with the exception of additional Queen data at 80+ years). There is close agreement between the F0 trends found in their study and those in Figure 3.3. Furthermore, they compare F0 measurements extracted from segmented vowels with those extracted from voiced frames, and conclude that there is no significant difference. Rhodes [164, 165] presents F0 measurements of The Up Series speakers (see Table A.1), extracted from segmented vowels. Again, there is close agreement between these trends and those presented in Figure 3.3. The comparisons with these previous studies validate the use of automatic extraction of F0 (and related measures) from voiced frames.

For females, the trend in the relative change over 10 year intervals, Figure 3.3, shows F0 to generally follow the predicted trend - a consistent decrease is observed for the majority of speakers. The smokers, Finucane, Lockwood and Lynn show a significantly sharper decrease in F0 than the other females. The F0 for these speakers drops by approximately 30% by their 50s. The absolute F0 for Finucane and Lockwood in their 50s are the lowest of all females, at approximately 130 Hz and 120 Hz respectively. This demonstrates, quite dramatically, the effect of smoking on fundamental frequency in females. This corroborates findings by Rhodes [165], whose F0 trend for Lynn is consistent with that presented here

The F0 for Thatcher is markedly more variable than the other speakers. However, her recordings are far more variable in terms of environment (room, hall & outdoors) and speaking style (relaxed interview & animated speech) than the other females. In her final two recordings, the influence of her declining health may be a factor.

The influence of environment is also observable in the F0 values of Queen from 80 onwards. These later samples are YouTube recordings supplementing the BBC studio recordings, and are in differing environments.

These individual cases aside, there is a consistent decrease in F0 with ageing, with the 40s

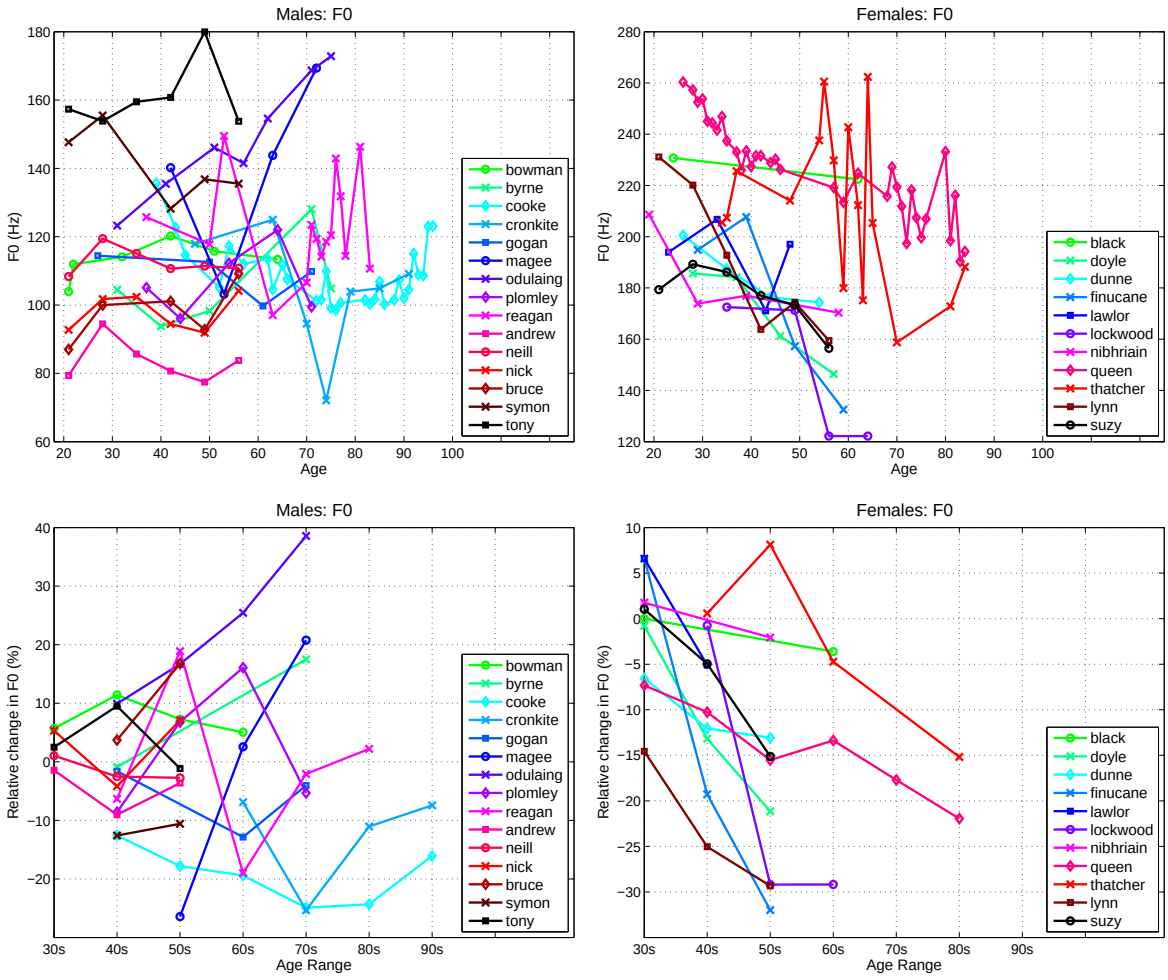

Figure 3.3: **Top:** Mean fundamental frequency (F0). **Bottom:** Percentage change in mean F0, relative to a speaker's youngest available sample, e.g. the y-axis value at Age Range = '50s' is the percentage difference between a speaker's mean F0 in their 50s and their mean F0 in their first decade of available data (their 20s or 30s in most cases). Note that there are different y-axis scales for males and females.

being the decade with most rapid decrease in general, which is slightly earlier than the predicted 50s.

Male speaker F0, Figure 3.3, is less consistent between speakers than with females. Observing the relative F0 changes in Figure 3.3 however, the predicted falling-rising pattern of F0 can be seen for most speakers, albeit at different ages.

The dramatic rise in F0 for O'Dulaing is apparent from listening to the relevant recordings. His later recordings possesses some of the attributes associated with an aged male voice: breathy, hoarse and weak.

Similarly to Thatcher in the female results, the recording environment and speaking style are influencing factors in the variability of F0 for Reagan and Cronkite. There are no smokers among the male subjects.

The predicted falling-rising of F0 in males with ageing is observed, with a wide variability in the age of the 'turning point'. Some males' F0 rises consistently from their 40s: O'Dulaing, Bruce, Byrne etc. Others rise only in their 70s (Cronkite) or 80s (Cooke). Thus, the intra-speaker variability in the onset of F0 rise among males is high. This may explain the discrepancy between studies on whether male F0 rises or falls with ageing. A possible reason for the late turning-point of F0 for Cooke is that he is a professional speaker, who through sustained voice use, have been able to delay the onset of F0 rise. This point was suggested by Reubold et al. [160]. However, many of the other males, who are also broadcasters, do not demonstrate this delay.

In Figure 3.4, the standard deviation of F0 with ageing is presented. In the percentage change plots at the bottom half of Figure 3.4, male std. dev. of F0 generally rises until the 50s where it levels off or drops slightly. For females there are conflicting trends between speakers. Speakers for which there is relativity a lot of data points, the queen and thatcher, show a rising trend. Interestingly, two of the three smokers, Finucane and Lockwood show a falling trend. The prediction of increased standard deviation with ageing is not fully observed, but there is a rising tendency for most speakers, male and female.

3.6.2 Speaking Rate

Speaking rate for each recording was estimated as the number of syllables divided by the speaking time (with silences removed). This measure is commonly referred to as the articulation rate, and was extracted with a Praat script using syllable nuclei detection to count syllables [46].

The predictions for articulation rate are that it will decrease for males and females with ageing [181], although it has been reported to decrease to a lesser extent in spontaneous speech than read speech [24, 100].

The articulation rate for males and females is given in Figure 3.5. Observing the trends in the plot of relative change, at the bottom half of Figure 3.5, the articulation rate is generally in line with predictions: it decreases for most speakers by their 50s, with one or two exceptions. The largest outliers in this trend are Suzy (female) and Nick (male). These are both sourced from The Up Series, and the recordings in question are interviews. Their later interviews are more relaxed and conversational in nature relative to their earlier recordings, which are more formal. This is the probable cause of increased articulation rate into their 50s. The relatively stable articulation rate of the Queen is likely a result of her content being exclusively prepared speeches delivered in a controlled manner. However this finding is contrary to that of Brückl et al. [24] who found that the reading rate of females decreased with ageing. The Queen's articulation rate in her 20s, in the top right plot of Figure 3.5, is relatively low, so perhaps this rate is maintainable with ageing. General quality-related variability of Thatcher, Queen (70s+), Cronkite and Reagan is again observable in the absolute articulation rates, top half of Figure 3.5.

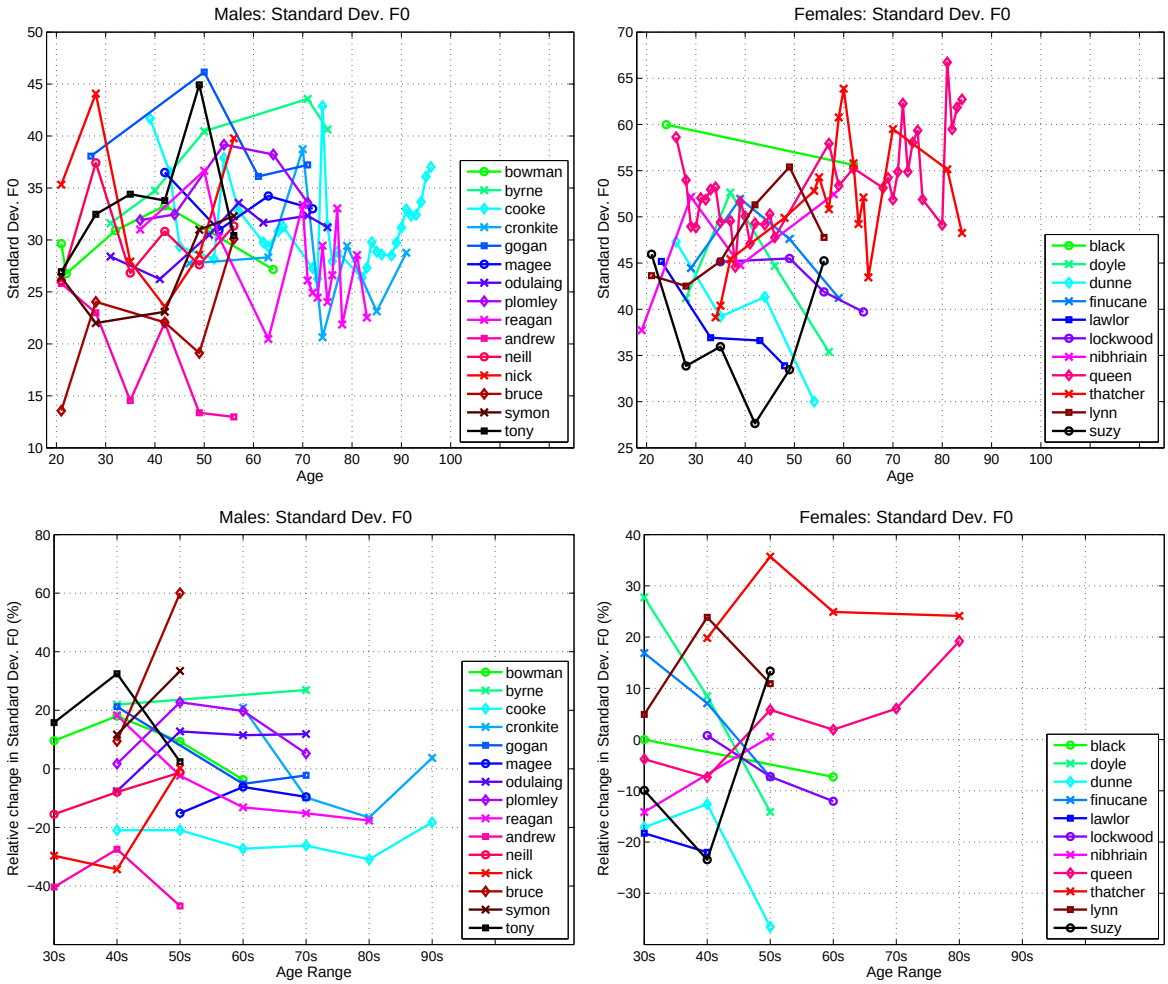

Figure 3.4: **Top:** Standard Deviation of F0. **Bottom:** Percentage change in standard deviation of F0, relative to a speakers youngest available sample, e.g. the y-axis value at Age Range = '50s' is the percentage difference between a speaker's std. dev. of F0 in their 50s and their std. dev. of F0 in their first decade of available data (their 20s or 30s in most cases). Note that there are different y-axis scales for males and females.

3.6.3 Jitter and Shimmer

The jitter and shimmer measures were extracted using Praat's local jitter and local shimmer routines. Local jitter is the percentage ratio of the average absolute difference between consecutive periods to the average period. Local shimmer is the percentage ratio of the average absolute difference between the amplitudes of consecutive periods to the average amplitude.

The local jitter rates for males and females are given in Figure 3.6. The findings are largely in line with expectations: jitter among males increases for the majority of subjects with age. The trend in females is less variable; all subjects' jitter values, with the exception of Lawlor, increase with ageing. Interestingly, two of the female smokers, Lockwood and Lynn, have the highest

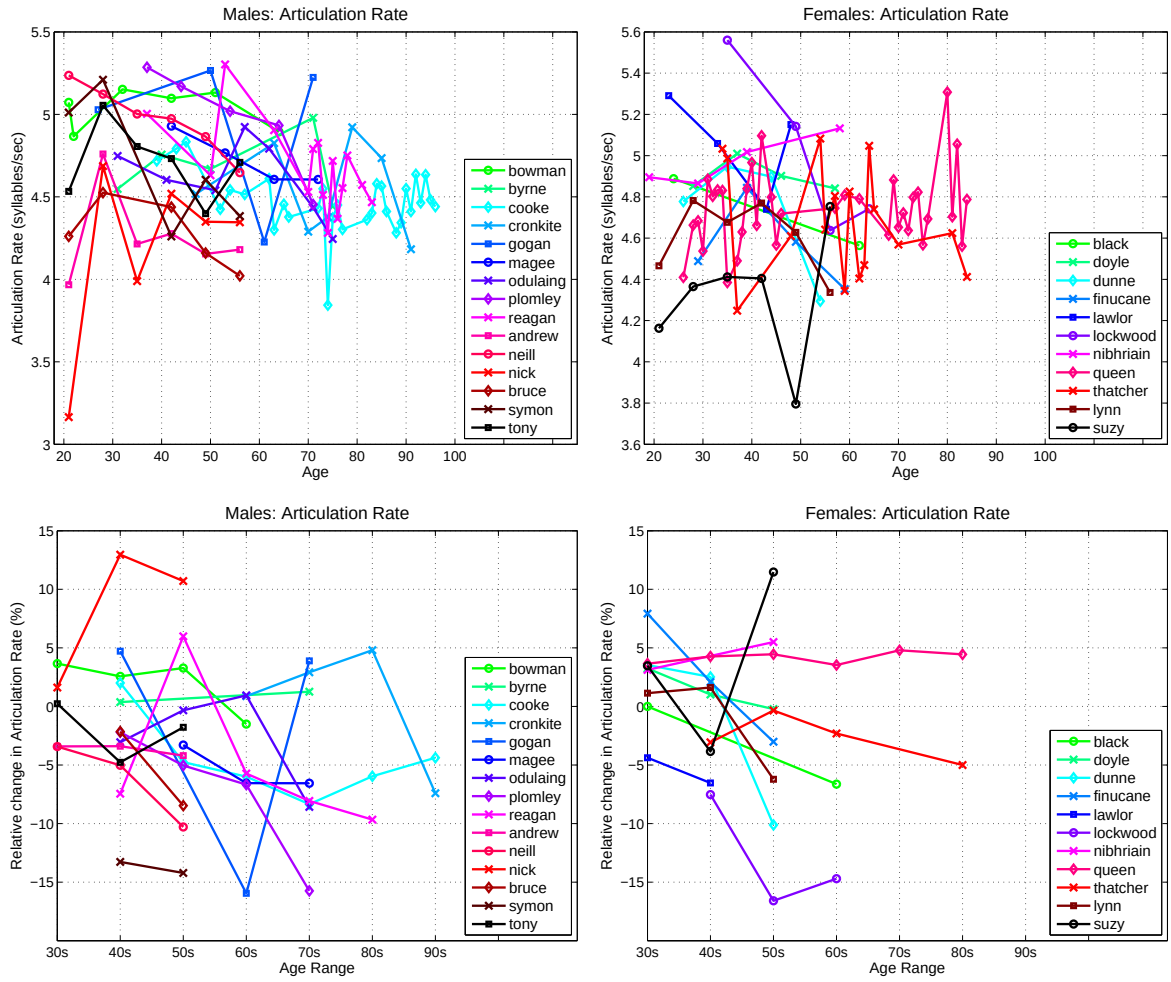

Figure 3.5: **Top:** Articulation Rate: the number of syllables divided by (silence-free) speaking time. **Bottom:** Percentage change in articulation rate, relative to a speakers youngest available sample, e.g. the y-axis value at Age Range = '50s' is the percentage difference between a speaker's articulation rate in their 50s and their articulation rate in their first decade of available data (their 20s or 30s in most cases). Note that there are different y-axis scales for males and females.

relative increase in jitter into their 50s, and also have the highest absolute jitter values (along with the non-smoker Suzy). In the males, Neill, who has ongoing mental health issues, has the sharpest rise in jitter, and the highest absolute jitter value, in his 50s. These individual cases add weight to the statements by Linville [131] and Mueller [144] that jitter is more indicative of general health than chronological age.

The local shimmer rates for males and females are given in Figure 3.7. The male trend is largely contrary to predictions: most speakers' shimmer values decrease by their 50s rather than increase. The relative shimmer change of females in their 50s is inconclusive, with some increasing and others decreasing. However, for the older speakers, there is a large increase in absolute shimmer into the 80s. There are again interesting individual trends: the smoker Lynn,

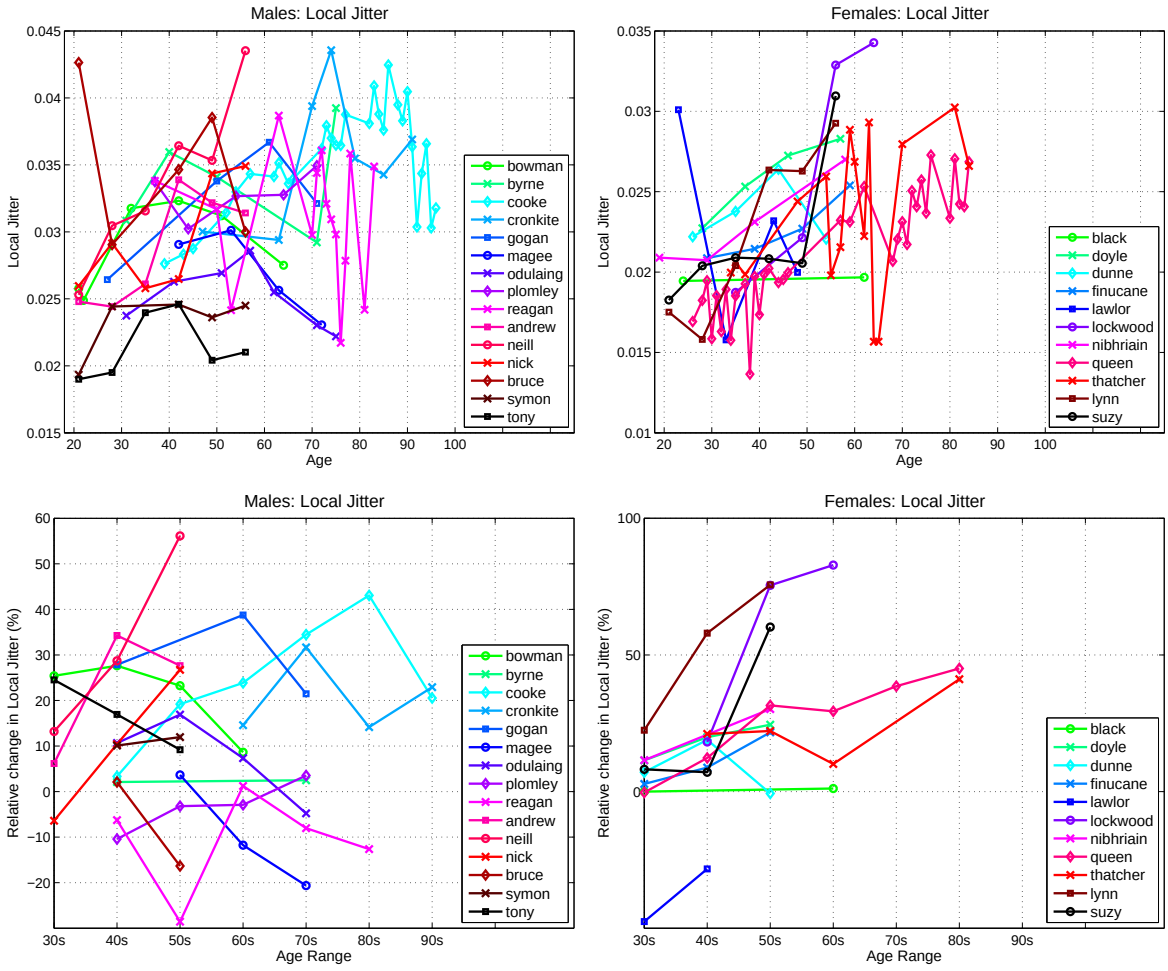

Figure 3.6: **Top:** Local jitter. **Bottom:** Percentage change in local jitter, relative to a speakers youngest available sample i.e, the y-axis value at '50s' is the percentage difference between a speaker's jitter in their 50s and their jitter in their first decade of available data (their 20s or 30s in most cases). Note that there are different y-axis scales for males and females.

has the highest relative increase in shimmer into her 40s, and increases again into her 50s. Neill has by far the highest absolute shimmer value among males, and the sharpest relative increase into his 50s.

3.6.4 Spectral Noise

Spectral noise was estimated as the Noise-to-Harmonics ratio (NHR). The NHR is computed as the ratio of the noise to the energy of the speech in the periodic part of the signal. The NHR values for males and females are given in Figure 3.8. There have been conflicting reports about the trends of NHR with ageing (Section 3.4.4). Here, the NHR trends of females are relatively consistent, with almost all increasing progressively with age. Male NHR is also quite

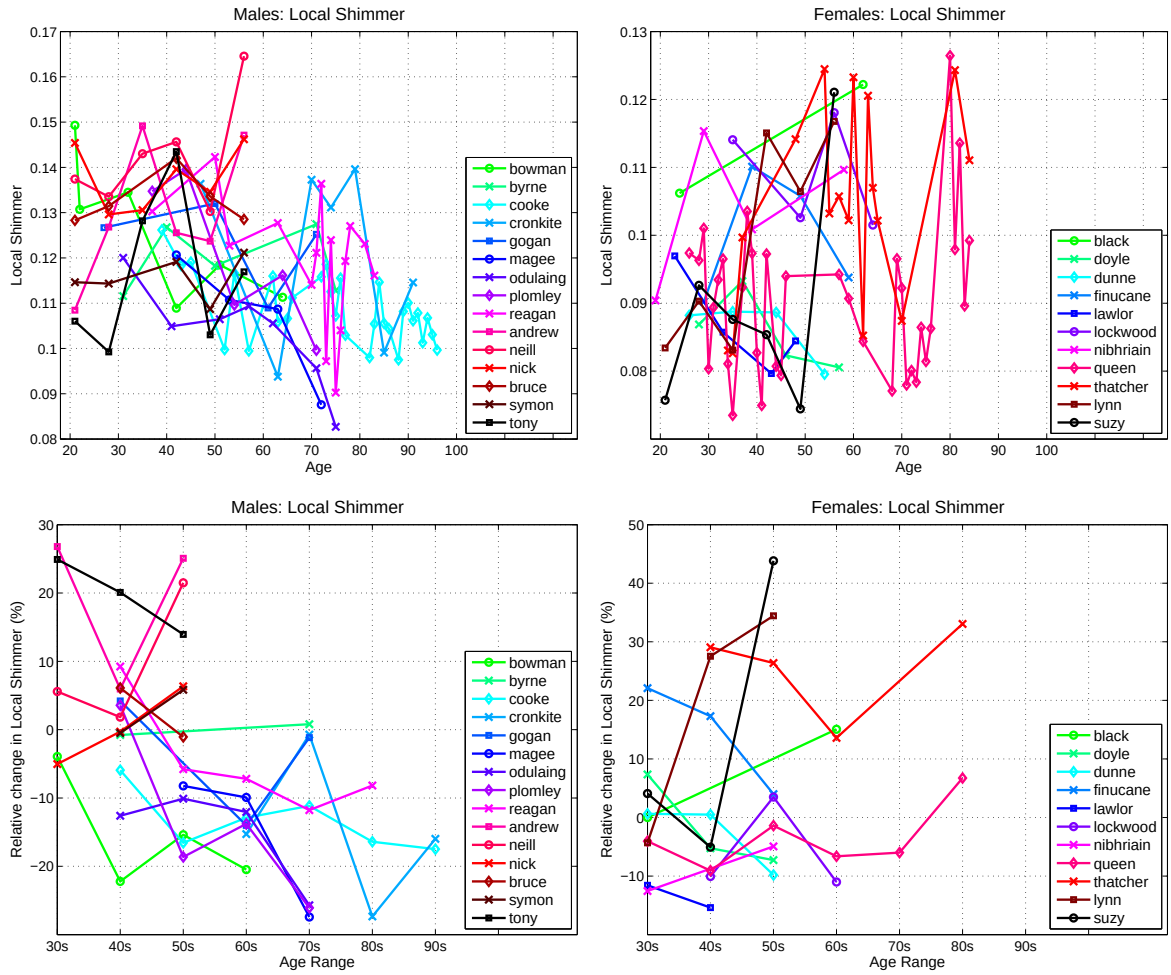

Figure 3.7: **Top:** Local shimmer. **Bottom:** Percentage change in local shimmer, relative to a speakers youngest available sample, e.g. the y-axis value at Age Range = '50s' is the percentage difference between a speaker's shimmer in their 50s and their jitter in their first decade of available data (their 20s or 30s in most cases). Note that there are different y-axis scales for males and females.

consistent, but in a decreasing direction however. This is an interesting finding, and has not been reported before - although there have been studies showing male NHR remaining stable [194] and female NHR increasing [201]. The smokers again display notable trends: Finucance, Lynn and Lockwood all have large relative increases in NHR into their 50s, and Lockwood has the highest absolute NHR value in her 50s. An interesting male trend is again with Neill, whose NHR increases progressively from his 30s to his 50s, going against the trend of most other males. He also has the highest absolute NHR of any speaker in their 50s.

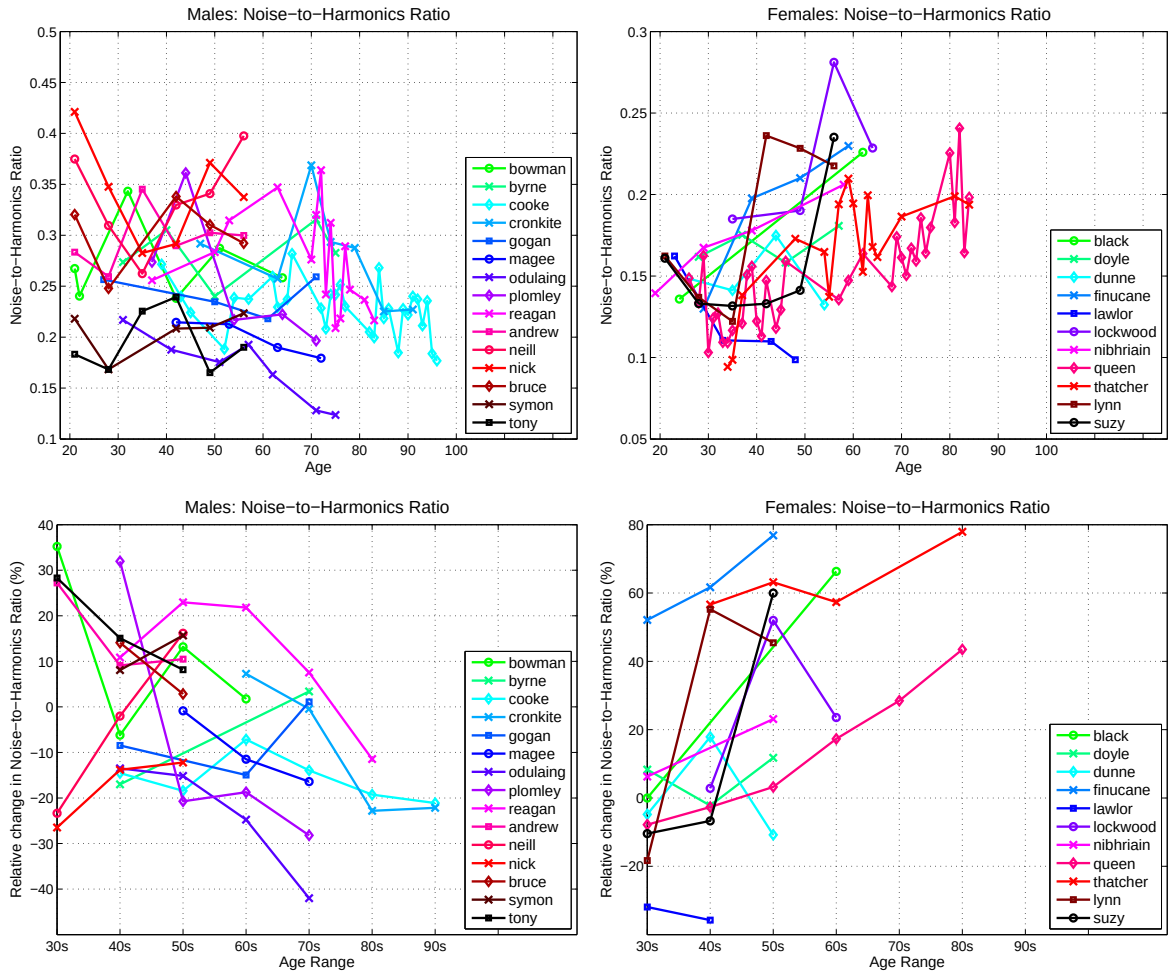

Figure 3.8: **Top:** Noise-to-Harmonics ratio (NHR). **Bottom:** Percentage change in NHR relative to a speakers youngest available sample, e.g. the y-axis value at Age Range = '50s' is the percentage difference between a speaker's NHR in their 50s and their NHR in their first decade of available data (their 20s or 30s in most cases). Note that there are different y-axis scales for males and females.

3.6.5 Summary of acoustic correlates of ageing experiments

The acoustic analysis of the TCDSA database speakers largely confirmed predictions for F0 change with ageing, while highlighting the variability of the 'turning point' of male F0. The standard deviation of F0 is not a reliable indicator of age, but shows more tendency to increase with age in females. Articulation rate decreases generally for ageing males and females, with variability likely due to speaking style in particular recordings, and read vs spontaneous speech. Local jitter aligns with expectations of a rising trend with ageing, in both males and females. Local shimmer observations are unclear: a general rising trend observed in females and a falling trend in males. Significantly, in both cases, speakers affected by smoking and health issues experience increased levels of jitter and shimmer increase. NHR ageing trends differ for males

and females: rising in females but falling in males. Smoking and health issues again have an clear influence on NHR.

This experiment validated several predictions regarding acoustics of the ageing voice, while also illustrating the variability in extent and onset of these acoustic changes. An interesting finding was the dramatic way in which smoking appears to accelerate the acoustic changes expected as a result of normal vocal ageing.

MFCC features, Section 2.1.1, capture frequency information and temporal information (via Δ and Δ^2 coefficients). The extent of the ageing-related acoustic changes observed here suggest that the MFCC feature distribution for a given speaker will be significantly affected by ageing. In the subsequent sections in this Chapter, the effect of ageing on a GMM-UBM speaker verification system utilising MFCC features is presented.

3.7 Effect of ageing on GMM-UBM speaker modelling

The first stage of investigation into the effects of ageing on Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification was to analyse the verification scores for a speaker whose model was trained and tested with recordings at different ages. Using the baseline GMM-UBM speaker verification system introduced in Section 2.2.1, and the TCDSA database, Section 3.5.1, experiments were designed to answer the following questions:

1. Does a speaker's verification score change significantly as their age in a test recording moves further away from their age in the training recording?
2. Does the age of the speaker at the time of model training influence the behaviour of verification scores?
3. How does any long-term ageing-related variability compare with inter- and intra-session variability?

The following subsections detail the GMM-UBM system and the experimental protocol used to address Questions 1 to 3. These experiments are largely documented in [110]. However, the results presented here have been supplemented with more recent additions to the TCDSA database and UBM database.

3.7.1 Pre-processing and feature extraction

The following pre-processing and feature extraction steps were applied to all recordings (see Section 2.1.2 for details): Energy-based silence removal and pre-emphasis were first applied. 12-dimensional MFCC vectors were then extracted over 20 ms windows with 50% overlap with 26 mel filters. RASTA filtering was applied to limit the influence of different channels. Delta and acceleration coefficients were then appended, resulting in a length 36 feature vector. Finally, mean and variance normalisation were applied.

3.7.2 GMM-UBM system

The first stage in developing a GMM-UBM system is creating a UBM (Section 2.2.1). To achieve a speaker-independent representation of the acoustic space, a large amount of data from a range of speakers is required for training the UBM [86, 175]. It is important that this data is balanced over sub-populations (e.g. male/female) in the data [163]. As described in Section 3.5.2, the TCDSA-UBM database was compiled to meet these requirements; it contains 60 minutes of speech from 60 males and 60 females, spread evenly across ages and accents (see Table A.2). To ensure a fully gender-balanced UBM, male and female (diagonal covariance) UBMs of 512 components were trained on a male-female split of the UBM database with 10 iterations of the Expectation-Maximisation (EM) [17] algorithm. A gender-independent UBM of 1024 components was then created by concatenating the component means and covariances from each of the gender-dependent UBMs, and halving their respective weights. This method of UBM training was previously applied by Reynolds et al. [163].

Speaker GMMs were then trained by adapting [163] the UBM component means given 30 seconds of speech. A relevance factor of 16 was used in the adaptation process. All GMM components (as opposed to a subset, e.g. 'top C scoring' [118, 163]) were considered in scoring. Adapting the GMMs from a UBM, as mentioned in Section 2.2.1, improves modelling by providing speaker-independent information not present in the training data. In the case of adapting from a gender-independent UBM, the cross-gender portion of the UBM may provide additional gender-independent information not in the same-gender portion of the UBM.

3.7.3 Long-term speaker verification score trends

In a standard speaker verification scenario, a user is first enrolled and then attempts to access the system via verification at a later date. In a forensic speaker recognition scenario, a comparison is made between a model for a suspected speaker and an evidential sample recorded some time *before* the model training. Although speaker verification is not suitable for the forensic evaluative mode, as discussed in Section 2.5, the underlying speech features and statistical models are shared across both approaches. With this in mind, an experiment was designed to answer Question 1 above, referring to the conventional and forensic scenarios as *forwards* and *backwards* verification respectively.

Two models were trained for each speaker, one using 30 seconds of speech from their earliest recording and the other using 30 seconds of speech from their most recent recording. For each speaker's N recordings, forwards verification scores were calculated by testing their model, trained with data from year 1, with recordings $[2, 3.., N]$. Backwards verification scores were calculated by testing their model, trained with data from year N, with recordings $[1, 2.., N-1]$. Each test involved computing the LLR (log-likelihood ratio) of a 30 second segment of a recording. Up to 5 segments per recording session were tested (dependent on data availability), and averaged to give one LLR per test year.

The resulting LLR scores are plotted against the time interval between training and testing in Figure 3.9. An initial assumption is made that the scores behave linearly with time and thus a linear least squares fit is plotted for each set of speaker scores. Despite some score variability, it is evident that there is a general and progressive decrease in LLR score as the age difference between training and testing increases.

If the slope of the line fit for each speaker is taken as an approximation of their rate of LLR score degradation, then the average slope across all speakers can be used to compare score degradation between genders and between forwards and backwards directions. The average slope in a forwards direction is -0.0078 for males and -0.0122 for females. In a backwards direction the average slope is 0.0087 for males and 0.0150 for females. Thus the LLR score degradation rate is consistent in forwards and backwards directions (As well as emulating a forensic scenario [63], backwards verification provides a form of reverse validation for the forwards direction [125]). LLR score degradation rate is almost twice as great in males than females however.

At low age difference values, there is significant variability between speaker LLRs. Considering the varied database content, and the fact no score normalization (e.g. Z-normalization, Section 2.2.1) was applied, this is to be expected.

There is general variability in the scores of speakers around the linear fit, particularly in the case of Thatcher, The Queen (both female) and Reagan (male). Variability was also observed in the acoustic measurement experiment for these speakers, Section 3.4, due to the varied nature of the environment and content of their recordings. A model-based measure of quality is introduced in Section 3.8.1 in an effort to filter the most 'acoustically different' recordings from the database.

3.7.4 Age-dependent long-term speaker verification score trends

As discussed in Section 3.3, and as observed from the acoustic measurements in Section 3.4 the rate of vocal ageing change generally increases in older age (50s-60s onwards). It would be expected then, that the trend of the LLR scores in Figure 3.9 would be dependent on the age of the speaker at training. To address this issue, raised in Question 2, the experiment in Section 3.7.3 was repeated, with models now trained for each speaker at multiple ages.

A model was trained for each speaker using data from year 1 and tested with recordings from year $[2, 3.., N]$. A new model was created with data from year 2 and tested with recordings from year $[3, 4.., N]$ and so on. The same testing protocol as in Section 3.7.3 was used to generate the LLR scores. Again, the assumption is made that the LLR score trend from each speaker model can be approximated linearly. The line fits for each score trend have been plotted in Figure 3.10 (the LLR scores have been omitted to aid visibility).

While there are some outliers, there is a general trend of increasing LLR degradation rate as age increases. An age of approximately 60 appears to be a turning point where the rate of LLR degradation accelerates. This behaviour aligns with expectations from the physiology of vocal ageing. For a clearer visualisation of this behaviour, the slope of each of these lines, from both

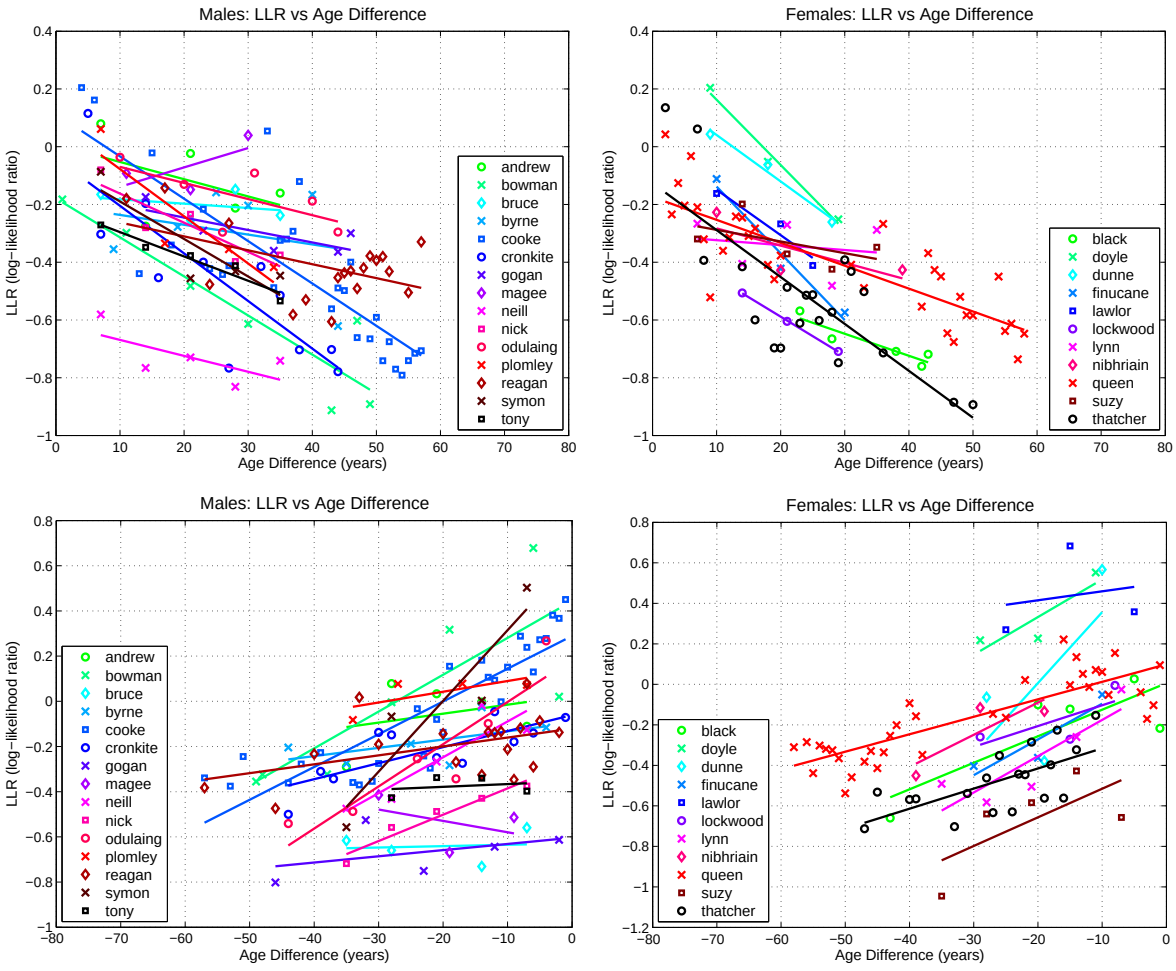

Figure 3.9: LLR scores produced by testing each speaker with his/her recordings as the age difference between training and testing increases. A linear fit is superimposed on each speaker's set of scores. **Top Left:** Males, Forwards verification, **Top Right:** Females, Forwards verification, **Bottom Left:** Males, Backwards verification, **Bottom Right:** Females, Backwards verification.

the male and female plots, Figure 3.10, is plotted in Figure 3.11. Exponential and quadratic fits have been superimposed on the slope values as possible models for the observed trend. The turning point at around age 60 is clearly indicated by both models.

3.7.5 Comparison of short-term and long-term session variability

The LLR scores in Figure 3.9 demonstrate a progressive decrease as the age difference between training and testing increases. This effect cannot be solely attributed to ageing however; short-term variability also causes degradation in speaker verification accuracy. As mentioned in Chapter 2, inter-session variability constitutes the biggest challenge facing speaker verification in general. Campbell et al. [34], discuss results from NIST SRE '05 showing a drop in verification accuracy over a period of one month, and conclude that inter-session variabilities

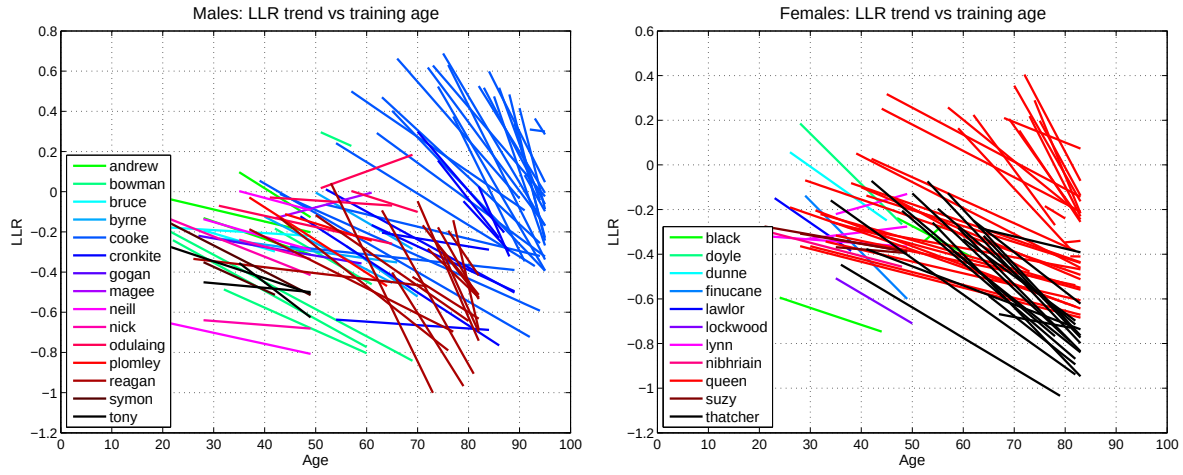

Figure 3.10: Each line represents a linear fit of the LLR scores produced when a speaker model at age A was tested with all their recordings at ages greater than A: $[A+1, A+2, ...]$ (forwards verification). The x-axis value at the beginning of each line indicates the age of the speaker at age of training (A), with the corresponding y-axis value given by the LLR of the first test sample, age $A+1$. **Left:** Male speakers, **Right:** Female speakers

other than ageing are responsible. Even over a longer time period of three years, the effect of ageing on verification scores has been shown to be negligible compared with typical inter-session variability [125].

It is reasonable to assume therefore, that the effect of ageing becomes apparent only when its influence on verification scores exceeds that of 'typical' inter-session variability. To investigate this, and address Question 3 above, an experiment was designed to compare short-term (less than one year) inter-session score variability with long-term (5-30 years) inter-session variability.

The amount of short-term inter-session data (recordings from different sessions within a given year) in the TCDSA database is very limited. Thus, the analysis was restricted to one speaker: Cooke (male). A model was first trained for each of Cooke's recording sessions with 30 seconds of speech.

Short-term inter-session verification scores were obtained by testing a model with 30 second segments from all other sessions from the same year. Long-term inter-session scores were obtained by testing a model with 30 second segments from all other sessions across all years. In addition, intra-session (within-session) scores were obtained by testing a model with 30 second segments from the same (training) session. The score distributions of these three sets of results are given in Figure 3.12. Note that the system configuration used for this experiment was that presented in [110], which differs slightly from the system described in Section 3.7.2, both in the composition of the UBM and the features used (no delta or acceleration coefficients). This is the reason for the difference in the range of Cooke's LLR scores in Figures 3.12 and 3.9.

As expected, intra-session scores have the highest LLR range, followed by short-term inter-

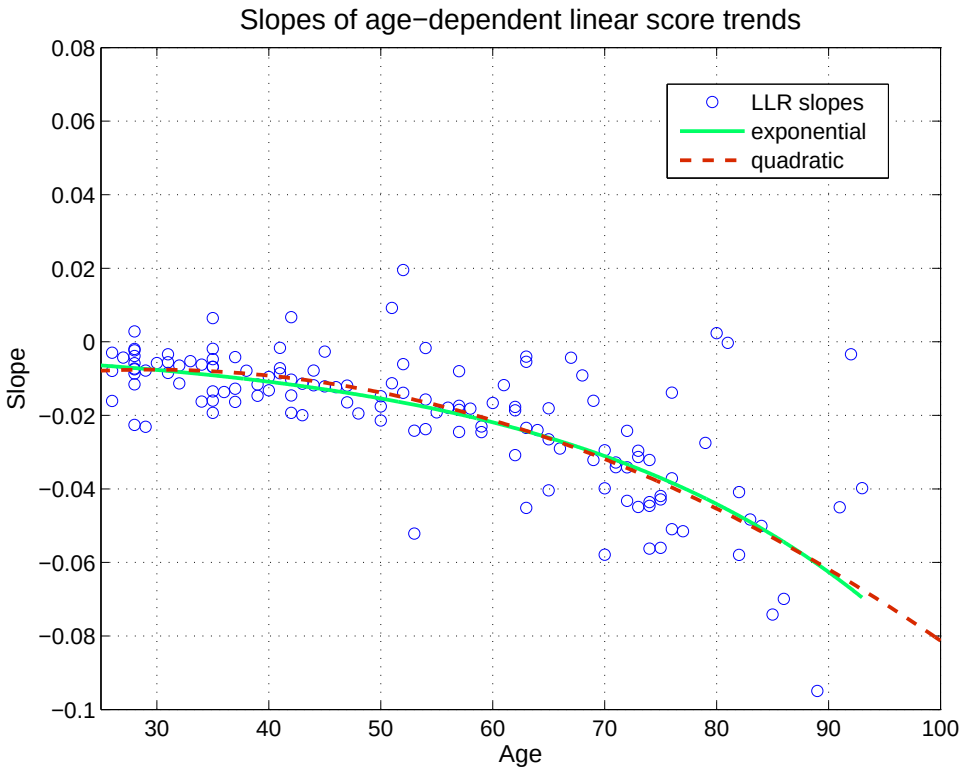

Figure 3.11: The slopes from male and female plots (forwards verification), Figure 3.10, combined and plotted against age. Quadratic and exponential fits have been added as possible models for the rate of ageing change.

session scores. Long-term inter-session scores at a time span of 5 years occupy a similar range to short-term inter-session results. This aligns with previous findings ([125]), which concluded the effect of ageing over three years is insignificant compared to typical inter-session variability. At time spans of 10, 20 and 30 years however, the verification score distribution shifts progressively downwards, beyond the range of the short-term inter-session score distribution. This suggests that at an age difference of greater than five years between training and testing, ageing-related variability exceeds that of typical inter-session variability.

3.8 Effect of ageing on GMM-UBM speaker verification

As discussed in Section 2.3, a decision threshold is determined for a system based on a trade-off between the misclassification rates of genuine-speaker and imposter scores. The experiments in Section 3.7 revealed that genuine-speaker scores decrease progressively as the age difference between training and testing increases. To evaluate the effect of ageing on speaker verification accuracy, imposter scores must be obtained. In this Section, an experiment involving long-term genuine-speaker and imposter scores is presented. The majority of this material was presented

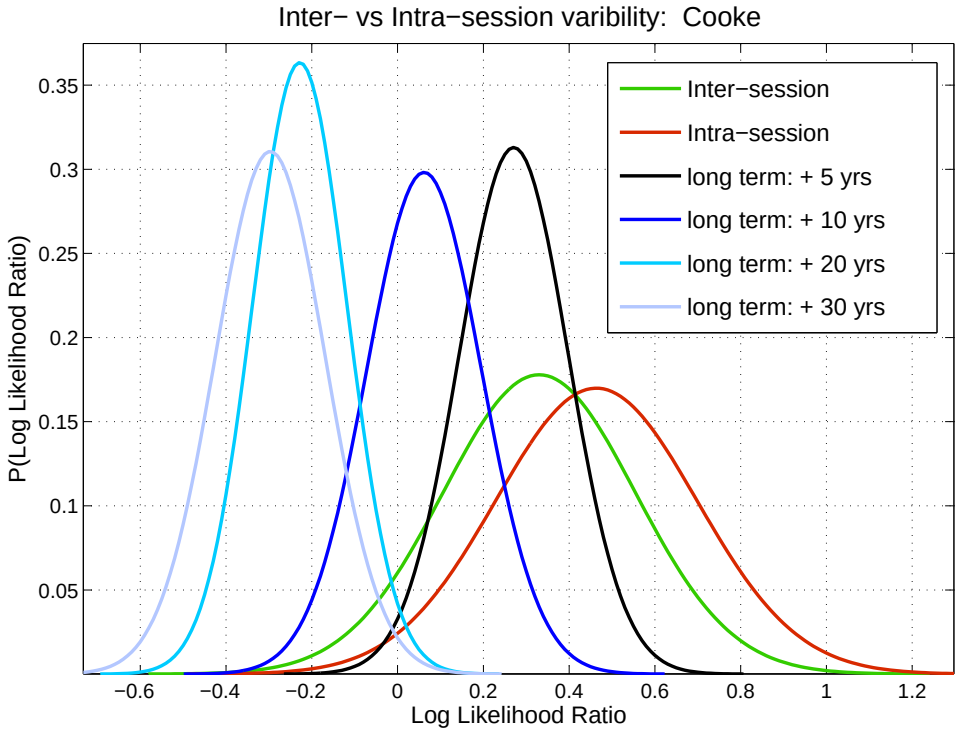

Figure 3.12: The distribution of LLR scores for Cooke, training and testing within session (intra-session), across sessions in the same year (inter-session), and across sessions in different years (long-term)

in [108, 109].

3.8.1 Restricting variability of TCDSA database recordings

In the compilation of the TCDSA database, detailed in Section 3.5.1.2, the aim was to limit non-ageing-related variability. However, the extent of the variability in acoustic measurement values and LLR scores (Sections 3.4 and 3.7), for a few speakers in particular, suggests that there remains unwanted variability in a portion of the database.

As an additional, objective measure of recording quality, a UBM-based measure was introduced. The likelihood of each recording in the TCDSA database, given an age-balanced UBM (Section 3.7.2) was used as a quality measure. This UBM likelihood approach has previously been applied for quality measurement by Harriero et al. [82]. The motivation for the measure can be understood by viewing the UBM as a representation of the acoustic features common to the speaker population. In the case of a degraded (or in some way acoustically different) test recording, its likelihood given the UBM will fall outside the range of scores from typical recordings. The notion of quality is addressed in depth in Section 4.2.

The UBM-based quality screening was applied to the database by dividing every recording into 30 second segments, extracting features as described in Section 3.7.1, and calculating the likelihood of each segment given the UBM. Any recording whose complete set of likelihood scores

fell outside 1.5 times the interquartile range of the set of all scores was deemed an outlier, and the associated recording was discarded from the database. A figure of 1.5 was chosen based on its use as a standard outlier measure in boxplots [64]. The assumption was that the set of all 'acceptable' UBM scores would fall within the limits of this range. Approximately 10% of the total number of recordings were discarded with this method. A schematic of the TCDSA database, with the removed recordings greyed-out, is given in Figure 3.13. All of the removed recordings are YouTube-sourced, and are all from the speakers with most variability in acoustic measures and LLR score.

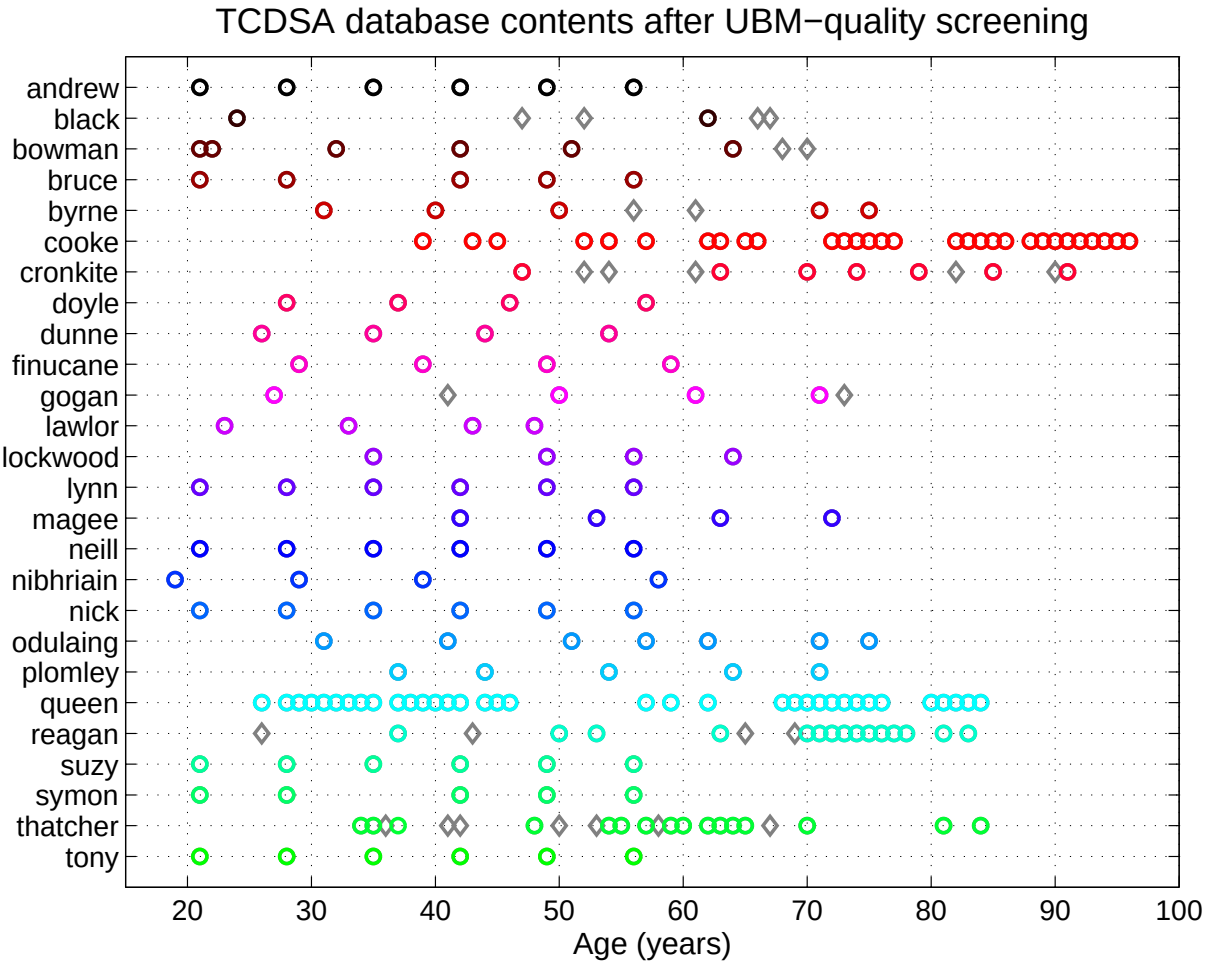

Figure 3.13: A schematic of the TCDSA database after a UBM-quality screening procedure was applied. Each circle indicates a year where a recording is available for a speaker. Grey diamonds indicate the recordings that have been removed. Speaker names are shown on the y-axis, with their age in each recording given on the x-axis

3.8.2 Long-term speaker verification evaluation

An evaluation of the TCDSA database with the GMM-UBM system was designed. A testing protocol similar to the evaluation in Section 3.7.3 was used, with the addition of imposter testing. A earlier version of the (UBM-quality screened) TCDSA database was used in the experiments in this Section; none of the eight Up Series speakers (Andrew, Bruce, Lynn, Neill, Nick, Suzy, Symon, Tony) were included, leaving nine males and nine females. The pre-processing, front-end and GMM-UBM parameters were the same as those in Sections 3.7.1 and 3.7.2.

3.8.2.1 Evaluation protocol

Two GMMs were trained for each speaker, one using 60 seconds of speech from their earliest recording and the other using 60 seconds of speech from their most recent recording. Each test score was based on the average of up to 10 (dependent on data availability) LLR scores from 30 second segments of a given recording.

In a *forwards* direction, for each speaker's N recordings, genuine-speaker scores were calculated by testing their model, trained with data from year 1, with recordings $[1, 2.., N]$. The set of all other (17) speakers were used as imposters. Imposter scores were calculated by testing the year 1 speaker model with all imposter recordings occurring *after* the year of the training recording.

In a *backwards* direction, for each speaker's N recordings, genuine-speaker scores were calculated by testing their model, trained with data from year N, with recordings $[1, 2.., N]$. The set of all other (17) speakers were used as imposters. Imposter scores were calculated by testing the year N speaker model with all imposter recordings occurring *before* the year of the training recording.

This protocol includes cross-gender trials (where 'a trial' is the calculation of the LLR of a recording given a model and UBM). Although cross-gender trials usually produce lower LLR scores than same-gender trials (and are excluded from NIST-SRE evaluations for that reason), they were included here given the limited number of same-gender imposters. Furthermore, it is not of concern if the error rates are lowered by the inclusion of cross-gender trials, as it is only the relative difference in error rates across different time intervals that is of interest in this investigation.

This protocol also includes same-session testing, i.e. a speaker's year 1 model tested with data from the same year 1 recording. This was necessary in order to obtain LLR scores suitable for setting a decision threshold for the system at the time of enrolment, i.e. a conventional decision threshold, Section 2.3. In these same-session tests, separate portions of the training recording were used for training and testing.

3.8.2.2 Z-normalization

Zero normalization, usually referred to as 'Z-norm', is a form of score normalization (Section 2.2.1). The purpose of Z-norm is to transform the verification scores from different speakers into a common range, enabling a speaker-independent threshold to be set [118]. To apply Z-norm, the statistics (mean and standard deviation) of the imposter score distribution are first estimated for each speaker. For each speaker in this experiment, all 17 other speakers were considered imposters. The set of 17 imposters was divided into two groups of 9 and 8. The speakers for each group were selected randomly, but kept balanced in terms of gender. Then, for each speaker model, Z-norm statistics were calculated from the imposter set of 9. These statistics were used to Z-normalize the scores of the other 8 imposters, along with the genuine-speaker scores from the speaker in question (as in Equation 2.10).

Finally, given that there were a variable number of data years available per speaker/imposter, the number of recordings from the 8 test imposters for which scores were included was limited to a random subset of 5 (in order to avoid biasing towards imposters with more data).

3.8.2.3 Long-term LLR trends

The resulting LLR scores (after Z-norm) for all 18 genuine-speakers and their corresponding imposters are plotted against the time interval between training and testing in Figure 3.14, in both forwards and backwards directions. Examples of individual speaker LLR scores for Cooke (male) and The Queen (female) are given in Figure 3.15.

It can be observed from the overall LLR trends in Figure 3.14 that there is a general and progressive decrease in LLR score for genuine-speakers as the time lapse between enrolment and verification increases. Imposter scores however, are relatively stable over the same time period. Assuming that recording conditions are unchanging, this behaviour is expected. Experimental findings in the domain of face ageing [56] are consistent with this finding. A possible reason for the slight drop in imposter LLR scores is that, as the time lapse between enrolment and verification increases, the age difference between the average imposter and the test speaker increases. As demonstrated in [50], there is greater separation between genuine-speaker and imposter verification scores as the age difference between them grows.

The individual speaker examples in Figure 3.15 demonstrate the same trend as the global case. The variability in the scores is also apparent in these examples. It is clear that a fixed threshold determined at the time of enrolment (e.g. where Age difference = 0, Figure 3.15) will result in an increasing verification error rate as time progresses.

In Figure 3.16, a DET (detection error tradeoff) plot (Section 2.3) was generated for both forwards and backwards verification cases. This illustrates the discrimination ability of the system given the genuine-speaker and imposter LLR scores in Figure 3.14.

The EER of the GMM-UBM system over an age difference of 60 years, as indicated in Figure 3.16, is approximately 18% in both directions. This is a much higher error rate than what is

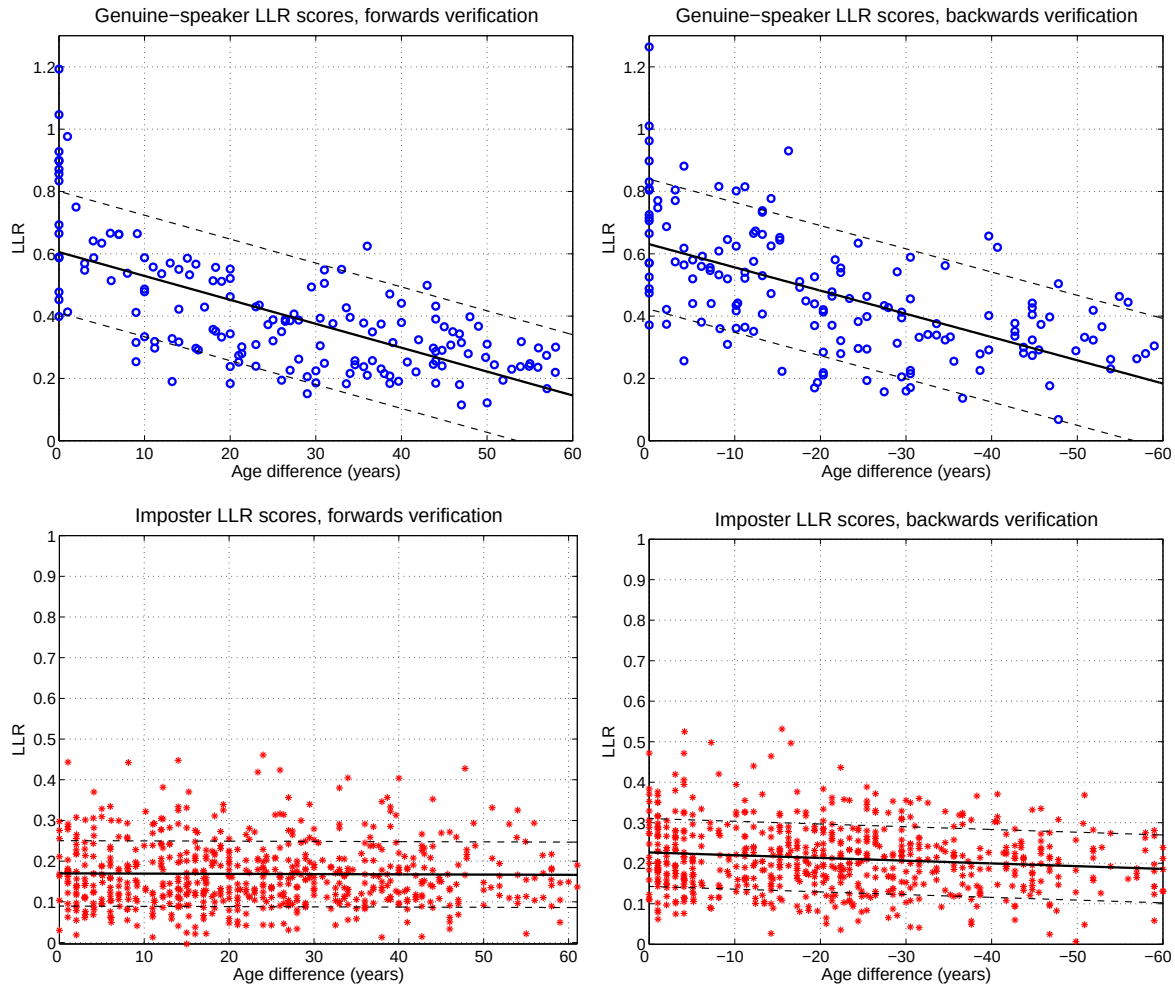

Figure 3.14: LLR scores against the time interval (age difference) between training and testing for all 18 speakers. A linear fit of the data is superimposed on each set of scores. The dashed lines denote the ± standard deviation around the linear trend. **Top Left:** Genuine-speakers, Forwards verification, **Top Right:** Genuine-speakers, Backwards verification, **Bottom Left:** Imposters, Forwards verification, **Bottom Right:** Imposters, Backwards verification.

typically obtained with a GMM-UBM system given contemporaneous data. For example, a benchmark EER of 8.45% for a GMM-UBM system on the NIST 2008 telephone-telephone task is quoted in [118].

The EER represents the performance if an *ideal* decision threshold was set (assuming both types of error, false acceptances and rejections, are of equal cost). In Section 3.8.3, the performance is evaluated with a decision threshold determined from an independent set of scores, resulting in a more realistic performance assessment.

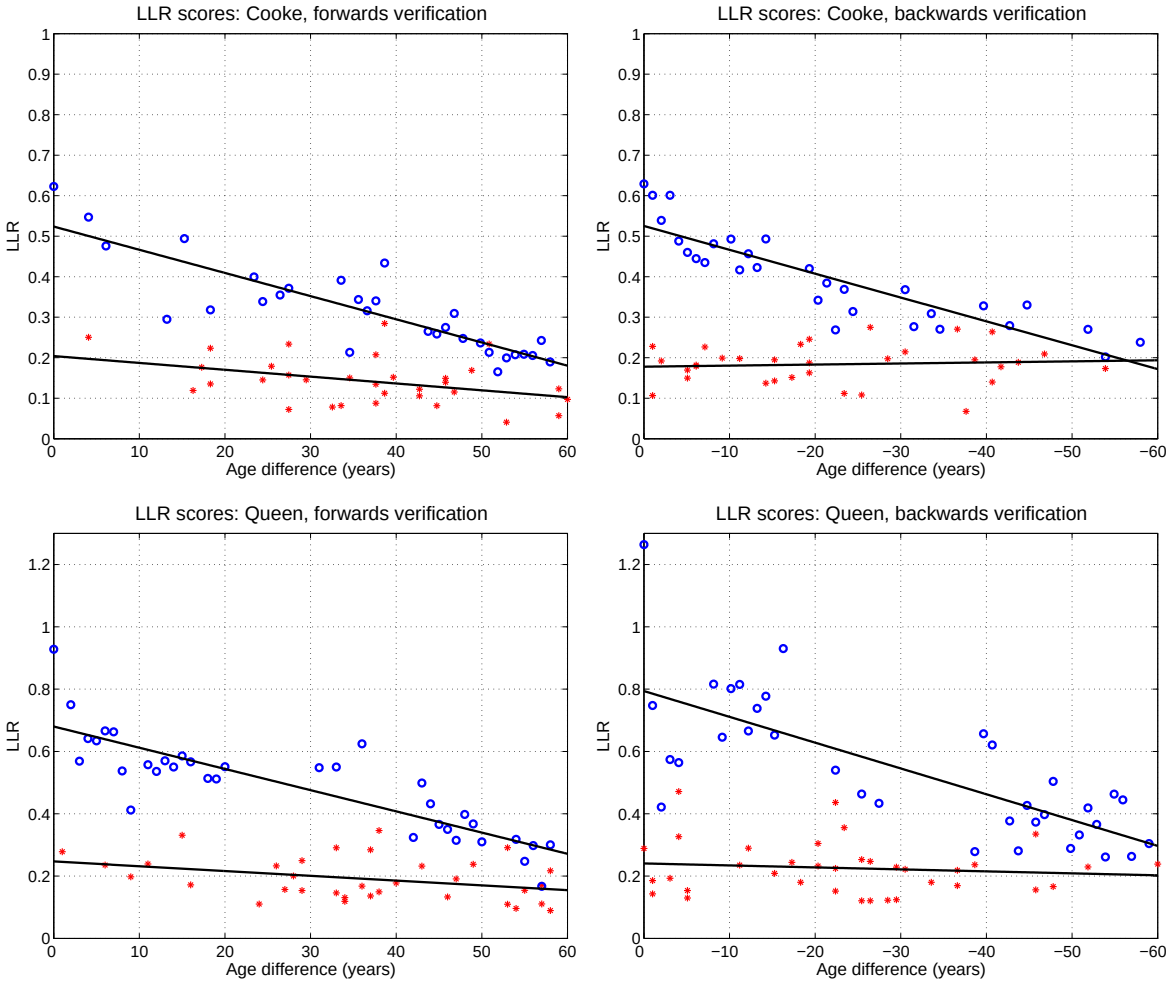

Figure 3.15: LLR scores against the time interval (age difference) between training and testing for an example male and female speaker. Blue circles denote genuine-speaker scores. Red asterisks denote imposter scores. A linear fit of the data is superimposed on each set of scores. **Top Left:** Cooke (male), Forwards verification, **Top Right:** Cooke (male), Backwards verification, **Bottom Left:** Queen (female), Forwards verification, **Bottom Right:** Queen (female), Backwards verification.

3.8.3 Long-term speaker verification performance

The baseline performance of the GMM-UBM system was evaluated on a per-speaker basis. For each of the 18 speakers, a decision threshold was calculated by pooling all genuine-speaker and training imposter scores corresponding to an age difference of zero. A threshold was determined such that the HTER (half total error rate) was minimised. The HTER is the average of the FAR and the FRR. This is the typical means of obtaining a decision threshold for a GMM-UBM system at the time of enrolment.

The decision threshold was applied to the scores of the test speaker and their test imposters, over all age differences. Any scores falling below the threshold were rejected and any above were

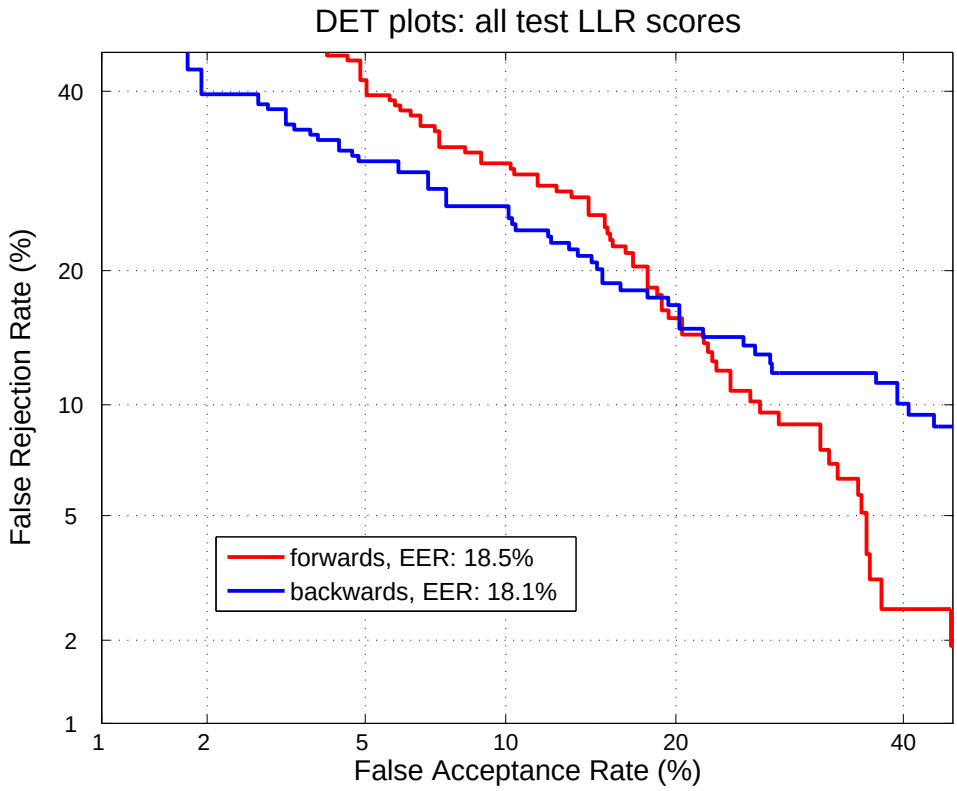

Figure 3.16: DET (detection error tradeoff) plots for the pooled sets of forwards and backwards test scores in Figure 3.14.

accepted. A HTER was obtained for each of the 18 speakers in this way. The average HTER in forwards and backwards directions is given in Table 3.1 for a range of age differences. The increase in HTER as the age difference between training and testing increases is evident from these results.

	HTER(%)						
Age Difference (years)	5	10	20	30	40	50	60
Forwards	2.45	6.12	13.62	18.58	21.07	23.43	24.98
Backwards	7.15	9.39	14.36	17.50	21.09	22.44	22.77
Mean	4.80	7.74	13.99	18.04	21.08	22.94	23.88

Table 3.1: Average HTER (half total error rate) of 18 speakers over increasing time lapses between training and testing. The HTER for forwards and backwards verification cases, and their mean, are indicated.

Individual speaker examples for Cooke (male) and Queen (female) are shown in Figure 3.17. The HTER for the full 60 year age difference is included on each plot. It is clear that the decision threshold, while providing good discrimination at the year of enrolment, leads to an increasing

amount of classification error as ageing progresses.

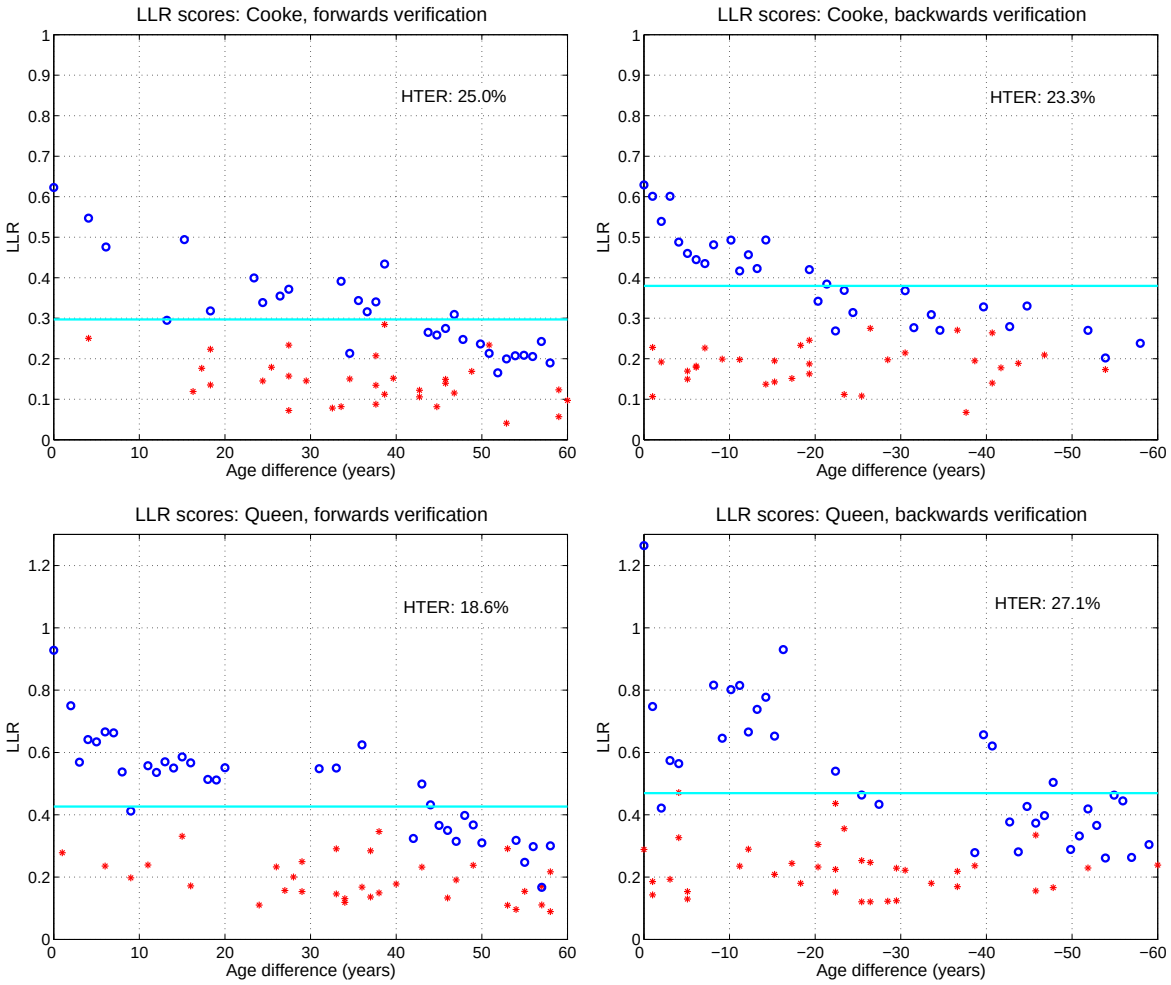

Figure 3.17: LLR scores against the time interval (age difference) between training and testing for an example male and female speaker. Blue circles denote genuine-speaker scores. Red asterisks denote imposter scores. The cyan line denotes the decision threshold, Section 3.8.3. **Top Left:** Cooke (male), Forwards verification, **Top Right:** Cooke (male), Backwards verification, **Bottom Left:** Queen (female), Forwards verification, **Bottom right:** Queen (male), Backwards verification.

3.9 Discussion

The TCDSA database introduced in this chapter is, at the time of writing, the largest (in terms of number of subjects, longitudinal range, and number of recordings) longitudinal speech database in the public domain. The subsequent experiments therefore offer a new level of insight into the effects of vocal ageing.

While largely in agreement with the literature, the analysis of the acoustic correlates of ageing present some interesting observations. Firstly, there appears to be more uniformity

in the ageing trends of females than males: Female F0, articulation rate, local jitter and NHR trends are more consistent between speakers than the corresponding male trends. It is presumed that this is as a result of the differences in physiology and ageing patterns between males and females. Another interesting finding was a noticeable increase in local jitter, local shimmer and NHR for speakers who smoked or had health problems. The general variability observed between the acoustic measurements of different speakers emphasises the recurring point raised in physiological studies, that vocal ageing is an individual process.

The speaker verification experiments demonstrate the significance of the ageing problem in a conventional speaker verification framework. Error rates progressively increase as the age difference between training and testing increases. Ageing variability appears to exceed typical inter-session variability after around 5 years, and the rate of score degradation increases above the age of 60, tying in with expectations from acoustic measurements. These findings motivate the development of specific strategies to improve speaker verification performance in the presence of ageing.

An existing approach for dealing with the ageing problem is model adaptation [61], whereby the parameters of a speaker model are updated using data from the speaker some time after enrolment. This allows the model to respond to changes in the speaker's voice over time. If the adaptation is based on speaker data from a verification attempt, a threshold on the verification score can be set, below which adaptation will not be executed. This is a means to prevent imposter attacks that operate by iteratively shifting the model parameters away from the speaker and towards the imposter. Even with such a threshold, a security weakness remains however - it may just take more verification attempts before adaptation is successful. Adaptation based on verification data would also increase the effectiveness of iterative 'hill-climbing' attacks, which have been used successfully in the domain of fingerprint [191] and signature [66] verification.

The alternative, more secure adaptation approach of re-enrolling the speaker at regular intervals is not a practical solution. In a large system this would be logistically difficult, and would result in a operating state where some speakers may have missed re-enrolment sessions, potentially making score normalization and/or calibration problematic.

A more favourable approach is a system that automatically adapts in response to ageing. In the subsequent Chapters 4 and 5, two novel methods of ageing-variability compensation for speaker verification are presented.

4

Stacked Classification for ageing speaker verification

In Chapter 3, a speaker verification experiment using the TCDSA database demonstrated a common tendency for genuine-speaker scores to degrade as the age difference between training and testing increased. Applying a standard decision threshold, optimised from the scores at the age of training and fixed for all subsequent ages, resulted in increasing verification error rates with increasing age difference. Based on this observation, a logical extension of the system is to employ a decision threshold that is ageing-dependent.

Such an approach could be viewed as a two-stage classification process; the first being to determine a verification *score* given the GMM-UBM system, and the second being to output a verification *decision* given the ageing-dependent threshold. A multi-stage approach to classification can be incorporated in a 'stacked classifier' framework.

Stacked classification (or stacked generalization) [188, 200], is an approach to classification whereby the outputs of two or more lower-level classifiers are used as input to a higher-level classifier. The goal is to improve classification performance by effectively combining the lower-level 'evidence'. Stacked classification has been proposed as a framework for biometric verification in general, as well as a solution to the specific problem of ageing in face verification.

Kryszczuk and Drygajlo [119] present a stacked classifier framework as a general solution to uni- and multi-modal biometric verification (a multi-modal system uses multiple biometric modalities to identify an individual, e.g. fingerprints and iris images). They present an effective framework for verification that combines the output from one or more biometric classifiers with one or more quality measures as lower-level inputs to a higher-level classifier. Here, a 'quality measure' is any metric extracted from the biometric sample that is a predictor of the output

score of the biometric classifier, but does not carry information as to the identity of an individual. An example of a quality measure for speaker verification is SNR (signal-to-noise ratio) [68]. In the experiments in Section 3.8.2, ageing information, in the form of the age difference between training and testing, could also be viewed as a quality measure. The notion of quality is dealt with in depth in Section 4.2.

Stacked classification has been applied to improve face verification performance in the presence of ageing. In this scenario, similar to the case of speaker verification, a degradation in verification accuracy is observed as the time interval between training and testing increases [54, 104, 150]. In several studies by Drygajlo, Li et al. [54–56, 128], different stacked classifier configurations are applied to the problem of ageing in face verification. The common approach across these studies is to combine scores from a standard face-image classifier with ageing information, in the form of the number of days elapsed between training and testing, as lower-level inputs to a higher-level classifier. Additionally, the scores from an second face-image classifier [129] or a quality measure (head rotation angle and deviation from an 'average frontal face') [55] are incorporated into the framework. The higher-level classifier used is a discriminative SVM (Support Vector Machine) [43] classifier. This framework consistently improves on the baseline long-term face verification performance ('long-term', in this instance, is a period of up to four years).

This research in the face verification domain motivates the use of a stacked classifier framework for dealing with the challenge of ageing in speaker verification. In this Chapter, a stacked classifier framework, combining scores from the GMM-UBM system with ageing and quality information is presented as a solution to the challenge of long-term ageing in speaker verification. The following sections describe work presented in [108]: combining GMM-UBM scores and ageing information, and [107, 109]: combining GMM-UBM scores, ageing information and quality measures.

4.1 Stacked Classifier for speaker verification

A schematic of the proposed stacked classifier framework is shown in Figure 4.1. Given a test speech sample, the GMM-UBM system is used to compute an LLR score, as in Section 3.7.2. The age difference is determined from the meta data of the recordings, and is defined as the difference (in years) between the recording date of the test speech sample and the training speech sample (i.e. the sample used to train the model under test). A number of quality measures are extracted from the sample; these will be detailed in Section 4.2. After a Z-normalization of the LLR score is applied, the score, ageing and quality values are concatenated into a new feature vector. In the stacked classifier scheme, the vector contains the lower-level information. This test feature vector is scaled to the range $[0, ..., 1]$ and input to the higher-level SVM classifier. The SVM, trained on a set of development feature vectors, outputs a accept/reject decision for the test feature vector.

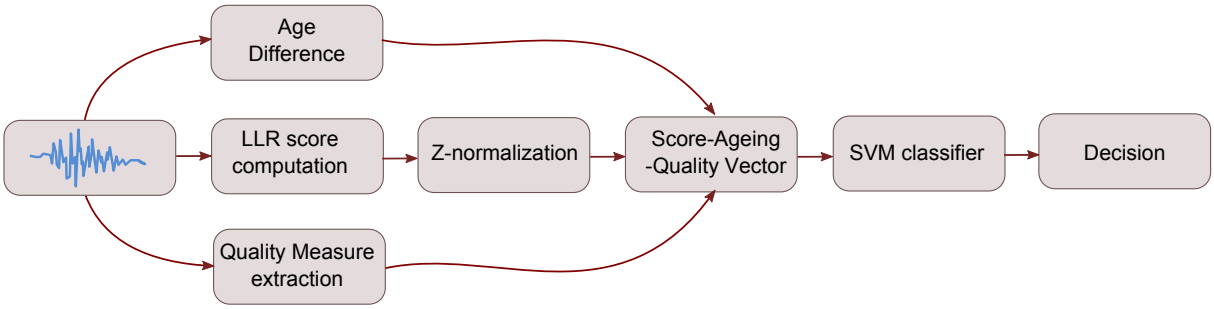

Figure 4.1: A schematic of the stacked classifier framework: The LLR score from the GMM-UBM system, ageing information and quality measures are combined as an input to an SVM classifer.

The details of the implementation and evaluation of this scheme are detailed in Section 4.1.1. In this Section, a lower-level feature vector of LLR score and age difference only will be considered. In the subsequent Section 4.3, the feature vector will be expanded to include quality measures.

4.1.1 Score-Ageing Stacked Classifier Evaluation

An evaluation of the score-ageing stacked classifier framework using the TCDSA database was designed. To create the lower-level feature vectors, the LLR scores from the GMM-UBM experiment in Section 3.8.2.3 were used. For each LLR score, the corresponding age difference was calculated. In this way a set of two-dimensional score-ageing vectors were created for each genuine-speaker and imposter trial of the 18 TCDSA database subjects.

4.1.1.1 Higher-level SVM classifier

Given a set of training score-ageing vectors (the training/testing protocol is described in Section 4.1.1.2), a scaling of the vectors to the range $[0, ..., 1]$ was applied based on the extreme values in the training set. Scaling (or normalization) is an important step in preparing data for input into an SVM classifier [98, 197], as it avoids attributes in large numerical ranges dominating those in smaller ranges. A linear SVM classifier was trained with the scaled data. The SVM training maximises the linear classification margin between the two classes (genuine-speaker and imposter) in score-ageing space such that the HTER on this training set is minimised. The *LIBSVM* [39] package was used for this purpose. A linear kernel was chosen for the SVM due to the limited amount of training data. A non-linear kernel has the potential to provide better discrimination, as it can capture non-linear dependencies between the elements of the feature vector. However in a face verification application, Drygajlo and Li found the performance gain by using a non-linear RBF (radial basis function) kernel over a linear kernel to be small [54]. The trained SVM boundary was then used to classify test data in score-ageing space.

	HTER(%)						
Age Difference (years)	5	10	20	30	40	50	60
GMM-UBM: Forwards	2.45	6.12	13.62	18.58	21.07	23.43	24.98
SC: Forwards	3.10	5.84	9.81	11.64	13.95	14.15	15.41
GMM-UBM: Backwards	7.15	9.39	14.36	17.50	21.09	22.44	22.77
SC: Backwards	7.50	6.81	11.27	11.81	15.8	16.47	16.90
GMM-UBM: Mean	4.80	7.74	13.99	18.04	21.08	22.94	23.88
SC: Mean	5.30	6.33	10.54	11.73	14.87	15.31	16.16

Table 4.1: Average HTER (half total error rate) of 18 speakers over increasing time lapses between training and testing. The HTERs for the ageing-score stacked classifier (SC) and the GMM-UBM baseline system are shown for forwards and backwards verification cases, as well as the mean of both directions.

4.1.1.2 Ageing Stacked Classifier experiment design

A cross-validation approach was taken for training and testing the score-ageing stacked classifier. For each of the 18 speakers in turn, the remaining 17 speakers were used for training. All of the genuine-speaker and imposter scores for the training set were converted to scaled score-ageing vectors and used to train the SVM classifier. The HTER for the test speaker was then evaluated by using the trained SVM to classify the test score-ageing vector. As described in Section 3.8.2.2, two independent sets (of 9 and 8) speakers were used as training and testing imposters. This scheme resulted in a stacked classifier evaluation for each of the 18 subjects, and hence 18 HTERs. These were averaged to provide an overall HTER, comparable with the baseline GMM-UBM approach, Section 3.8.3.

4.1.1.3 Ageing Stacked Classifier performance

The average HTER in forwards and backwards directions is given in Table 4.1 for a range of age differences. The baseline GMM-UBM HTERs, Table 3.1, are reproduced in Table 4.1 for comparison. An increase in HTER as the age difference between training and testing increases is observed for both systems. However, the score-ageing stacked classifier provides significantly lower error rates than the GMM-UBM baseline. This is most noticeable at the extreme age difference of 60 years, where the relative improvement in mean HTER offered by the stacked classifier is 32%. However, at an age difference of 5 years, the stacked classifier performs slightly worse than the baseline, in both directions.

Training LLRs for two example speakers, Cooke (male) and Queen (female), in forwards and backwards directions, are plotted against age difference in Figure 4.2. The score-ageing decision threshold, trained on this data, is superimposed on each plot. It is clear that these thresholds track the genuine-speaker score trend far more closely than a one-dimensional decision threshold

determined at enrolment. The application of these thresholds to the test data of the two example speakers is shown in Figure 4.3. The standard GMM-UBM decision boundary is included for comparison. The HTER determined with the score-ageing decision threshold for the test speaker is indicated on each plot. Compared with the corresponding baseline HTERs for these example speakers, shown in Figure 3.17, the stacked classifier approach reduces the individual speaker HTERs by 5.4 - 15.8% in absolute terms.

From the training examples, Figure 4.2, it is evident that the stacked classifier threshold performs slightly worse than the baseline at age differences of less than 5 years due to a greater number of false acceptances over this interval. Given the constraint of a linear threshold, there appears to be a trade-off between short-term and long-term classification performance.

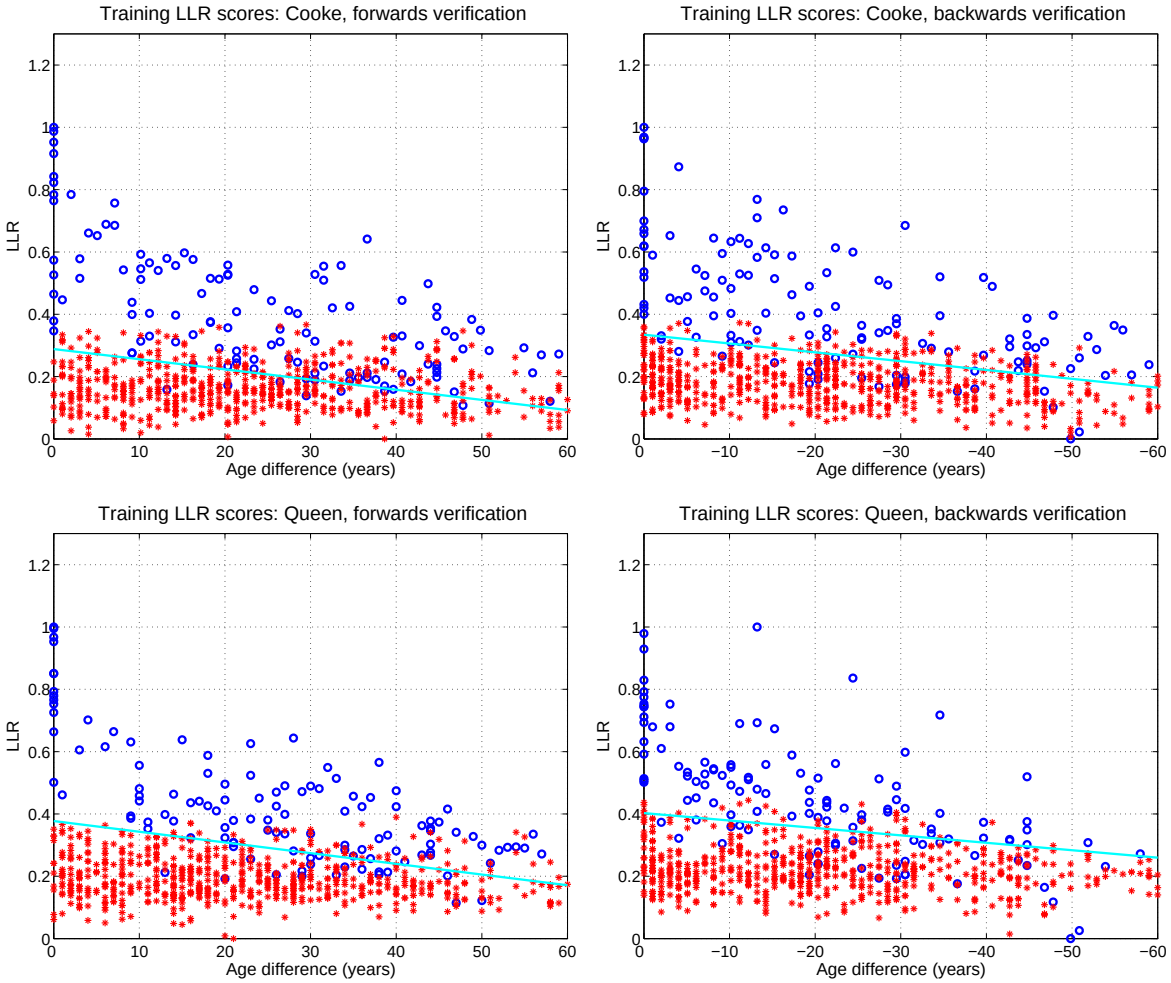

Figure 4.2: LLR scores used to train a decision threshold for Cooke (male) and Queen (female), consisting of the genuine-speaker scores for all *other* 17 speakers, and all scores from 9 training imposters. The cyan line denotes the score-ageing dependent decision boundary that minimises the HTER on the training set. **Top Left:** Cooke, Forwards, **Top Right:** Cooke, Backwards, **Bottom Left:** Queen, Forwards, **Bottom right:** Queen, Backwards.

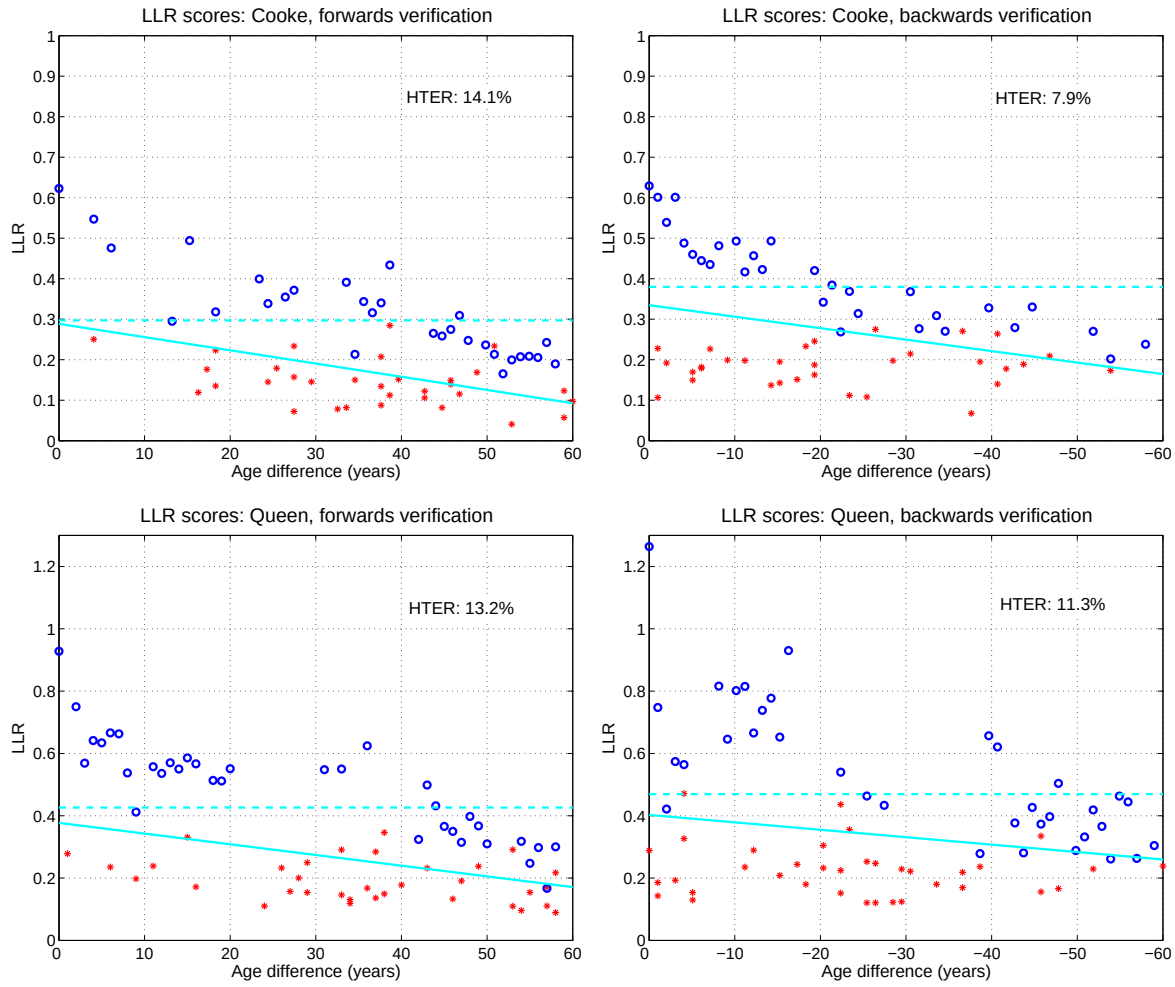

Figure 4.3: Test LLR scores for Cooke (male) and Queen (female), and their 8 testing imposters. The cyan line denotes the score-ageing dependent decision boundary, as determined in Figure 4.2. **Top Left:** Cooke, Forwards, **Top Right:** Cooke, Backwards, **Bottom Left:** Queen, Forwards, **Bottom right:** Queen, Backwards.

4.1.1.4 Absolute age vs Age difference

The effect of vocal ageing is not uniform across the lifespan, as evidenced both from physiological studies reviewed in Section 3.3 and from the experiments examining the acoustic correlates of ageing in Section 3.4. In Figure 3.11, a trend is observed of accelerating LLR score degradation as the age of the speaker at model training increases. Therefore, it is of interest to investigate the effect of using the absolute age of a speaker in the stacked classifier framework, rather than the age difference between training and testing. The experiment detailed in Section 4.1.1.2 was repeated, but with the absolute age of a speaker in each test recording taking the place of age difference in the score-ageing feature vector (after scaling to the range $[0, ..., 1]$). When it comes to imposters, the use of absolute speaker age assumes that the imposter age is known or provided.

An implementation of the system could operate by requesting the user input their age; in this case it is possible that an imposter inputs a false age (it is also possible that a genuine user do the same, but that case is not considered here). If the system instead operates by automatically calculating the age of a speaker by the difference between today's date and the date of enrolment, then the same problem arises in the imposter case, as they are automatically assigned the age of the speaker represented by model they are attempting (unwittingly or otherwise) to access. For this reason, three cases are considered in a stacked classifier experiment:

1. **Actual:** Both genuine-speaker and imposter LLR scores are associated with the actual age of the speaker in the test sample.
2. **Close:** Genuine-speaker LLR scores are associated with the actual age of the speaker in the test sample. Imposter LLR scores are associated with an age close (± 5 years) to the actual imposter age.
3. **Random:** Genuine-speaker LLR scores are associated with the actual age of the speaker in the test sample. Imposter LLR scores are associated with a random age (within bounds of all actual imposter ages).

The average HTERs for these three cases, along with the original HTERs from Table 4.1 are shown in Table 4.2. There is a reduction in HTER in the forwards direction by using the actual speaker age in the stacked classifier: 14.9% compared with 15.41%. As the relationship between absolute age and LLR score was found to be non-linear (e.g. Figure 3.11), using a non-linear SVM kernel in this experiment may exploit this relationship to a greater extent.

The absolute age of a speaker in their *training* recording was not considered in this experiment. In the forwards direction this simplification does not likely have a significant effect, as most speakers are within the age-range 20-30 at the year of training. In the backwards direction however, there is a greater difference between the training ages of speakers. This provides an explanation for the greater discrepancy between forwards and backwards HTERs in the actual age case and the age difference case.

An important observation is that the three permutations of imposter age assignment do not significantly effect the average HTER - the result of random imposter age assignment has the greatest effect, increasing the HTER from 16.16% to 17.15%. Thus imposter attacks that operate by manipulating age information do not greatly effect the system. This is not surprising, given the almost uniform imposter LLR score distribution, e.g. Figure 4.2.

4.2 Effect of Quality on Speaker Verification

As discussed in Section 3.5, quality variation is an unavoidable attribute of long-term longitudinal data like that in the TCDSA database. Therefore, although the aim of this research is to observe and compensate for ageing variation, quality variation must also be considered; ageing

	HTER(%)			
Age info.	Age diff.	Actual	Close	Random
Forwards	15.41	14.90	14.97	15.85
Backwards	16.90	18.02	18.57	18.44
Mean	16.16	16.46	16.77	17.15

Table 4.2: Average HTER of 18 speakers over a 60 year age difference using a score-ageing stacked classifier, with ageing information incorporated in different ways. Age diff. is the inclusion of ageing information via the age difference between training and testing samples, as in Table 4.1. Three possibilities for assigning imposter age values are then considered: actual age of imposter, close-to-actual age of imposter, and a random age. Genuine-speaker scores are associated with the actual age of the speaker in all of these cases.

and quality go hand in hand in uncontrolled data. To attempt to separate the influences of quality and ageing on the TCDSA data, it is important to be able to quantify the quality of a recording.

A *quality measure* of a speech sample is a measurable factor known to influence the classifier behaviour [167]. Thus, the quality measures from a set of speech samples should exhibit a relationship with their verification (LLR) score [167]. In this sense, quality measurement can be defined as the comparison of a speech sample to a predefined criterion known to influence the performance of the speaker verification system [79]. As noted by Grother and Tabassi [79], this concept of quality is distinct from a human perception of quality, i.e. a recording that sounds of 'good quality' to a human ear may not possess attributes that benefit the performance of the automated classifier. This is particularly true with speech: while some factors like additive or channel noise that affect the classifier output are audible, other factors such as the speech content or the attributes of a particular speaker in the population (e.g. the "Doddington Zoo effect" [50]), which may affect the classifier to the same extent, may not be noticeable to a listener.

The relationship between a quality measure and verification score can be exploited to improve the performance of a biometric classifier: In the domain of face verification, Drygajlo et al. [55] jointly model face quality measures and verification scores within a stacked classifier framework, improving classification performance. Richiardi et al. [169] use joint modelling of verification score and quality measures for speech, face, fingerprint and signature verification modalities, again in a stacked classifier framework, to reduce classification error.

Quality measures have also been used to optimise fusion of biometric modalities: A comprehensive framework for using quality measures to inform the combination of multiple biometric classifier outputs was presented by Poh and Kittler [152]. Kryszczuk et al. [120, 121] use quality measures for speech and face to optimise the performance of multimodal fusion.

Quality measures can also provide an estimate of the reliability of a verification score or deci-

sion; Richiardi uses quality measures for both speaker [167, 170] and signature [167] verification to quantify the confidence and reliability of the classifier output.

4.2.1 Quality Measures for Speech

An early, but comprehensive, examination of quality measures for speech can be found in [155]. Quality measures can be split into modality-dependent and modality-independent categories [169]. Modality-dependent measures exploit specific domain knowledge, e.g. a measure derived directly from a speech waveform, and thus are not generally applicable to a different modality. Modality-independent measures do not rely on specific domain knowledge, e.g. the distance from a verification score to a decision threshold, and thus are applicable to multiple modalities. The majority of speech quality measures previously applied in speaker verification fall into the modality-dependent category. Several modality-dependent measures are considered in this Chapter:

4.2.1.1 Signal-to-noise ratio (SNR)

The use of SNR as a quality measure was studied in [68, 82, 155, 168, 170]. Here, SNR was calculated using an energy-based voice activity detector. A sample is divided into 20ms frames, which are designated as either speech or non-speech by an energy threshold. The SNR is then given by:

$$SNR = 10 \log \frac{E_s}{E_{ns}} \tag{4.1}$$

Where E_s and E_{ns} are the mean energies of the speech and non-speech frames respectively. NIST provide a robust implementation of an SNR extractor in their 'speech assurance algorithm' [145]. However, in the subsequent experiments in this Chapter, there was little difference observed in the normalised SNR estimates of the NIST routine and the energy-based SNR calculation in Equation 4.1. Thus, all subsequent mentions of SNR refer to the definition in Equation 4.1.

4.2.1.2 Kurtosis

As clean speech has a distinctive distribution, higher-order statistics of a speech sample can be used as estimates of noise in the signal. The kurtosis of a distribution is its degree of 'peakedness', i.e. the lower the kurtosis, the flatter the distribution. Kurtosis was used as a quality measure in [82, 168, 169]. Kurtosis is given by:

$$Kurtosis = \frac{1}{T} \sum_{t=1}^{T} \sum_{x=1}^{X} \left(\frac{s_{xt} - \mu_t}{\sigma_t} \right)^4 \tag{4.2}$$

Where s_{xt} is the x^{th} element of the t^{th} frame of the speech sample and μ_t and σ_t are the mean and variance of this frame. The frame length was taken as 20ms.

4.2.1.3 Skewness

The skewness of a distribution is the degree to which it 'leans' to one side, i.e. negative skew indicates that the left tail of the distribution is longest, and the mass is centred on the right. Skewness was used as a quality measure in [82, 168, 169], and is given by:

$$Skewness = \frac{1}{T}\sum_{t=1}^{T}\sum_{x=1}^{X}\left(\frac{s_{xt}-\mu_t}{\sigma_t}\right)^3 \tag{4.3}$$

Where s_{xt} is the x^{th} element of the t^{th} frame of the speech sample and μ_t and σ_t are the mean and variance of this frame. The frame length was taken as 20ms.

4.2.1.4 Speech Intelligibility Index (SII)

The Speech Intelligibility Index (SII) is a quality measure that predicts the intelligibility of speech by a human listener under a range of adverse conditions. It is defined in the ANSI standard [7]. The SII is calculated as a weighted sum of the SNR in different frequency bands across the spectrum of a speech sample. The SII extraction was implemented based on the details in the ANSI standard, using the 'critical band' frequency windows and the critical band importance function 'short passages' condition. SII has not previously been used as a quality measure for speaker verification.

4.2.1.5 P.563

The final modality-dependent quality measure considered is P.563 (The ITU-T Standard for Single-Ended Speech Quality Assessment) [135]. This has been used as a quality measure for speaker verification in [68, 82]. The algorithm uses models of voice production and perception to output a mean opinion score in the range 1-5, from worst to best quality. ITU provides an implementation of this algorithm, which was used in this paper. The algorithm is designed for narrowband (3.4 kHz) speech assessment at an 8 kHz sampling rate. The data in the TCDSA Database is wideband, sampled at 16 kHz. It was therefore necessary to downsample data to 8 kHz prior to P.563 extraction. Ideally a single-ended quality measure designed for wideband speech would be applied, but to date no such measure has been standardised.

4.2.1.6 Modality-independent measures

Few modality-independent quality measures have been applied to speaker verification. Richiardi et al. [169] suggests two possible measures based on GMM covariance matrices. However, in the case of a GMM-UBM system with mean-only adaptation (the dominant approach), all GMMs and the UBM share common covariance matrices. Measures based on covariance matrices are thus not applicable.

Harriero et al. [82] proposes using the UBM log-likelihood score as a quality measure. The UBM log-likelihood (UBML) is an intermediate step in the LLR score calculation, given by:

$$UBML = \log\left(s|UBM\right) \tag{4.4}$$

where s is the current speech sample.

While this can be used as an indicator of quality - used to 'screen' the TCDSA database content in Section 3.8.1 - the UBM score is inherently contained in the LLR (the difference between speaker GMM log-likelihood and the UBM log-likelihood). In addition, Harriero et al. note that the UBM likelihood score reflects speaker-specific traits along with quality variation. In Section 4.2.2, a new modality-independent measure of quality, 'Wnorm', is proposed.

4.2.2 Proposed quality measure: Wnorm

A new modality-independent measure of quality is proposed, which measures quality in a similar way to the UBM score, while remaining independent from the LLR. During verification, for every test speech sample, a GMM is trained by UBM mean-only adaptation. The difference between the mean vectors of this GMM and the UBM is calculated. The difference is multiplied by the corresponding component weights and the matrix (Frobenius) norm is taken. The motivation is that a recording with a higher quality speech will result in greater 'movement' of important (higher weighted) means in training and will therefore produce a higher quality value. This measure is referred to as Wnorm (weighted norm), and is given by:

$$Wnorm = \sqrt{\sum_{m=1}^{M}\sum_{d=1}^{D}\left(w_m\left(\mu_{kd}^{T} - \mu_{kd}^{UBM}\right)\right)^2} \tag{4.5}$$

Where μ^T and μ^{UBM} are matrices of M component means, each of dimension D, of the test sample GMM and the UBM respectively. w is the vector of component weights of length M. A 30 second sample is used to train the GMM in the implementation of Wnorm.

The proposed approach uses a distance measure between GMMs to normalize verification scores. In this respect, the technique is related to the 'D-norm' (distance normalization) class of normalizing methods [13, 202]. In these works, the distance between each speaker GMM and a UBM is estimated via approximations of the Kullback-Leibler (KL) divergence. The speaker-dependent distance is then used to normalize the verification scores. D-norm is similar to Z-norm, but without the need for additional imposters for calculation of the normalization statistics. Wnorm differs in that the distance measure used for normalization is calculated using the test data. In this respect, it is more similar to T-norm (Section 2.2.1).

Since Wnorm is calculated from the mean matrices of a GMM and the UBM, it can be defined as a distance between mean 'supervectors' [118]. Thus, it is related to some common procedures in supervector-based speaker verification; there are numerous supervector kernels based on distance measures between GMMs, e.g. approximations of the KL divergence [36] and

bhattacharyya distance [40]. Wnorm is a similar distance measure, but operates in a different setting: as a means of test normalization for a GMM-UBM system.

4.2.3 Quality Measure Evaluation

Since the TCDSA database is influenced by a combination of ageing and quality factors, a separate dataset was used to evaluate the utility of the proposed Wnorm measure, along with the other quality measures described in Section 4.2.1, independently of ageing. The CSLU database, Section 3.5.5.1, was chosen for this purpose

From a total of 91 speakers, a gender-balanced UBM development set of 24 was selected, leaving 67 speakers as test subjects. With the same GMM-UBM configuration as in Section 3.7, a UBM was trained from the set of 24 speakers, and a GMM was adapted for each of the 67 test speakers using their first session of data. Using the remaining 11 sessions for testing, a set of genuine-speaker and imposter scores were generated for each speaker. Only the spontaneous speech samples from the database were used, all of which are at an 8 kHz sampling rate.

In addition, each of the seven quality measures in Section 4.2.1 were extracted for every test speaker speech sample. All measures were bounded between zero and one, where zero represents the worst possible quality and one the best. This mapping was done based on the extreme values across all speakers.

A framework for evaluating the utility of quality measures in biometric recognition, outlined by Grother and Tabassi [79], is to rank trials according to their quality score. Beginning with the lowest quality, trials are progressively removed and the effect on the recognition error is recorded. Where a useful quality measure is employed, there should be a decrease in error as lower quality samples are discarded.

Here, trials from the speaker verification experiment were ranked according to each of the quality measures in turn. The worst quality trials were progressively removed (1% at a time) from the evaluation until 20% of the samples remained. The overall equal error rate (EER) was assessed after each trial-removal.

Since a trial involves two speech samples, one for training and the other for testing, a quality measure extracted from each of the samples can be combined to determine a single measure of quality for a trial [79]. In this experiment, there were four methods considered for determining the overall quality Q of a trial, given a quality measure extracted from the training sample q_{tr} and a quality measure extracted from a testing sample q_{te}:

1. **Test sample only:** the quality of a trial is given by the quality of the test sample only, i.e. $Q = q_{te}$.

2. **Minimum:** the quality of a trial is given by the minimum of the train and test sample qualities: $Q = min\,(q_{tr}, q_{te})$.

3. **Mean (arithmetic):** the quality of a trial is given by the arithmetic mean of the train

and test sample qualities: $Q = (q_{tr} + q_{te})/2$.

4. **Mean (geometric):** the quality of a trial is given by the geometric mean of the train and test sample qualities [62]: $Q = \sqrt{q_{tr} \cdot q_{te}}$.

Thus, for every trial, the set of seven quality measures (Section 4.2.1) were determined for each of the four combination methods. Given the scores from the speaker verification evaluation, the effectiveness of each quality permutation was assessed by following the Grother and Tabassi [79] 'ranking and removal' approach.

In Figure 4.4, for each combination method, EER is plotted against the percentage of trials excluded according to each quality measure. For the 'test sample only' and 'minimum' combination methods, a decrease in EER is observed for most quality measures as the first 20% of trials are excluded. For the 'arithmetic mean' and 'geometric mean' combination methods, the Wnorm, UBML and SNR measures bring about a reduction in EER after 20% of trials are excluded. The other measures however, lead to a slight increase or minimum change in the EER.

For all four combination methods, after 20-30% of the trials are excluded, the proposed Wnorm measure brings about the largest reduction in EER (1-2%) of any of the quality measures. Kurtosis proves effective in the 'test sample only' case, reducing EER by 3% after 60% of trials are excluded, and by 5% after 80% of trials are excluded. SNR performs well in the 'test sample only' and 'minimum' combination methods.

Based on these trends, it would appear that Wnorm is the best indicator of quality in this evaluation, followed by Kurtosis and SNR. In terms of the combination methods, a similar rate of EER reduction is observed with Wnorm for both 'test sample only' and 'minimum' methods. However, the EER rate continues to decrease in the 'test sample only' case after 40% trial exclusion while remaining constant after this point in the 'minimum' case. Considering this Wnorm trend, along with that of Kurtosis and SNR, the 'test sample only' method seems to be the most effective approach. As such, all further uses of quality measures in this Chapter follow this method.

A similar 'EER vs trials excluded' quality measure evaluation using the TCDSA database would not be meaningful, as it contains limited ageing-independent data (i.e. multiple same-year sessions). Therefore, to observe the potential effectiveness of the quality measures for use with the TCDSA database, the correlation between genuine-speaker LLRs and quality measures for both the TCDSA and CSLU databases is compared in light of the 'EER vs trials excluded' evaluation in Figure 4.4.

The linear correlations between quality measures (extracted from test sample only) and genuine-speaker LLR scores from both the CSLU and TCDSA (forwards) evaluations are included in Table 4.3. In the TCDSA case, the relationship between ageing and the quality measures is also of interest. Thus, the linear correlations between the quality measures and the age difference (for the corresponding genuine-speaker trial) are also included in Table 4.3.

Considering the CSLU case, the measures with the strongest positive correlation, Wnorm

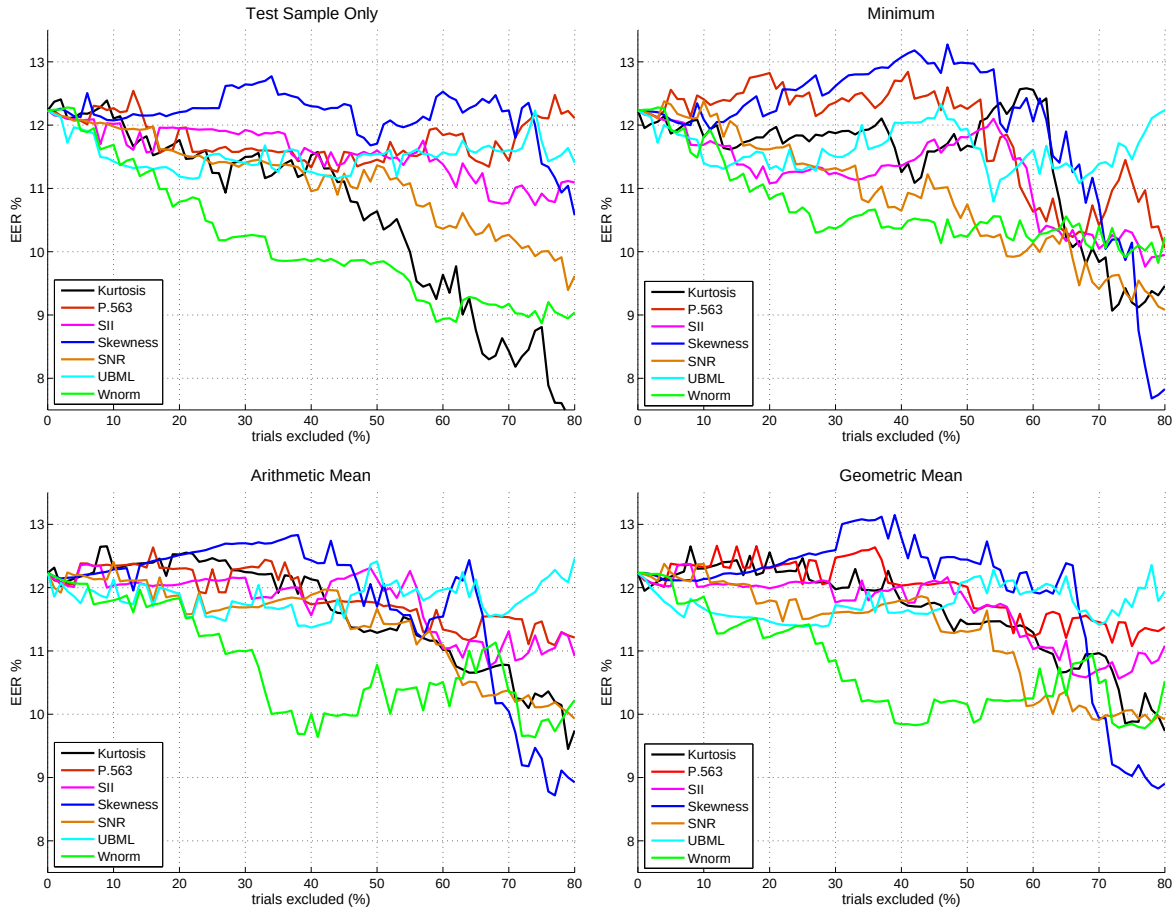

Figure 4.4: Speaker verification evaluation of the CSLU database: EER vs % of trials excluded, where trials are ordered from worst to best quality according to a range of quality measures extracted from their corresponding speech samples. Each plot demonstrates the effect of combining quality measures from both training and testing samples in different ways. **Top Left:** quality measures extracted from testing sample only, **Top Right:** quality measures given by the minimum value of training and testing sample quality measures, **Bottom Left:** quality measures given by the arithmetic mean of training and testing sample quality measures, **Bottom Right:** quality measures given by the geometric mean of training and testing sample quality measures.

and Kurtosis, were the most effective in the 'EER vs trials excluded' experiment, while P.563 and UBML, which performed poorly in that experiment, have small negative correlations. In the TCDSA case, the LLR has stronger correlations with Wnorm and Kurtosis than with the other measures (aside from SNR), and lower correlations with P.563 and UBML. These trends point towards Wnorm and Kurtosis being useful predictors of LLR on the TCDSA database. Although SNR has a low correlation with the CSLU LLR, it was effective in the 'EER vs trials excluded' experiment. It also has a high (relative) correlation with the TCDSA LLRs. This discrepancy suggests that the insight gained from this correlation analysis may be limited.

It is also evident from Table 4.3, that the TCDSA age difference has low (relative) cor-

relation with the most promising quality measures, Wnorm and Kurtosis. This implies that these measures have complimentary information to age difference, making them suitable for joint modelling with age difference and LLR score to improve classification.

	Kurtosis	P.563	SII	Skewness	SNR	UBML	Wnorm
CSLU: LLR	0.38	-0.09	0.11	0.05	0.08	-0.03	0.20
TCDSA: LLR	0.19	0.10	-0.03	0.04	0.26	-0.03	0.37
TCDSA: age diff.	-0.10	0.13	-0.07	-0.04	-0.19	0.15	-0.11

Table 4.3: Correlation between quality measures (extracted from test sample only), genuine-speaker LLR scores from the CSLU and TCDSA (forwards) evaluations, and TCDSA age difference.

4.2.4 Quality Stacked Classifier

Before proceeding to incorporate the evaluated quality measures into the score-ageing stacked classifier, Section 4.1, it is of interest to evaluate the effect of a score-quality stacked classifier (i.e. with no ageing information). The experiment in Section 4.1.1.2 was repeated with the lower-level feature vectors now comprised of the LLR score and a quality measure. As was the case with the inclusion of ageing information, the quality measures were mapped to the range $[0, ..., 1]$ based on their extreme values from all recordings in the database.

The application of these thresholds to the test data of the example speakers is shown in Figure 4.5. Quality is represented by the Wnorm measure in this example. The HTER determined with the score-quality decision threshold for the test speaker is indicated on each plot. Compared with the corresponding individual HTERs in Figure 3.17, the inclusion of quality information into the decision threshold reduces the individual HTERs by 5.2 - 11.7% in absolute terms.

The average HTER in forwards and backwards directions, for each of the seven quality measures, is given in Table 4.4. Compared with the baseline GMM-UBM HTERs, Table 4.1, the inclusion of each quality measure lowers HTER. The reduction in HTER with any of the measures is not a great as that achieved by including ageing information however (a full results comparison is shown in Table 4.5).

Wnorm is the best performing measure in both directions. It was anticipated to perform well given the correlation analysis in Section 4.2.3. However, the use of kurtosis, which was also anticipated to be effective, results in the second highest mean HTER. SNR and P.563 on the other hand, although anticipated to perform poorly, outperform kurtosis, SII, skewness and UBML in terms of mean HTER.

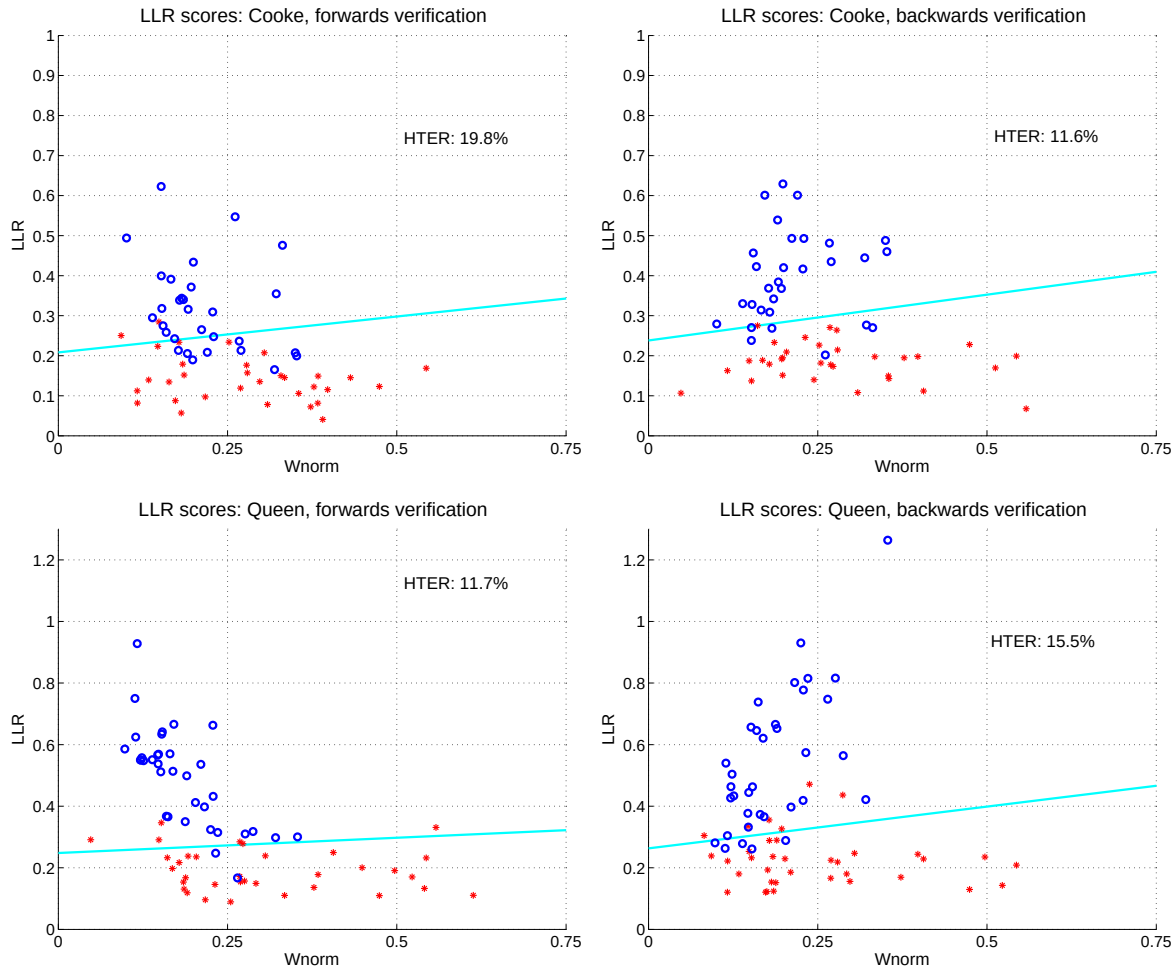

Figure 4.5: Test LLR scores for Cooke (male) and Queen (female), and their 8 testing imposters, plotted against the quality measure Wnorm. The cyan line denotes the score-quality dependent decision threshod. **Top Left:** Cooke, Forwards, **Top Right:** Cooke, Backwards, **Bottom Left:** Queen, Forwards, **Bottom right:** Queen, Backwards.

	HTER(%)						
	Kurtosis	P.563	SII	Skewness	SNR	UBM	Wnorm
Forwards	20.13	18.04	17.33	19.08	18.00	19.99	16.98
Backwards	19.49	17.32	19.78	18.97	17.45	19.97	17.08
Mean	19.81	17.68	18.56	19.03	17.73	19.98	17.03

Table 4.4: Average HTER of 18 speakers over a 60 year age difference using a score-quality stacked classifier for the different quality measures (extracted from the test sample only).

4.3 Score-Ageing-quality Stacked Classifier

In Sections 4.1.1 and 4.2.4, both ageing and quality information have been shown to independently reduce the baseline HTER. In this section, the effectiveness of combining ageing and

quality information with the baseline LLR scores in a stacked classifier framework is evaluated. The experiment in Section 4.1.1.2 was again repeated, now with a three-dimensional score-ageing-quality feature vector. Each of the quality measures in turn were used as the quality entry in this vector. As before, the quality measures were mapped to the range $[0, ..., 1]$.

Training LLRs for the two example test speakers, in forwards and backwards directions, are plotted against quality and age difference in Figure 4.6. Quality is represented by the Wnorm measure in this example. A three dimensional score-ageing-quality decision threshold, trained on the data in the plot, is also shown. The application of the trained thresholds to the test data of the example speakers is shown in Figure 4.7. The HTER determined with the threshold for the test speaker is indicated on each plot. The score-ageing-quality decision threshold reduces the individual HTERs by a further 1.4 - 7.7% compared with the minimum HTER obtained with either a score-ageing or score-quality HTER, Figures 4.5 and 4.3.

A full comparison of the different score-ageing-quality stacked classifier permutations is given in Table 4.5. The best overall performance is achieved by compensating for both ageing and quality. Examining the mean of the forwards and backwards HTER for the score-ageing-quality case, Wnorm is the overall best performing measure, reducing HTER to 13.72%. SNR is the second best performing measure, reducing HTER to 15.30%.

The analysis of quality measures on the CSLU database, Figure 4.4, correctly predicted the effectiveness of both Wnorm and SNR. Kurtosis, which was also expected to perform well, resulted in the highest mean HTER however. The use of a separate database to evaluate quality measures likely led to this disagreement. The correlation analysis suggested that Wnorm and Kurtosis would be the most effective measures. The fact they were the most effective and least effective measures respectively, underlines that the linear correlation of a quality measure with genuine-speaker LLR score is not a reliable indicator of its utility in classification.

While improvements in the baseline HTER were observed with the score-quality classifier for all quality measures, the addition of ageing information to the score-quality combination results in further reductions in the HTER in all cases. The relative improvements with each measure after the addition of ageing, followed by quality information, illustrate that although the two factors are not independent, they have separate influences on score variability, and therefore compensating for both results in the best classification performance.

To assess the significance of the differences between the HTERs for each quality measure, the fact that the HTER is a combination of two proportions, the False Acceptance Rate (FAR) and the False Rejection Rate (FRR), must be accounted for. Bengio and Mariéthoz [14] extend a number of statistical tests to HTERs. Their methodology to express the difference between two HTERs, assuming dependence of the two underlying distributions, was applied here. The percentage confidence δ that the HTERs of two models, A and B, are significantly different is given by:

$$\delta = \Phi(z) \tag{4.6}$$

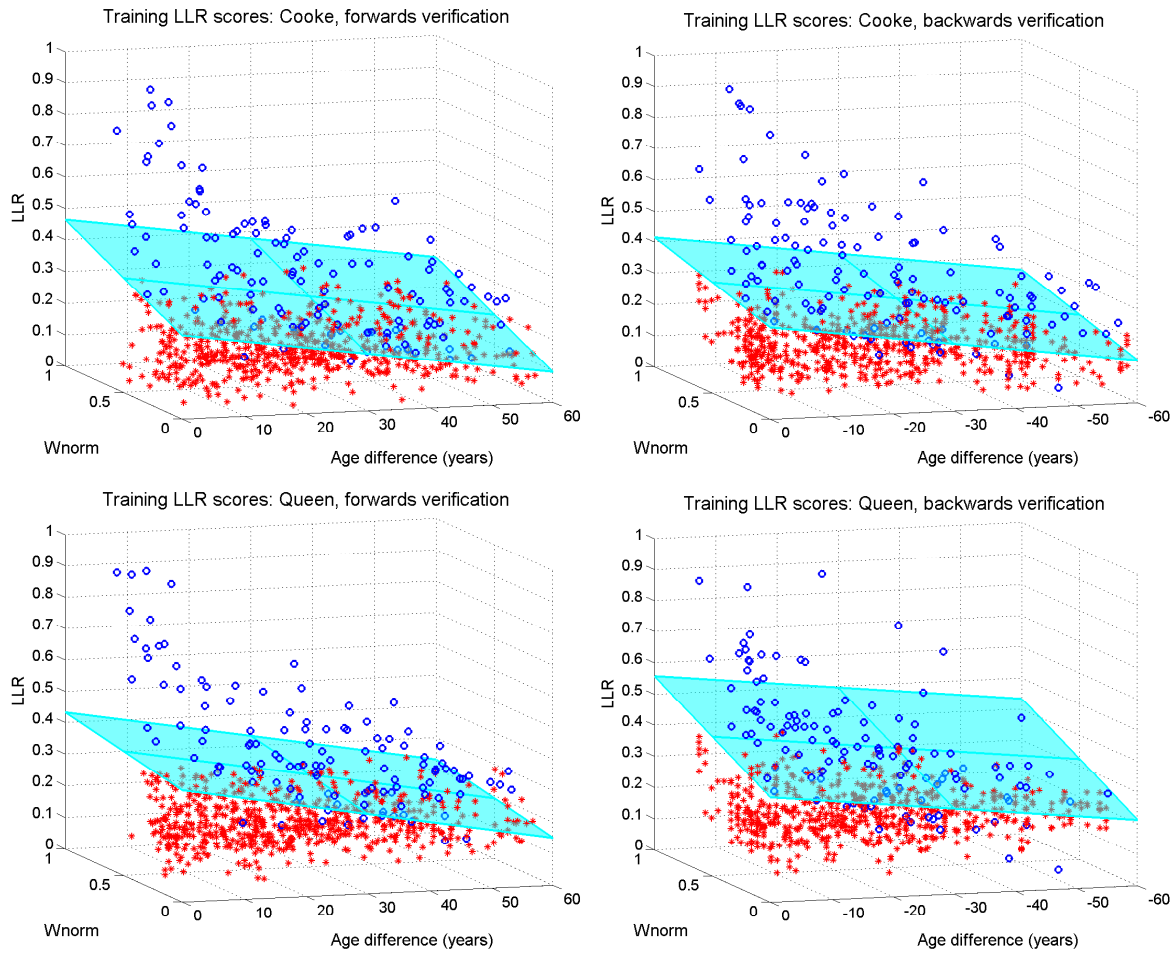

Figure 4.6: LLR scores used to train a decision threshold for Cooke (male) and Queen (female), consisting of the genuine-speaker scores for all *other* 17 speakers, and all scores from 9 training imposters. The cyan plane denotes the score-ageing-quality dependent decision boundary that minimises the HTER on the training set (the quality measure is Wnorm). **Top Left:** Cooke, Forwards, **Top Right:** Cooke, Backwards, **Bottom Left:** Queen, Forwards, **Bottom right:** Queen, Backwards.

Where Φ represents the normal cumulative distribution function, and:

$$z = \frac{|FAR_{AB} - FAR_{BA} + FRR_{AB} - FRR_{BA}|}{\sqrt{\frac{FAR_{AB} + FAR_{BA}}{4NI} + \frac{FRR_{AB} + FRR_{BA}}{4NC}}} \tag{4.7}$$

Where NI and NC are the number of imposter and client (genuine-speaker) trials, and

- $FAR_{AB} = NI_{AB}/NI$, where NI_{AB} is the number of imposter trials correctly rejected by system A and incorrectly accepted by system B.
- $FAR_{BA} = NI_{BA}/NI$, where NI_{BA} is the number of imposter trials correctly rejected by system B and incorrectly accepted by system A.

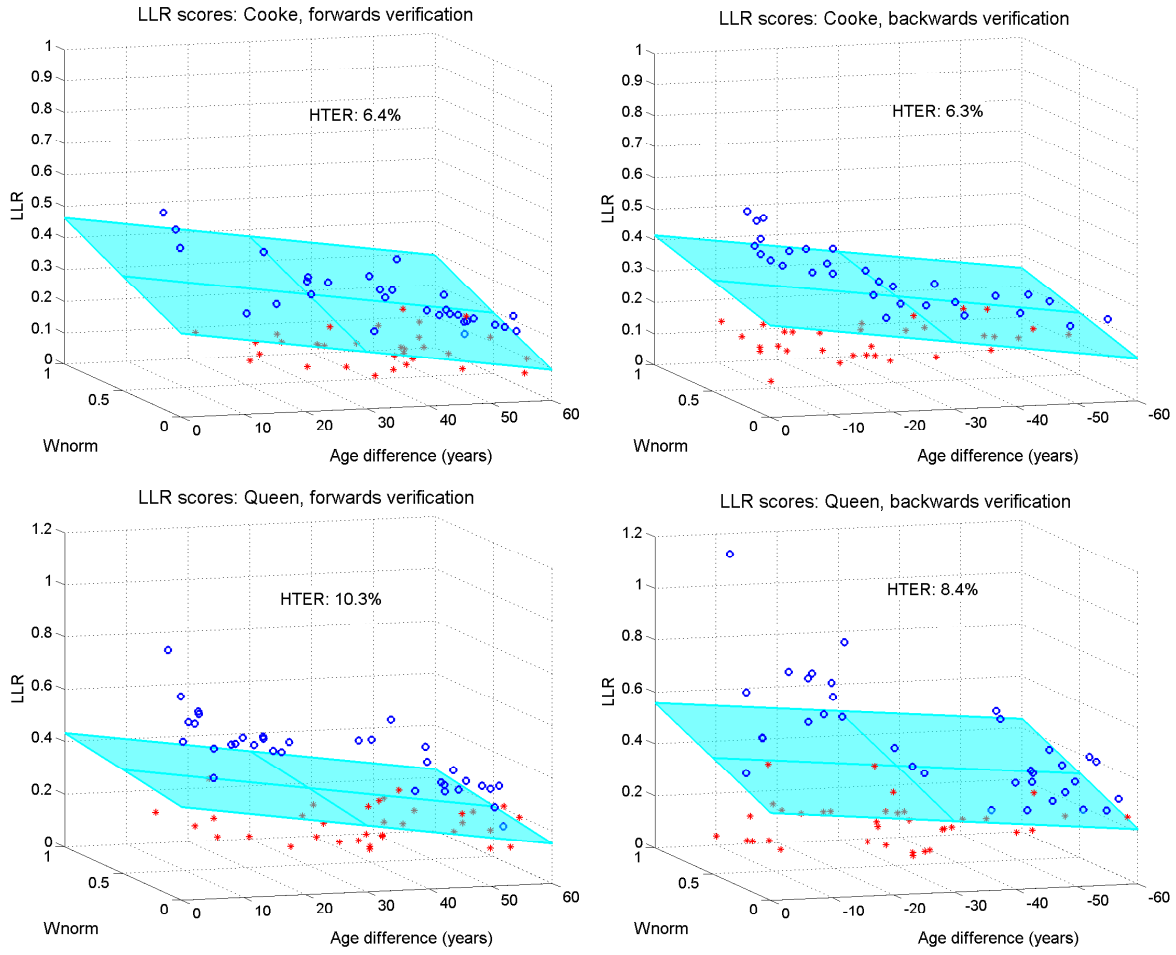

Figure 4.7: Test LLR scores for Cooke (male) and Queen (female), and their 8 testing imposters. The cyan plane denotes the score-ageing-quality dependent decision boundary, as determined in Figure 4.6 (the quality measure is Wnorm). **Top Left:** Cooke, Forwards, **Top Right:** Cooke, Backwards, **Bottom Left:** Queen, Forwards, **Bottom right:** Queen, Backwards.

- $FRR_{AB} = NC_{AB}/NC$, where NC_{AB} is the number of client (genuine-speaker) trials correctly accepted by system A and incorrectly rejected by system B.
- $FRR_{BA} = NC_{BA}/NC$, where NC_{BA} is the number of client (genuine-speaker) trials correctly accepted by system B and incorrectly rejected by system A.

Table 4.6 shows the percentage confidence that the HTER of Wnorm is statistically different to the HTERs of the other quality measures. A positive percentage denotes the case where the HTER of Wnorm is lower than the compared measure, while a negative percentage denotes the case where the compared measure has a lower HTER. From this assessment, there is moderate confidence in the improvement offered by Wnorm over the other three measures in the forwards direction, particularly in the 'GMM-UBM + Ageing + Quality' case, where the confidence is

GMM-UBM (baseline)	HTER(%)						
Forwards	24.98						
Backwards	22.77						
Mean	23.88						
GMM-UBM + Ageing							
Forwards	15.41						
Backwards	16.90						
Mean	16.16						
GMM-UBM + Quality	Kurtosis	P.563	SII	Skewness	SNR	UBML	Wnorm
Forwards	20.13	18.04	17.33	19.08	18.00	19.99	16.98
Backwards	19.49	17.32	19.78	18.97	17.45	19.97	17.08
Mean	19.81	17.68	18.56	19.03	17.73	19.98	17.03
GMM-UBM + Ageing + Quality	Kurtosis	P.563	SII	Skewness	SNR	UBML	Wnorm
Forwards	16.01	15.43	16.31	16.27	15.57	16.29	12.10
Backwards	17.54	16.52	15.84	16.75	15.02	16.06	15.34
Mean	16.78	15.98	16.55	16.51	15.30	16.18	13.72

Table 4.5: Average HTER (%) for all 18 speakers over a 60 year age difference, for the baseline GMM-UBM system and all stacked classifier score/ageing/quality configurations presented in this Chapter

approaching 90%. In the backwards direction, the confidence in the improvement offered by Wnorm is less significant. SNR outperforms Wnorm for the 'GMM-UBM + Ageing + Quality' condition, hence the negative confidence. Overall, this assessment offers support for Wnorm as the best performing measure on this dataset.

4.4 Discussion

A GMM-UBM speaker verification evaluation of the TCDSA database demonstrates that the LLR scores of genuine-speakers decrease progressively as the age difference between their training and testing samples increase. In this Chapter, a stacked classifier framework has been applied to the problem, enabling a decision threshold dependent both on ageing and LLR score. The best performing configuration succeeds in reducing mean HTER from 23.88% to 13.72% over a 60 year age range.

The study of vocal ageing requires long-term longitudinal data. The nature of such data guarantees the presence of non-ageing-related variability, due to varying recording environments and the evolution of recording technology, for example. To investigate the effect of this general quality variation on the LLR scores from the speaker verification experiment, a number of quality

	Confidence(%): GMM-UBM + Quality					
	Kurtosis	P.563	SII	Skewness	SNR	UBML
Forwards	74.53	74.68	51.26	87.00	52.89	89.63
Backwards	81.41	61.09	84.39	74.42	58.10	84.87
	Confidence(%): GMM-UBM + Ageing + Quality					
	Kurtosis	P.563	SII	Skewness	SNR	UBML
Forwards	85.13	88.23	84.39	81.81	85.06	83.06
Backwards	74.77	61.20	59.49	61.10	-67.00	65.83

Table 4.6: Percentage confidence that the HTERs obtained with Wnorm, for the 'GMM-UBM + Quality' and 'GMM-UBM + Ageing + Quality' cases (Table 4.5) are statistically different from each of the other quality measures. A positive percentage denotes the case where the HTER of Wnorm is lower than the compared measure, while a negative percentage denotes the case where the compared measure has a lower HTER. The only negative entry, '-67.00', indicates 67% confidence in the difference between SNR and Wnorm HTERs (where SNR outperforms Wnorm) being significant.

measures were extracted from the TCDSA data. Along with the established quality measures of kurtosis, P.563, SII, skewness, SNR and UBML, a new model-based measure of quality, Wnorm, was proposed.

A decision threshold dependent on quality measures and ageing information resulted in a decrease in the baseline HTER. This (expected) result confirmed that there are other variabilities intertwined with that of ageing. However, since a greater reduction in HTER was obtained with with a score-ageing threshold than with a score-quality threshold, the claim that the dominant long-term variability in the TCDSA database is ageing is supported. The combination of both ageing and quality information in a score-ageing-quality decision threshold brought about greater reductions than each factor independently. Thus, these sources of variability provide complementary information.

The proposed quality measure, Wnorm, proved the most effective of the quality measures for both score-quality and score-ageing-quality decision thresholds. As Wnorm is model-based, its quality estimates may be more 'robust' than those of signal-based measures. As a consequence, these quality estimates can then be exploited more effectively in classification. Wnorm was also found to be an effective predictor of LLR score in the CSLU experiment, Figure 4.4, demonstrating that it is not database-dependent. There was a general discrepancy between the signal-based measures in terms of predictions from the CSLU 'EER vs trials excluded' evaluation, the correlation analysis, and the actual HTER reduction in the stacked classifier framework. For example, the mean HTER for kurtosis for the score-ageing-quality decision threshold is the highest of all quality measures (16.78%). However, kurtosis performed well in the 'EER vs trials excluded' evaluation and displayed a correlation with LLR score. This demonstrates that

the utility of signal-based measures varies depending on the data, and that linear correlations between genuine-speaker score and quality are not very informative.

The stacked classifier method was evaluated over a very large age difference range. Errors, particularly false acceptances, increase towards the upper end of this range, e.g. Figure 4.3 and Table 4.1. With the linear decision threshold, there is also a trade-off between reducing the false rejection rate over the long-term (>20 years), and reducing the false acceptance rate over the short-term (<5 years). In practice, operating ranges of biometric systems are far shorter than 60 years, in which case the performance of the stacked classifier would be more effective. In addition, the experiment of assigning ages to imposters in different ways suggested that the system would be relatively robust to imposter attacks that operate by manipulating age information.

As observed in Chapter 3, the effects of vocal ageing are gender-dependent, absolute-age-dependent and vary per-individual. The relationship between these factors and the LLR score from a verification experiment is therefore a complex one. Ageing information was exploited in the stacked classifier by modelling a linear relationship between LLR score and the age difference between two samples in a speaker- and gender-independent way. Although this was effective, it is a very simple model considering the complexity of the process. A more complex modelling of the relationship between LLR scores and ageing could be achieved by incorporating additional features into the stacked classifier framework (e.g. multiple quality and ageing measures) and employing a non-linear kernel in the SVM classifier. The scope for such extension is limited however, given the size of the TCDSA database.

An alternative approach is move from the level of scores to the *model-level*, where the effect of ageing on the components of a speaker model may present a more effective means of compensation. This idea is explored in Chapter 5, where a new approach to ageing-variability compensation for speaker verification, at the model-level, is presented.

5

Eigenageing compensation for ageing speaker verification

In Chapter 4, an ageing compensation method for speaker verification, utilising a stacked classifier framework to implement an ageing-dependent decision threshold, was presented. This method applies ageing compensation at the *score-level*. In this Chapter, a new approach, applying ageing compensation at the *model-level*, is presented.

The new method, *eigenageing* compensation, compensates for the effect of vocal ageing by adapting a speaker model towards a test sample based within a predetermined ageing variability subspace. The term eigenageing has been adopted as a reference to the related method of *eigenchannel* compensation, which compensates for the effect of channel variability by adapting a speaker model to a test sample according to a predetermined channel variability subspace.

In Section 5.1, the eigenageing compensation algorithm is introduced in detail. In Section 5.2, experimental evaluations of the technique are presented on different datasets. The experiments in Section 5.2.3 were previously presented in [106]. In Section 5.3, a new approach to age estimation derived from eigenageing compensation is presented, along with an experimental evaluation.

5.1 Eigenageing compensation background

Eigenchannel compensation was introduced by Kenny et al. [115,116], and subsequently used by several groups for speaker verification evaluations [25,27,29,113]. As summarised in Section 2.4,

eigenchannel compensation suppresses the effects of inter-session variability by adapting a model trained under one 'channel' (or recording condition) towards a different channel condition present in the test data. The adaptation is constrained to a 'subspace' representing channel variability. Thus, adaptation from one speaker to another is minimised.

Eigenageing compensation is analogous to eigenchannel compensation, with *ageing* taking the place of *channel* as the source of variability. Thus, the aim is to adapt a model trained at one age to the age of the speaker in a test recording. This adaptation is constrained to a limited vocal ageing subspace, so that the speaker model is shifted toward the age rather than the identity of the test speaker. After adaptation, the model can be tested in the usual way, resulting in an age-compensated LLR (log-likelihood ratio).

Eigenageing compensation was integrated into the baseline GMM-UBM system described in Section 3.7.2. The vocal ageing subspace is estimated with a set of ageing speaker GMMs, translated into a 'supervector' representation.

5.1.1 GMM supervector representation

As discussed in Section 3.7.2, in a GMM-UBM system, individual speaker GMMs are trained by adapting the statistics of the UBM given new data. It has been shown experimentally that adapting only the UBM means is effective in terms of verification performance [118], and thus the weights and covariances are not typically adapted in practice. In the GMM-UBM system used here, only mean adaptation is applied. Thus, each speaker GMM differs only in the means of its components and can therefore be described by the concatenation of its mean vectors. This representation of a GMM is commonly referred to as a supervector [27,118]. Before concatenating the GMM means, each is normalised by its corresponding standard deviation [25]. The conversion from a GMM to supervector (SV) can be expressed as:

$$SV = \left[\frac{\mu_1}{\sqrt{\Sigma_1}}, \frac{\mu_2}{\sqrt{\Sigma_2}}, \ldots, \frac{\mu_M}{\sqrt{\Sigma_M}}\right] \tag{5.1}$$

where M is the number of GMM components, and μ_m and Σ_m are the mean and covariance of the m_{th} component. Each component is D dimensional, such that $\mu_m = (\mu_{m1}, \mu_{m2}, ..., \mu_{mD})$. For example, a GMM with 512 components trained from a 24 dimensional feature vector is represented by a supervector with $MD = 512 * 24 = 12288$ elements.

5.1.2 Ageing Subspace Estimation

The aim of ageing subspace estimation is to find a low-rank (relative to the supervector dimensionality) matrix that defines the directions of greatest change in the models of ageing speakers. To create a set of models of ageing speakers, a subset of recordings from the TCDSA database were extracted, according to several constraints: The effect of vocal ageing generally accelerates in old-age, as discussed in Section 3.7.4. To constrain the ageing subspace to 'typically ageing adults', the data used for modelling was restricted to recordings of speakers within the age range

of 20 to 60. The upper range was increased to 63 for one male speaker, Cronkite, so that there were two recordings of him in this subset (the other at age 47). To balance the contribution of different speakers, the number of recordings per speaker was limited to 6, with a 5 to 10 year age difference between recordings. Finally, since male and female voices change differently with ageing (as discussed in Section 3.3), it was considered appropriate to train separate ageing subspaces for males and females. A schematic of the resulting male and female subsets is provided in Figure 5.1.

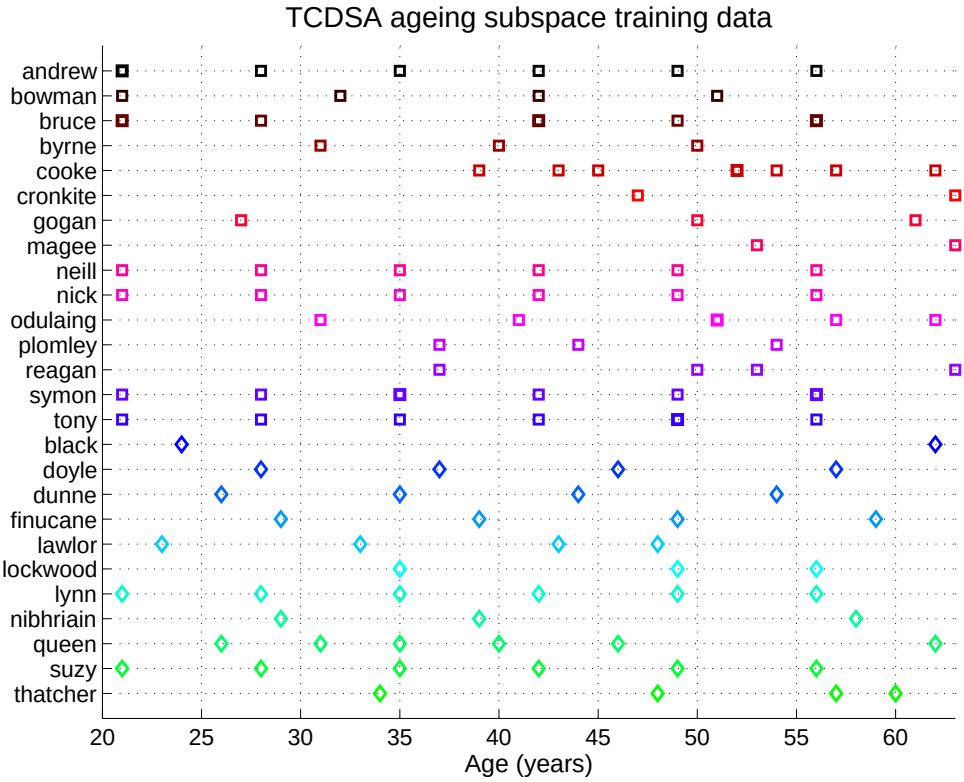

Figure 5.1: A schematic of the TCDSA data used for ageing subspace estimation. Speaker names are shown on the y-axis, with their age in each recording given on the x-axis. Each point indicates a recording used for training the subspace. The speakers are ordered by gender, where squares indicate males and diamonds indicate females.

For each recording in this reduced dataset, depicted in Figure 5.1, a GMM was trained and converted to supervector representation as detailed in Section 5.1.1. To remove speaker-dependent information from the supervectors, they must be normalised in some way. The approach in eigenchannel subspace estimation is to subtract each speaker's mean supervector from the set of all their supervectors [27]. This removes most of the speaker variability, but leaves session variability.

Ageing variability is a 'special case' of session variability, in that its source, the physiological change in the voice, is known. It is assumed that ageing variability manifests itself as cumulative

change in the voice. Thus, the effects on the model of an ageing speaker will be progressive and acting primarily in one direction, i.e. the ageing process is non-reversible. With this in mind, the speaker supervectors were normalised in a slightly different way to the mean-subtraction approach. For each speaker's set of supervectors, a set of ageing-difference supervectors were found by subtracting each speaker's age 1 (i.e. 'youngest') supervector from their subsequent (i.e. age 2, age 3, $\cdots$, age N, where $N \leq 6$) supervectors. A second and third set of ageing-difference supervectors were found by subtracting each speaker's age 2 and age 3 supervectors from their subsequent supervectors. The number of difference stages were reduced for speakers with an insufficient number of recording ages. The resulting three-stage ageing-difference supervectors from have much of their inter-speaker variability removed, but have progressive ageing-related variability remaining. This operation is depicted in Figure 5.2, using the male speakers' supervectors as an example.

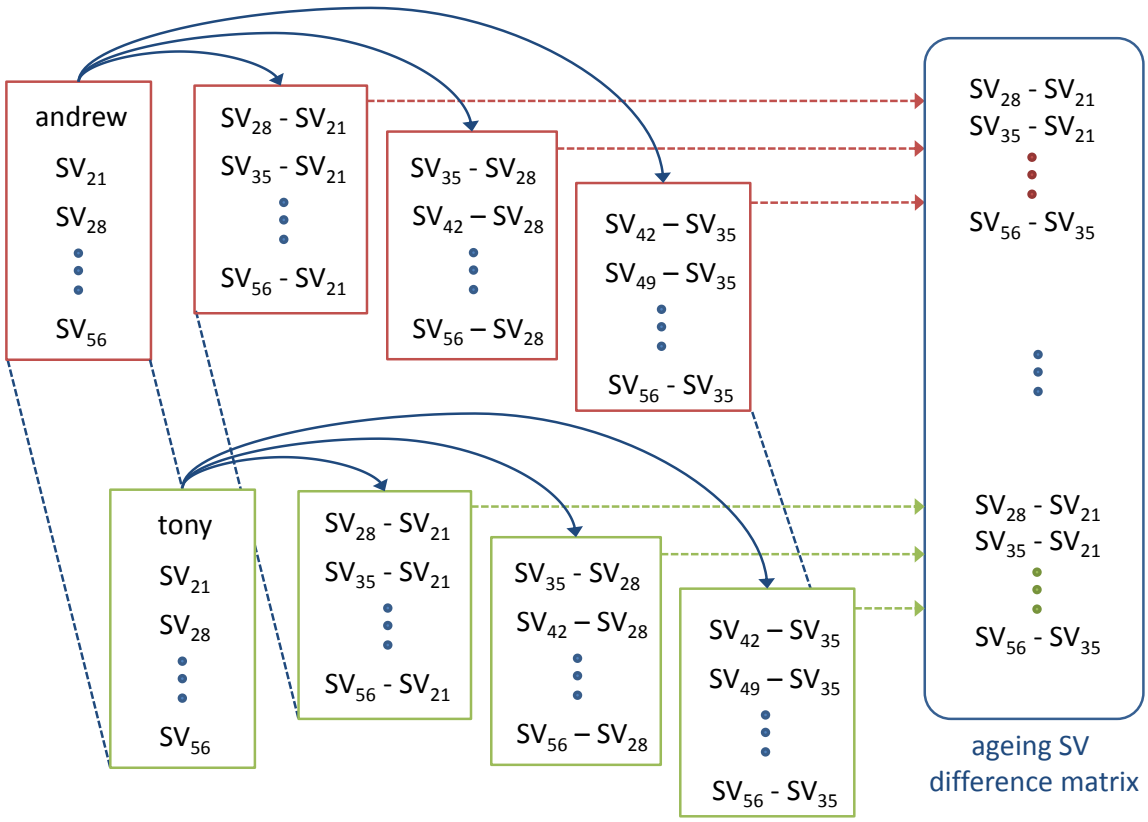

Figure 5.2: The three-stage difference approach to supervector normalisation. Taking Andrew (male) as an example: GMM supervectors (SVs) are created for each of his reordings at the ages denoted in Figure 5.1. This results in a set of SVs from age 21-56: [SV_{21},...SV_{56}]. The differences between the youngest three SVs and the subsequent SVs are pooled together in the ageing SV difference matrix. This process is repeated for every other male speaker in Figure 5.1, with each contribution included in the overall ageing SV difference matrix.

To determine the principal directions of ageing variability, an eigen-analysis is then performed on the ageing SV difference matrix. The ageing SV difference matrix SV_{age} is of dimension $MD \times J$, where MD is the supervector dimension, a product of the number of GMM components M and the feature dimension D. J is the total number of three-stage ageing difference supervectors, which was approximately 100 for males and 80 for females. A principal components analysis (PCA) is applied by extracting the R principal eigenvectors of the within-speaker covariance matrix $\frac{1}{J} SV_{age} SV_{age}^{\top}$. The R principal eigenvectors form the ageing subspace matrix V, which is of dimension $MD \times R$

A plot of the 50 largest eigenvalues from this analysis, sorted in decreasing order, is given in Figure 5.3 for both the male and female case. The decrease in magnitude of the eigenvalues is approximately exponential. This indicates that ageing variability is low-dimensional within the ageing difference supervector matrix, and thus can be described with a relatively small number of eigenvectors. In the subsequent experiments in this Chapter, the number of principle eigenvectors R used to compose the ageing subspace V, is taken as 20.

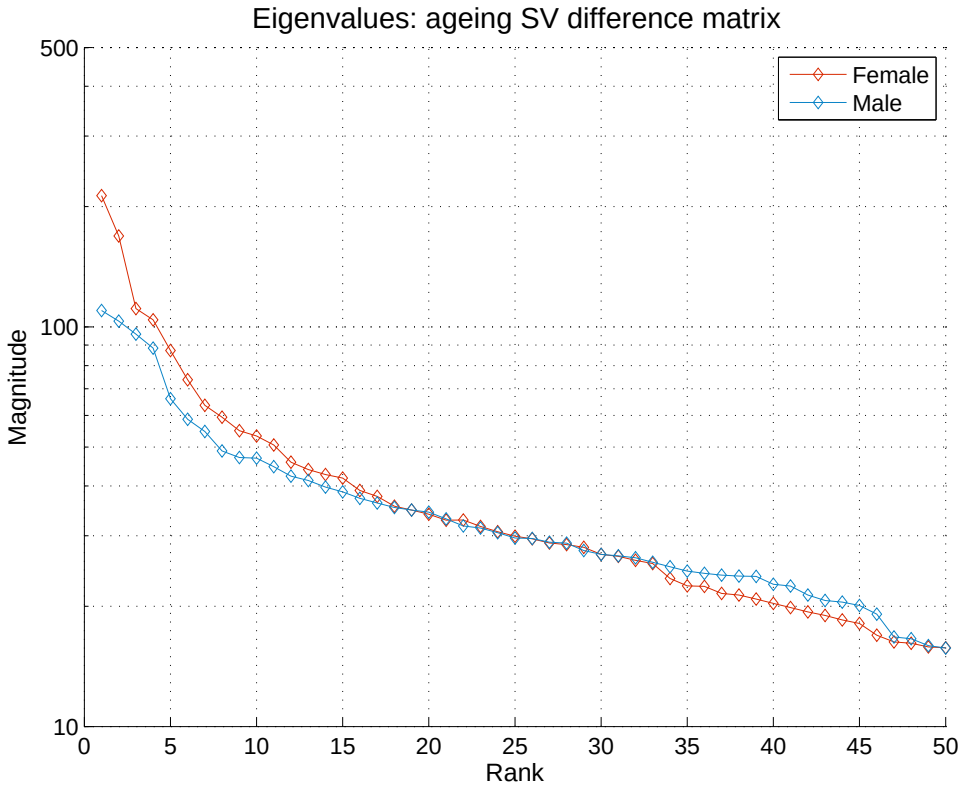

Figure 5.3: Eigenvalues determined from the set of female and male ageing-difference supervectors

the first two eigenvalues are significantly larger in magnitude (approximately double) than the first two male eigenvalues.

Kenny et al. [113] present a graph of the 50 largest eigenvalues corresponding to an eigen-channel matrix. They base the graph on female data, and state that the graph in the male case

is similar. Their graph is similar in magnitude and rate of decrease to the male case in Figure 5.3. In the female case in Figure 5.3 however, the magnitudes of the first two eigenvalues are significantly larger than the first two male eigenvalues. Since the male and female data share the same sources, the most likely explanation for this discrepancy is the difference between the vocal ageing patterns in males and females. As concluded in Chapter 3, trends in vocal features (F0, jitter, NHR etc.) are more consistent between female speakers than males. This finding goes towards explaining the larger female eigenvalues; the principle directions of ageing change in female voices are more concentrated than those in male voices.

The assumption made in generating the ageing subspace is that the only source of variability between speaker sessions is age. As discussed in the introduction to the TCDSA database, Section 3.5.1, although it was compiled such that ageing was the main source of variability, there are other unavoidable sources of variability due to the different recording environments and technologies involved. We assume here that ageing variability exceeds that of other variabilities, and therefore that the R principal eigenvectors are descriptors of ageing variability. While undoubtedly some non-ageing-related variability will be modelled by the ageing subspace, the discrepancy between male and female eigenvalues supports the assumption that ageing is the dominant source of variability.

5.1.3 Eigenageing compensation

Eigenageing compensation is applied by shifting a speaker model towards a test sample, within the constrained subspace specified by the ageing subspace matrix V. Given a speaker model λ (in supervector representation) and a set of test features $O = [o_1, o_2, \ldots, o_T]$, the ageing-adapted model is given by maximising the following criterion, with respect to the low-dimensional vector x [25]:

$$p\left(O|\lambda + Vx\right) \tag{5.2}$$

The adaptation shown in Expression 5.2 is a constrained maximum likelihood (ML) adaptation. A maximum a posteriori (MAP) adaptation is applied in several works on eigenchannels [29,113], by placing a normal prior on x (zero mean and unit covariance). However, Brümmer et al. [27] found that the simpler ML criterion performed as well as MAP. As shown in [25], x is maximised by:

$$x = A^{-1} \sum_{m=1}^{M} V_m^{\top} \sum_{t=1}^{T} \gamma_m\left(t\right) \frac{o_t - \mu_m}{\sigma_m} \tag{5.3}$$

where o_t is the tth feature frame, $V_m^{\top}$ is the transpose of the $D \times R$ block of matrix V corresponding to the mth mixture component, $\gamma_m\left(t\right)$ is the probability of occupation of mixture component m at time t. μ_m and σ_m are the component mean and standard deviation vectors. A is given by:

$$A = \sum_{m=1}^{M} V_m^{\top} V_m \sum_{t=1}^{T} \gamma_i(t) \tag{5.4}$$

The occupation probabilities $\gamma_i(t)$ are computed using the test speaker GMM, λ. In the experiments in this Chapter, x was initialised as 0 and three iterations of the maximisation in Equation 5.2 were computed, each time using the updated GMM to calculate $\gamma_i(t)$. The ageing-compensated LLR (log-likelihood ratio) score for each trial was then calculated in the usual manner, using the UBM and the adapted speaker model λ_a translated back into GMM representation:

$$LLR_{compensated} = \log p(O|\lambda_a) - \log p(O|UBM) \tag{5.5}$$

5.2 Eigenageing compensation experimental evaluation

The full set of 26 TCDSA speakers were used for the experiments in this Chapter. In the speaker verification experiments in Chapters 3 and 4, a cross-validation approach was taken, allowing the remaining TCDSA speakers to be used as either imposters or Z-norm speakers. Since the eigenageing compensation approach relies on the TCDSA database to create the ageing subspace, a separate data source is required for imposters and Z-norm speakers.

A male-only experiment is presented in Section 5.2.2, using the (male-only) TCDSA Forensic Development (TCDSA-FD) database speakers, Section 3.5.3, for imposter and Z-norm purposes. Following this, a male/female experiment is presented in Section 5.2.3, utilising the (male/female) CSLU database [42], Section 3.5.5, for imposter and Z-norm purposes.

5.2.1 Feature extraction and GMM-UBM system configuration

The GMM-UBM configuration is that used in [106], which differs from the configuration in previous Chapters in that all data was first downsampled to 8 kHz, only delta coefficients were appended to MFCCs (resulting in a 24 dimensional feature vector), and 512 GMM components were used. All other aspects of the system were consistent with that in Section 3.7.

The reduction in sampling rate and feature and model dimensionality, compared with that in Section 3.7, does not significantly affect the baseline performance of the GMM-UBM system. This is observable in Section 5.2.2. The use of this 'reduced' configuration was prompted by the need to use the CSLU database, which contains data sampled at 8 kHz only. The reduced amount of data available for the male-only UBM, required in Section 5.2.2, also influenced this decision. In addition, this configuration more closely aligns with other speaker verification studies; the NIST speaker recognition evaluations for example, use 8 kHz speech data, as do the previously referenced studies on eigenchannel compensation, e.g. [25].

5.2.2 Eigenageing compensation male evaluation

A male-only evaluation was designed using a cross-validation approach. A schematic of the full experiment is provided in Figure 5.4. Taking each one of the 15 TCDSA males at a test speaker, in turn, the experiment proceeded as follows:

- A UBM was trained with 30 minutes of male speech from the TCDSA-UBM database.
- A GMM was adapted from the UBM using one minute of data from the test speaker's youngest recording.
- An ageing subspace was estimated using the *other* 14 male speakers in the (constrained) TCDSA dataset, Figure 5.1. In estimating the subspace, GMMs were adapted from the UBM with one minute of speech.
- Genuine-speaker scores were obtained for the test speaker with the remainder of their training recording, and all of their subsequent (i.e. older) recordings. The 'quality-screened' version of the database, Figure 3.13, was used. No age restrictions were placed on test recordings. Thus, the full age range of 21 to 96 was used (as opposed to the constrained set used for subspace estimation). All trials were 30 seconds in duration. Scores were obtained for each trial before and after applying eigenageing compensation.
- Imposter scores were obtained for the test speaker with two separate sets of 25 male speakers from the TCDSA-FD database. The speakers in each set were equally distributed across the five age groups in the database: 25-35, 36-45, 46-55, 56-65 and 66+. All trials were 30 seconds in duration. Scores were obtained for each trial before and after eigenageing compensation was applied. The scores from one set were used to calculate Z-norm statistics. The scores from the other were used as imposter trials in the evaluation.

A DET plot generated from the pooled scores of all speakers, before and after applying eigenageing compensation, is shown in Figure 5.5. The corresponding EERs (equal error rates) before and after eigenageing compensation are 14.68% and 11.40% respectively. There is good separation between the curves below a false rejection rate of 20%, demonstrating that eigenageing compensation is effective in reducing classification error over a range of operating points. The baseline EER is lower than in the evaluation in Chapter 3, where the EER in a forwards direction is 18.5%. Although a different set of imposters were used in each evaluation, the fact the performance is in the same range indicates that the reduced configuration (8 kHz sampling rate and 512 GMM-UBM components) in this experiment does not have a significant effect. In Section 5.2.3, an eigenageing compensation experiment involving males and females, with a comparison to the stacked classifier approach, is presented.

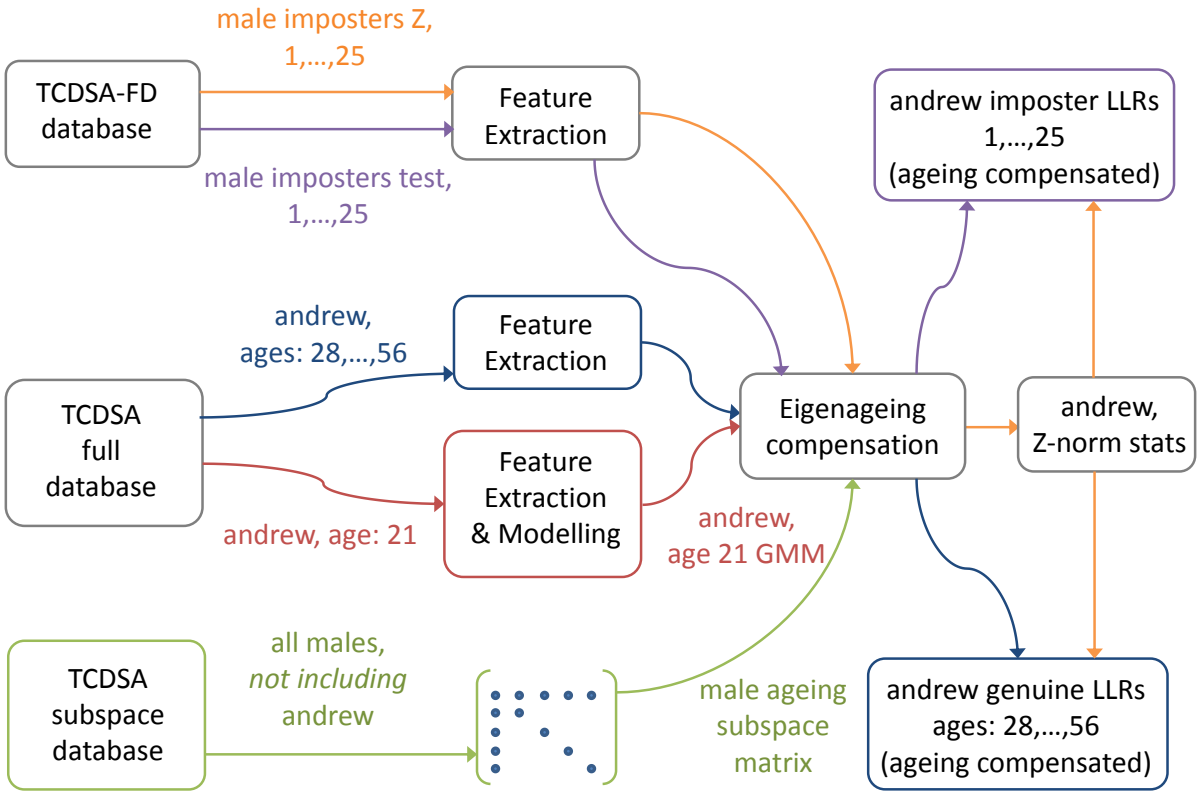

Figure 5.4: The eigenageing compensation training and testing framework for one example male speaker, Andrew. The colour of the arrows correspond to the processing chain for each dataset. The TCDSA full database refers to the full range of data, Figure 3.13. TCDSA subspace database refers to the constrained database, Figure 5.1. In Andrew's case, age 21 is his youngest recording, and hence used for training. The male ageing subspace is estimated as in Section 5.1.2, and eigenageing compensation is applied as in Section 5.1.3. Z-norm statistics are obtained for Andrew, GMM age 21, from a set of 25 TCDSA-FD males. A separate set of 25 TCDSA-FD males provide imposter trials. With the cross-validation approach, this process was repeated for each of the 15 TCDSA male speakers in turn.

5.2.3 Eigenageing compensation male and female evaluation

An evaluation with comparable experiments for both males and females was then designed. As mentioned in the previous male-only experiment, an additional database for imposters and Z-norm speakers is required due to the use of the TCDSA speakers both as ageing subjects and for ageing subspace modelling. The male-only TCDSA-FD database was used for this purpose in Section 5.2.2.

As a source for comparable male and female data, spontaneous speech passages from the CSLU (Section 3.5.5.1) database were used. As CSLU contains exclusively telephone speech of American-accented males and females, it is not ideally matched to the TCDSA database. However, the purpose of the experiment is to compare the relative performance of speaker verification with and without eigenageing compensation. The effect of 'database mismatch' will

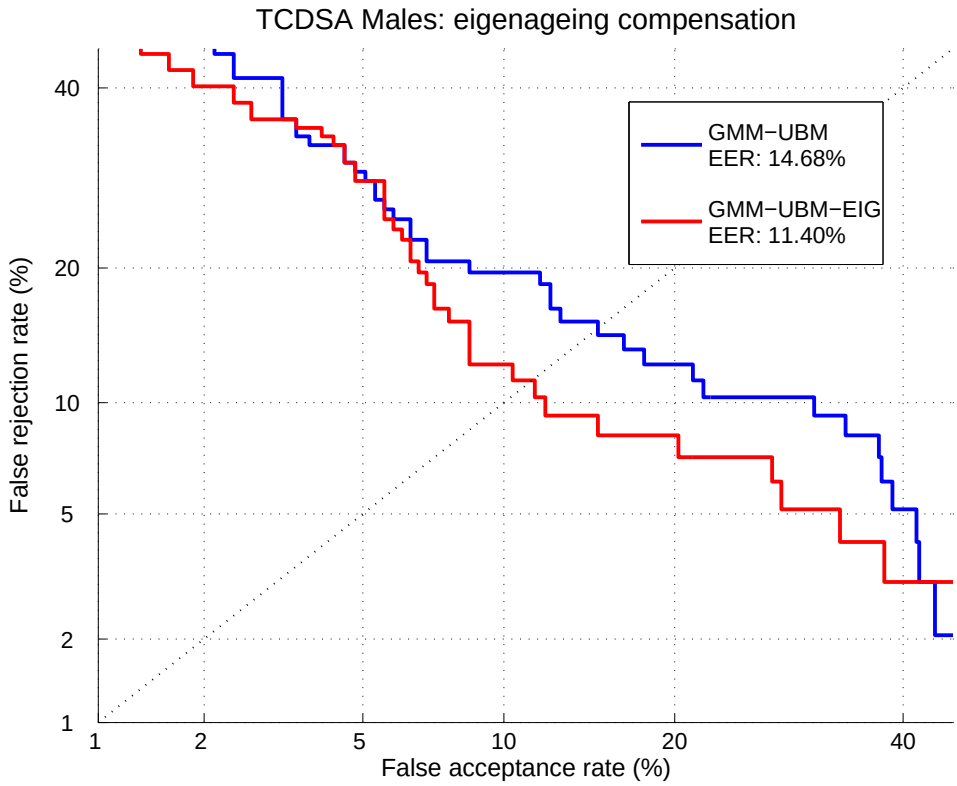

Figure 5.5: DET plot of the pooled scores of the TCDSA males for the baseline GMM-UBM system and for the GMM-UBM system incorporating eigenageing compensation (GMM-UBM-EIG)

be present in the scores of each of these configurations, allowing for a fair comparison. The effect of the this mismatch on the absolute error rates is of less importance. The influence of the CSLU data on the experiment is discussed further in Section 5.2.5.

The experimental design (and results) are those presented in [106]. The protocol is similar to the male-only evaluation in Section 5.2.2, with the main difference being in the databases used. The TCDSA males (15) and females (11) were used both as subjects and for ageing subspace estimation. Gender-dependent UBMs were trained with one hour of data, with 30 minutes sourced from the TCDSA-UBM database and 30 minutes sourced from CSLU (3 minutes from each of 10 speakers). Additional sets of CSLU speakers were used as imposters and Z-normalisation speakers. There was no overlap between the UBM subset and these additional subsets.

A cross-validation approach was again taken: taking each of the 15/11 TCDSA males/females as test speakers, in turn, the remaining 14/10 were used to train a gender-dependent ageing subspace matrix. A schematic of the experiment is provided in Figure 5.6.

A DET plot generated from the pooled scores of all speakers, before and after applying eigenageing compensation, is shown in Figure 5.7. DET plots generated by pooling the scores of each gender separately are given in Figure 5.8. The EER in each case is provided on the plots.

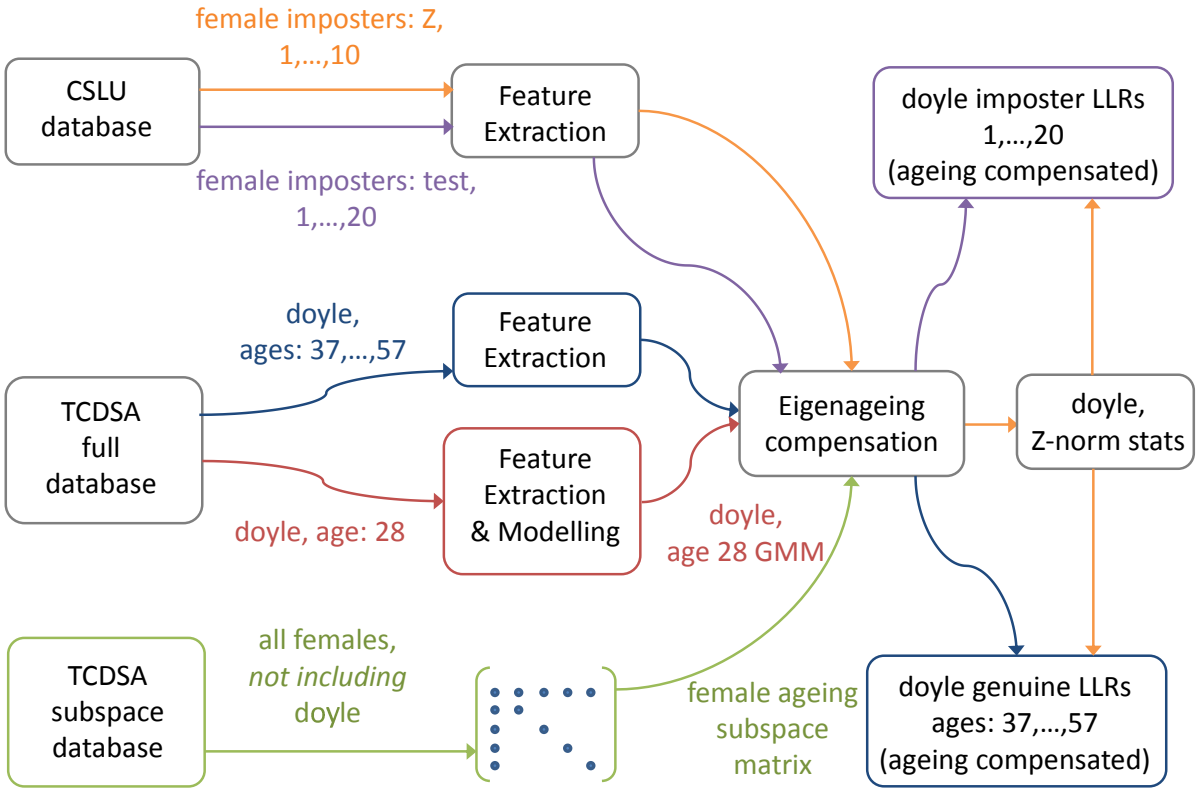

Figure 5.6: The eigenageing compensation training and testing framework for one example female speaker, Doyle. The colour of the arrows correspond to the processing chain for each data source. The TCDSA full database refers to the full range of data, Figure 3.13. TCDSA subspace database refers to the constrained database, Figure 5.1. In Doyle's case, age 28 is her youngest recording, and hence used for training. The female ageing subspace is estimated as in Section 5.1.2, and eigenageing compensation is applied as in Section 5.1.3. Z-norm statistics are obtained for doyle, GMM age 21, from a set of 10 CSLU females. A separate set of 20 CSLU females provide imposter trials. With the cross-validation approach, this process was repeated for each of the 11 TCDSA females in turn (and in the male case, for each of the 15 TCDSA males).

As observed in the male-only case in Section 5.2.2, below a false rejection rate of 20%, the corresponding false acceptance rate is in general significantly lower after eigenageing compensation for males, females and a combination. The separation of DET curves is particularly large in the female case, where a relative reduction of 56% in EER is achieved.

As mentioned previously, the DET plot and associated EER describe the general discrimination ability of the system given the genuine-speaker and imposter classes. To obtain a more realistic measure of performance, a decision threshold was determined from a subset of the trails using a cross-validation approach, similar to that in the stacked classifier experiments in Chapter 4. For each test speaker, the scores from the genuine-speaker and imposter trials of all *other* speakers of the same gender were used to train the threshold. From the set of genuine-speaker

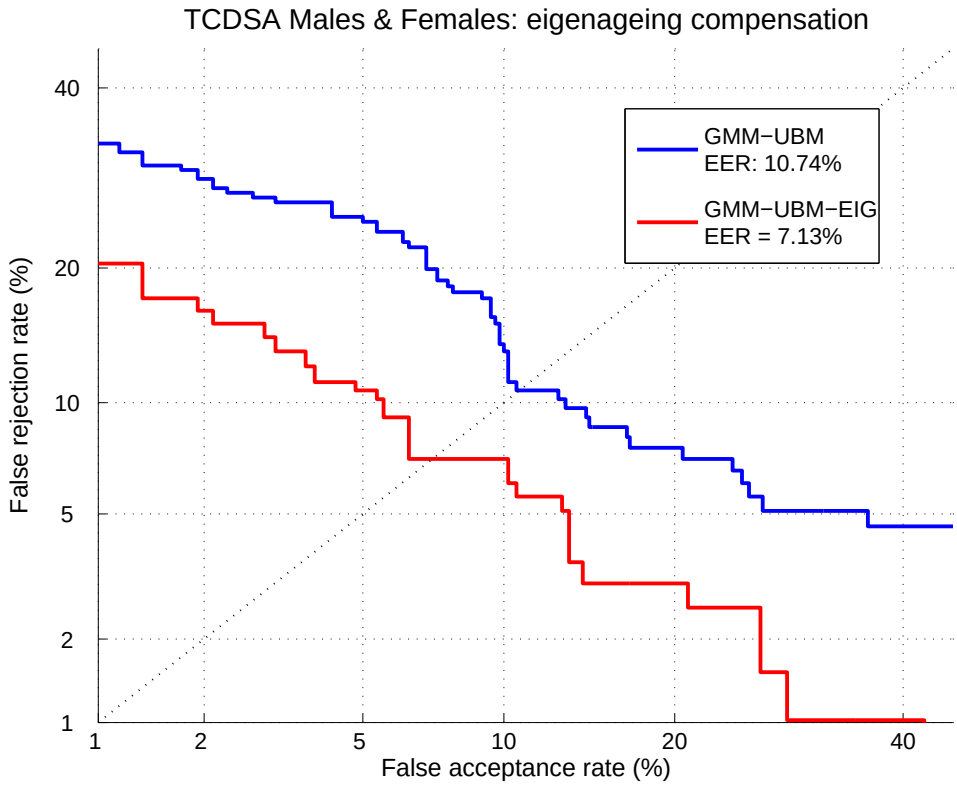

Figure 5.7: DET plot of the pooled scores of males and females for the baseline GMM-UBM system and for the GMM-UBM system incorporating eigenageing compensation (GMM-UBM-EIG)

scores, only those from the same year as the training age, 'age 1', were used. This ensured that no ageing information was present in the decision boundary training set. A threshold was found such that the HTER (half-total error rate) on the training set was minimised.

The LLR scores of an example male and female speaker, before and after eigenageing compensation, are given in Figure 5.9. Also indicated on each plot is the decision threshold and the resulting HTER. The positive shift in the scores of genuine-speaker trials compared with imposter trials, after eigenageing compensation, is evident. An associated improvement in performance can be seen in HTERs.

The average HTER for all 15 males and 11 females is given in the first two rows of Table 5.1. Also included are gender-independent HTERs, determined by pooling male and female LLRs at the decision stage. It is clear that there is a significant relative improvement in HTER after eigenageing compensation: 36% for males, 39% for females, and 27% for a combination.

5.2.4 Comparison of Eigenageing Compensation and Stacked Classification

An evaluation was designed to directly compare the performance of the stacked classifier approach from Chapter 4 with eigenageing compensation. The protocol is the same as that in Section 5.2.3, with the difference being in the decision thresholding step. The stacked classifier

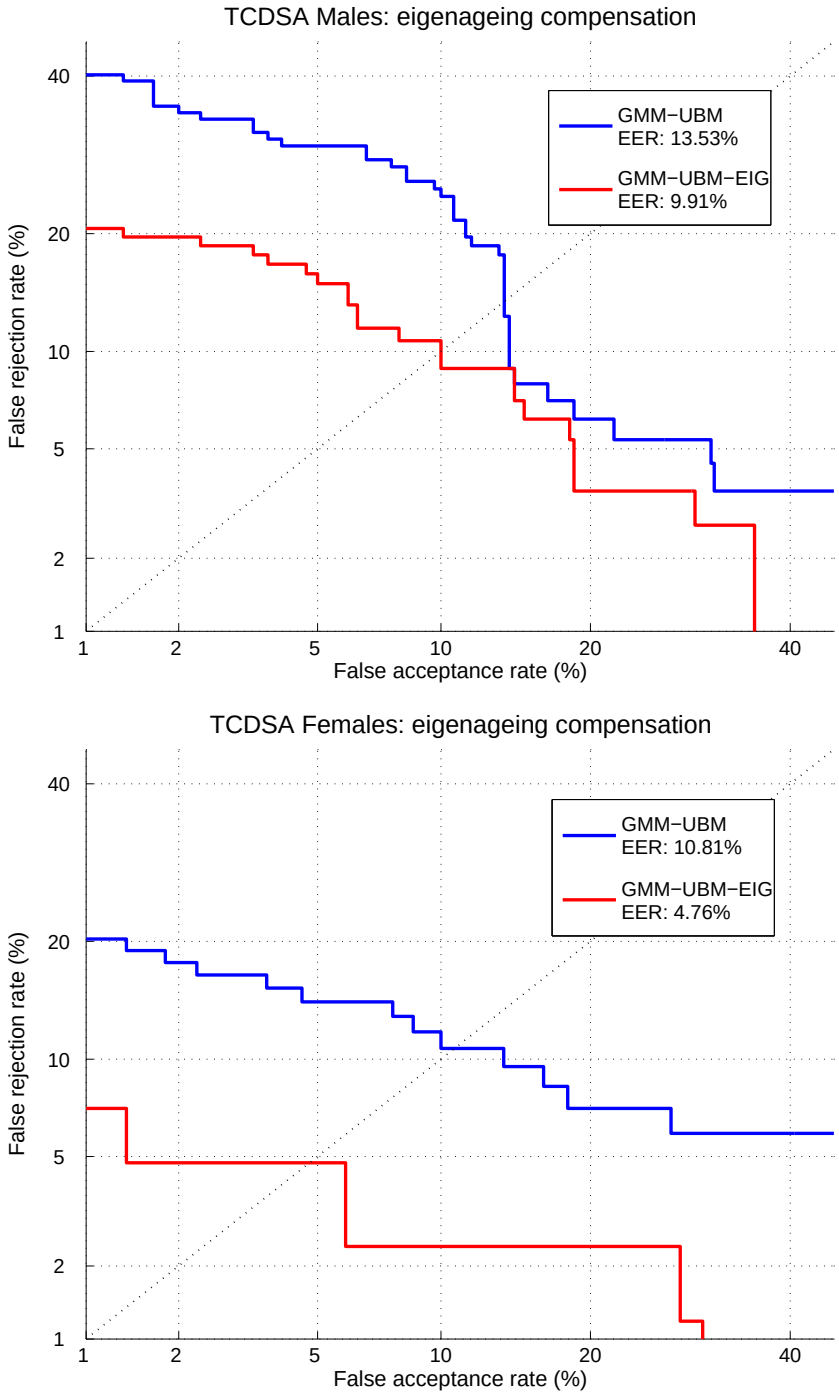

Figure 5.8: DET plots for the baseline GMM-UBM system and for the GMM-UBM system incorporating eigenageing compensation (GMM-UBM-EIG). **Top:** Males, **Bottom:** Females

experiments presented in this Section are the same as those in [106].

The stacked classier was evaluated with a cross-validation decision threshold training scheme, similar to that in Section 4.1.1.2, with the same set of training and testing scores as Section

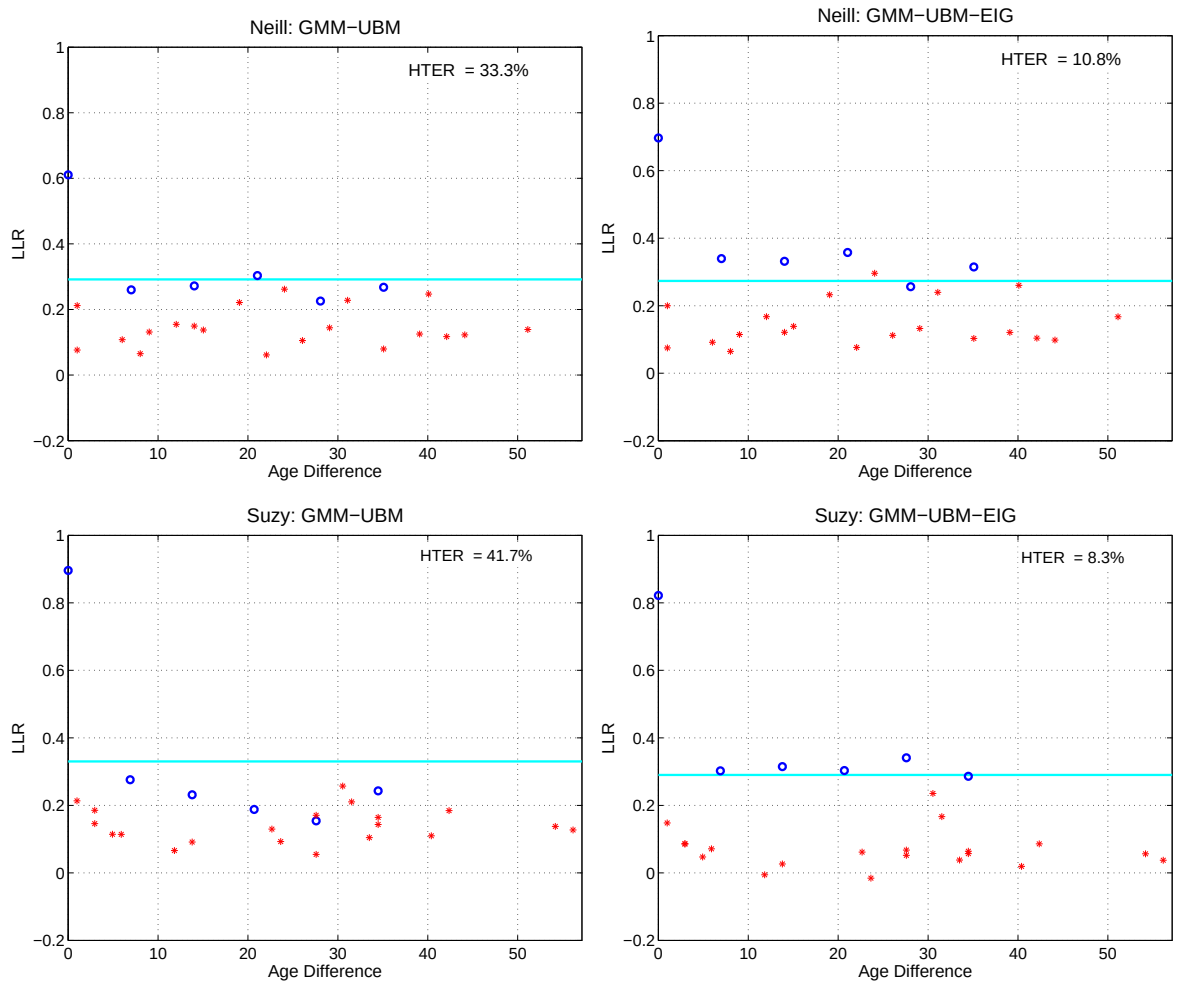

Figure 5.9: Examples of individual speaker LLR scores before and after eigenageing compensation. Genuine-speaker scores (blue circles), Imposter scores (red asterisks) and the decision threshold are shown on each plot: **Top Left:** Neill (male), GMM-UBM, **Top Right:** Neill, GMM-UBM + eigenageing, **Bottom Left:** Suzy (female), GMM-UBM, **Bottom Right:** Suzy, GMM-UBM + eigenageing

5.2.3. Unlike Section 5.2.3 however, for each test speaker, *all* scores from the genuine-speaker and imposter trials of all other speakers of the same gender we used. Thus, ageing information was present in the decision threshold training set. Genuine-speaker scores were associated with their corresponding age differences (i.e. the time-lapse in years between training and testing).

Imposter trials were each assigned a random age-difference within the range of the genuine-speaker trial age differences. To smooth any effects of the random assignment of age-differences to imposters, each of the stacked classifier experiments involving age-difference was repeated with 10 different random age-difference assignments. HTERs presented here are the average of those 10 iterations.

A score-ageing decision threshold was trained with a linear SVM (support vector machine)

and used to classify the test speaker scores. The average HTER for all 15 males and 11 females, and their pooled scores, is given in the third row of Table 5.1. The score-ageing decision threshold significantly improves over the baseline HTERs. A relative improvement of 28% for males, 20% for females, and 18% for a combination is achieved. However, the improvement in all cases is not as great as that achieved with eigenageing compensation. This is particularly the case with females; the relative improvement in HTER with eigneageing compensation is almost double that achieved with the stacked classifier.

	HTER(%)		
	male	female	all
GMM-UBM	19.8	21.4	18.0
GMM-UBM-EA	12.7	13.0	13.2
SC: GMM-UBM + ageing	14.2	17.2	14.8
SC: GMM-UBM + ageing + Wnorm	5.9	7.7	6.6
SC: GMM-UBM-EA + Wnorm	3.9	2.5	3.5

Table 5.1: Average HTER (%) for the baseline GMM system (GMM-UBM), the GMM system with eigenageing compensation (GMM-UBM-EA), and three stacked classifier (SC) configurations combining GMM-UBM/GMM-UBM-EA scores with ageing progression (ageing) and quality (Wnorm)

In Section 4.3, the best performing quality measure was the 'Wnorm' metric 4.2.2. A three-dimensional decision threshold, incorporating score, age-difference and Wnorm was evaluated in the same manner as the score-age threshold. The average HTER for all 15 males and 11 females, and their pooled scores, is given in the fourth row of Table 5.1. These HTERs are significantly lower than those with the score-ageing boundary. The large reduction in HTER by incorporating Wnorm is due to the use of the CSLU database for imposters. In Section 5.2.5 this point is addressed further.

To compare this score-ageing-quality stacked classifier with eigenageing compensation, eigenageing compensated scores were input into the stacked classifier, along with Wnorm. Assuming that eigenageing compensation reduces the ageing variability, then these scores combined with quality can be directly compared with a score-ageing-quality stacked classifier. The average HTER for all 15 males and 11 females, and their pooled scores, is given in the bottom row of Table 5.1 for this configuration. Again, very low HTERs are achieved by incorporating Wnorm into the stacked classifier. The performance of interest here is not the improvement over the baseline, but rather the relative improvement of the eigenageing score-quality stacked classifier over the baseline score-ageing-quality stacked classifier. In this comparison, the combination of eigenageing scores and Wnorm brings a relative reduction in HTER of 34% for males, 68% for females, and 47% for a combination.

5.2.5 Discussion

Eigenageing compensation provides significant improvement over a baseline GMM-UBM system at the task of verifying ageing speakers, as evident from the HTER improvements in Table 5.1 and the DET plots in Figures 5.5 and 5.8.

It achieves lower HTERs than the stacked classifier approach, with the advantage that no ageing meta-data is required at verification time, i.e. the age of the speaker at time of verification does not need to be known. This also protects against the possibility of an imposter attack by manipulating ageing information (although the experiments investigating this in Section 4.1.1.4 suggest that it would not be an very effective means of attack). Furthermore, the eigenageing subspace training in the experiments here used less data (a maximum of 6 recordings in the age range 20-60 per speaker) than used for the stacked classifier boundary training.

The performance of eigenageing compensation is likely to improve with additional data. Considering that implementations of the related eigenchannel compensation technique have used thousands of recordings of speakers in different conditions to model a channel subspace [29], and here there is at most 100 ageing-difference supervectors used to model the ageing subspace, there is certainly scope for refinement.

The low HTERs achieved with the addition of Wnorm, Table 5.1 are very optimistic, as the use of different datasets for imposter (CSLU) and genuine-speaker trials (TCDSA) maximises the discrimination ability of the quality measure Wnorm; TCDSA is exclusively microphone data and CSLU is telephone. As mentioned in the experiment outline, the reason that the TCDSA speakers were not used as imposters in a cross-validation approach was that they were already in use in the cross-validation scheme for ageing subspace training. With a more closely matched test, the improvement with Wnorm would not be as dramatic. However, it is the relative comparison between Wnorm and eigenageing scores, and Wnorm and baseline scores with age-differences, that is of interest.

The DET plots presented in this Chapter were generated from pooled sets of genuine-speaker and imposter scores from all speakers. Since the distribution of recordings in the TCDSA database is not equally spread across speakers, some speakers have a greater influence on the DET curve (and hence the EER) than others. Thus, while the DET plots describe the overall discrimination given the test scores, the performance may be different given balanced contributions from individual speakers. The HTER however, was calculated on a per-speaker basis, and averaged to arrive at a speaker-average HTER. The contributions of different speakers are therefore balanced in the HTER, making it a 'fairer' metric for performance assessment in these experiments. In both male experiments, Figures 5.5 and 5.8 (top plot), the DET curves before and after eigenageing compensation converge for a subset of the error range. In Figure 5.5, this occurs above rejection rate of 20%, and in Figure 5.8 (top plot), this occurs at a rejection rate of between 5 and 10%. In the female case however, Figure 5.8 (bottom plot), there is consistent separation of the DET curves. It is possible that the influence of individual speakers is

responsible for the discrepancy between these trends.

An additional explanation is that difference between male and female ageing patterns, presumed responsible for the discrepancy between principal eigenvalues (Figure 5.3), is at play. In addition to the female DET curves, there is a slightly greater relative HTER improvement with eigenageing compensation in females than males (39% compared to 36%), despite the fact there were less females than males used to train the gender-dependent ageing subspaces (11 compared to 14). This a further indication that ageing variability nay be more readily modelled in females than males.

5.3 Eigenageing for age estimation

In the application of eigenageing compensation, the ageing subspace constrains a speaker model to be shifted towards the age of the speaker in the test sample. It is therefore reasonable to assume that this 'shift' represents the age difference between the model and the test sample in some way. Based on this observation, a method of automatic age estimation is presented in this Section.

5.3.1 Age estimation proposal

As detailed in Section 5.1.3, in estimating the ageing-compensated likelihood of a test sample given a speaker model, a low-dimensional vector x is found such that the following expression is maximised: $p\left(O|\lambda + Vx\right)$, where $O = [o_1, o_2, \ldots, o_T]$ is a set of test features, λ is the speaker model and V is the ageing subspace matrix. The dimension of x is $J \times 1$, where J was taken as 20 (Section 5.1.2) and corresponds to the number of principal eigenvectors contributing to V. The 20×1 vector x can be considered an *age factor* in the above expression, as it represents the 'shift' of the speaker model λ towards a test sample O in an ageing subspace. A proposal to use the age factor x for age estimation is presented here.

When eigenageing compensation is applied to speaker verification, a speaker model is adapted given a test sample and an ageing subspace. The age factor in this case represents the relative difference in age between the speaker in training and testing samples. Thus, to estimate the age of the speaker in the testing sample, the age of the speaker in the training sample would have to be known. A more useful age estimation system would operate without prior knowledge of the speaker. To achieve this, a GMM can be trained from the recordings of a set of 'young' speakers, whose mean age is known. Then, if the young GMM is adapted towards the test sample of the unknown speaker, the age factor x can be assessed relative to the mean age of the young GMM. An age estimation experiment, using an age factor determined in this way, is presented in Section 5.3.2.

5.3.2 Age estimation experimental evaluation

A male-only experiment was designed, whereby the ageing subspace was estimated from TCDSA males, and a separate set of males were drawn from the TCDSA-FD database (Section 3.5.3) as age estimation subjects. The pre-processing, feature extraction and GMM-UBM system configurations were the same as described in Section 5.2.1.

5.3.2.1 Extracting age factors for age estimation

As proposed in Section 5.3, a 'young' speaker model, GMM_y, was created, using the pooled recordings of all speakers in the TCDSA-FD database in the age range 25-35. In a similar manner to the testing approach in previous Chapters, a *backwards* test was also considered, by training an 'old' speaker model, GMM_o, using the pooled recording of all speakers in the TCDSA-FD database over the age of 66. Both GMM_y and GMM_o contained 512 components and were each trained with approximately one hour of speech.

An ageing subspace was trained using *all* TCDSA males in the same manner as Section 5.1.2. This is referred to as the forwards ageing subspace V_f, and was paired with GMM_y in a forwards age estimation test.

An second subspace was trained, for the backwards direction, again using *all* TCDSA males. A backwards ageing subspace was estimated given a backwards ageing supervector (SV) difference matrix, created as shown in Figure 5.10. It is essentially the reverse of the forwards procedure (Figure 5.2): for each speaker's set of SVs, a set of backwards ageing difference SVs were found by subtracting from each speaker's age N (i.e. 'oldest') SV their younger (i.e. age 1, age 2,$\cdots$, age $N-1$, where $N \leq 6$) SVs. A second and third set of ageing-difference SVs were then found by subtracting from each speaker's age $N-1$ and age $N-2$ SVs their younger SVs. A backwards ageing subspace V_b was then derived with the same PCA analysis as in Section 5.1.2, and was paired with GMM_o in a backwards age estimation test.

In the forwards age estimation experiment, given GMM_y and the forwards ageing subspace V_f, the remaining males in the TCDSA-FD database not used for training GMM_y (and hence all over the age of 35) were used as test subjects. For each test, GMM_y was eigenageing adapted given the forwards ageing subspace V_f and a 30 second speech sample.

In the backwards age estimation experiment, given GMM_o and the backwards ageing subspace V_b, the remaining males in the TCDSA-FD database not used for training GMM_o (and hence all below the age of 66) were used as test subjects. For each test, GMM_o was eigenageing adapted the given the forwards ageing subspace V_f and a 30 second speech sample.

This resulted in a set of forwards and backwards age factors, x_f and x_b respectively, for which the corresponding actual speaker ages were known.

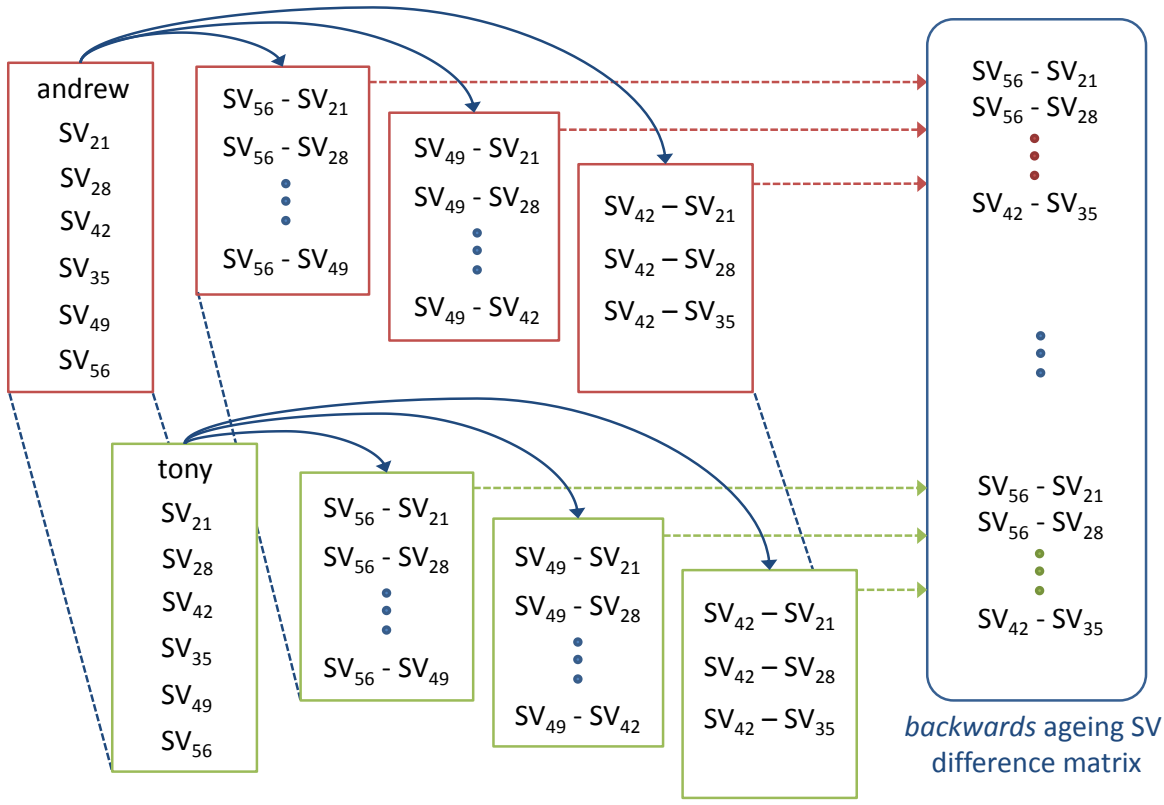

Figure 5.10: The three-stage difference approach to supervector normalisation, backwards direction: Taking Andrew (male) as an example: GMM supervectors (SVs) are created for each of his reordings at the ages denoted in Figure 5.1. This results in a set of SVs from age 21-56: [SV_{21},...SV_{56}]. The differences between the oldest three SVs and the previous SVs are pooled together in the ageing SV difference matrix. This process is repeated for every other male speaker in Figure 5.1, with each contribution included in the overall backwards ageing SV difference matrix.

5.3.2.2 Estimating age from age factors

The set of forwards and backwards age factors were divided into training and testing sets. Given the training age factors and their corresponding actual ages, Gaussian Mixture Regression (GMR) was applied to estimate the ages of the speakers represented in the age factor testing set.

Gaussian mixture regression (GMR) [31, 186] is a means of multivariate regression utilising GMMs. An example of an application for GMR is to predict a spatial value given a temporal value [31]; A GMM is first estimated from a set of training data (temporal and spatial data vectors). The expected spatial values given a range of temporal data samples are estimated from the GMM components. A generalised relationship between temporal and spatial values is established in this way. Given test data with one dimension missing, e.g. temporal data only, GMR outputs the expected value of the remaining dimension, e.g. the spatial data. In

the context of the age estimation experiment, a GMM trained from a set of age factors and corresponding actual ages can be used to estimate age given a set of test age factors. Prior to training the GMM, age factors (vectors of size 20×1) were reduced in dimensionality by extracting several 'features':

1. **first:** the first element of the age factor
2. **mean:** the mean of all elements in the age factor
3. **first & mean:** the first element on the age factor concatenated with the mean.
4. **top 5:** the first 5 elements of the age factor
5. **top 10:** the first 10 elements of the age factor

Each feature, along with the actual speaker age, were used to train a full-covariance GMM. The test age factor features were used to output a set of estimated ages. The mean absolute error (MAE), a standard error metric for speaker age estimation [10, 49], was calculated given the estimated and actual ages.

In the forwards and backwards directions there were 88 and 94 age factors respectively. In each case, a leave-one-out cross-validation protocol was applied, whereby each age factor was held out in turn, and the remaining 87/93 were used for training the GMM. The experiment was also run for the combination of forwards and backwards age factors.

5.3.2.3 Experimental Results

The MAE, averaged over all iterations, is shown in Table 5.2 for the different age factor features. Also shown is a 'baseline' MAE, which represents the performance of a simple age estimator that always outputs the mean age of the training data. The number of GMM components was optimised experimentally in each case, and varied between 1 and 5.

	MAE(years)		
	combined	forwards	backwards
N	182	88	94
baseline	11.07	10.38	10.41
first	10.62	10.34	10.52
mean	11.08	10.03	8.66
first, mean	10.43	10.49	9.38
top 5	10.79	10.57	10.45
top 10	10.47	10.93	8.93

Table 5.2: Average MAE of age estimation using forwards and backwards age factors and their combination. N is the number of leave-one-out cross-validation iterations.

In Figure 5.11, the pooled estimated ages from all cross-validation iterations are plotted against their corresponding actual ages for a selected configuration from Table 5.2.

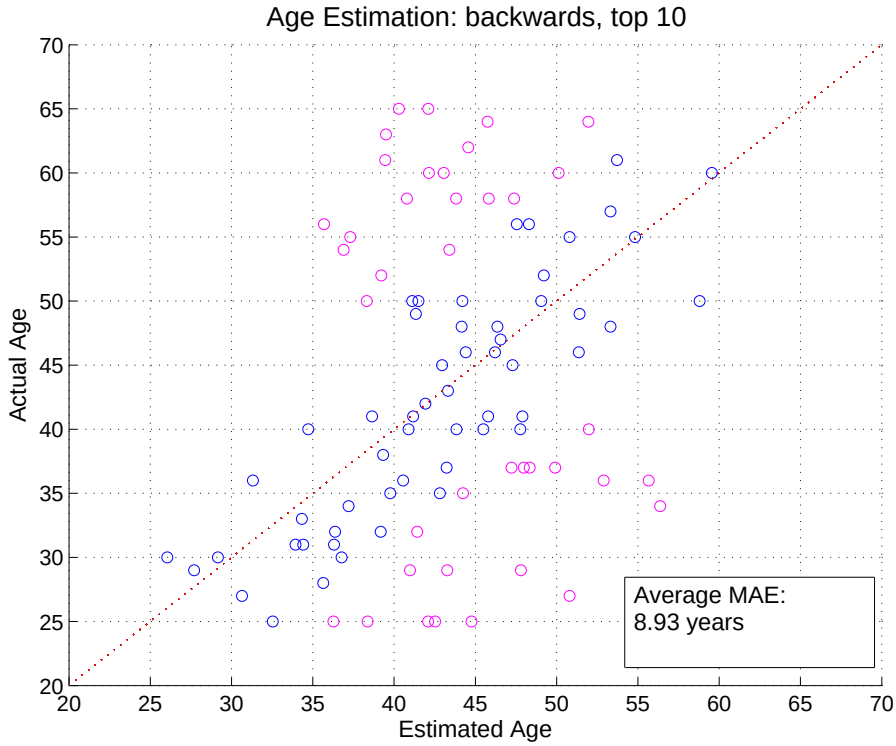

Figure 5.11: Age estimation using the 'top 10' backwards age factors as features for GMR. Each point on the plot represents the output of a cross-valdation iteration. The average MAE across all iterations is 8.93 years in this case. The blue circles indicate estimates below or equal to 8.93, and the magenta circles indicate estimates greater than 8.93

In the combined, forwards and backwards tests, a number of features achieve better ageing estimates than the baseline estimate. This confirms that the age factors carry absolute age information. In both forwards and backwards tests, the mean of the age factor is the best performing feature, with an MAE of 10.03 and 8.66 years respectively. The 'top 10' feature also performs well in a backwards direction, with an MAE of 8.93 years.

5.3.3 Discussion

The successful application of an age factor, derived from eigenageing compensation, for age estimation, demonstrates that ageing progression is modelled by the ageing subspace. As discussed in relation to eigenageing compensation for speaker verification, additional data to improve the ageing subspace model is likely to improve age estimation performance.

There is a noticeable difference between forwards and backwards test results in Table 5.2, particularly in the case of the larger 'top 5' and 'top 10' features. A likely reason for this is

the presence of older speakers (66+) in the forwards testing set. All speakers in the backwards testing set were under the age of 66. As discussed previously, there is generally larger variability between the voices of older speakers, making the forwards case a more challenging test set.

A recent study on automatic speaker age estimation by Bahari et al. [10] presented an i-vector approach and compared it to other recent approaches using GMM supervectors, including those of Dobry et al. [49]. The test data was sourced from the NIST SRE databases 2010 and 2008. On a set of 3999 male test utterances, the baseline MAE (determined in the same way as the result in Table 5.2) was 10.39 years, the MAE for a GMM-SVR (support vector regression) system was 8.02, and the MAE for the best performing i-vector system was 7.63 years. Thus the relative percentage improvement over the baseline system was 22.8% with the GMM-SVR approach and 26.6% with the i-vector system. The best performing GMR approach here, resulted in an 16.8% improvement over the baseline result.

The i-vector approach in [10] used a feature set of 20 MFCCs with Δ and Δ^2 coefficients and "over 30,000" speech recordings for training the UBM and total variability subspace. Considering the vastly smaller quantity of data used here for ageing subspace training, the results are promising. In addition, the datasets used by Bahari et al. are biased towards speakers within the age-group 25-35 (illustrated by the histograms, Figure 2 in [10]). Thus, the performance of the system in their evaluation is not strongly affected by the age estimation accuracy of older, more challenging, speakers.

An interesting observation from [10, 49], is that the age estimation of female speakers is more accurate than that of males, for almost all system configurations evaluated. As mentioned in Section 5.2.5, indications from the eigen-analysis and subsequent eigenageing compensation evaluation of females is that the female ageing subspace is a closer representation of female ageing progression than the male case. An evaluation of this age estimation approach with female speakers is therefore a priority for future work.

This Chapter has demonstrated that a model of ageing progression can improve long-term speaker verification performance, and enable a promising approach to automatic speaker age estimation. In Chapter 6, the impact of ageing on forensic automatic speaker recognition is assessed, and the effectiveness and suitability of eigenageing compensation for the forensic domain is evaluated.

6

Forensic speaker recognition and ageing

In Chapters 3, 4, and 5, the effect of ageing on speaker verification was explored, and two approaches for compensating for its detrimental effect on performance were presented. In this Chapter, the effect of vocal ageing in a forensic speaker recognition context is investigated.

Forensic speaker recognition is the task of determining whether two (or more) recordings of speech originate from the same speaker. Thus, the process involves the comparison of a recording of an unknown speaker in an evidential recording with one or more recordings of a known suspected speaker [52].

As discussed in Section 2.5.1, the application of automatic speaker recognition methods to forensic applications is well established [6, 52, 53, 74, 76, 141, 143, 158, 172, 173]. In forensic automatic speaker recognition (FASR), acoustic parameters of an unknown voice in an evidential recording are compared with a statistical model derived from the acoustic parameters of a suspected speaker's voice [52]. The comparison results in a likelihood ratio (LR) representation of the strength of the evidence, which is becoming increasingly accepted as the "logically and legally correct" [73] way for forensic experts to present evidence to the court.

As detailed in Section 2.5.1, other approaches in *technical* forensic speaker recognition are the 'auditory-perceptual' method, which is essentially a careful listening analysis undertaken by a trained phonetician, resulting in a subjective measure of the strength of similarity between two voices, and the 'acoustic-phonetic' method, which involves comparison of the statistical characteristics of various acoustic features, resulting in an objective measure of the strength of similarity between two voices. *Naïve* speaker recognition refers to the everyday ability of people to recognize voices they have heard previously.

Ageing is of direct relevance to forensic speaker recognition, as the speech samples involved are always separated by a period of time. In the event of a large time-lapse between an evidential recording and a suspected speaker recording, it is important to establish to what extent ageing affects the outcome of the forensic comparison. This is particularly important for FASR, where the influence of ageing on the LR must be established.

In Section 6.1, a review of previous studies involving forensic speaker recognition and ageing is presented. In Sections 6.2.1 and 6.2.2, the male TCDSA database speakers are used to establish the impact of ageing on a GMM-UBM FASR system.

In Section 6.2.1, a general evaluation of the effect of ageing on long-term LR estimation is considered, using all TCDSA males as subjects and the TCDSA-FD database for background modelling and reference populations. These experiments were previously presented in [111]. In Section 6.2.1, a detailed evaluation of the effect of ageing on LR estimation, replicating a realistic forensic scenario, is presented: for all Irish-accented males in the TCDSA database, the LR is evaluated for all combinations of suspected speaker age and evidential recording age, using an age-dependent reference population. The effect of eigenageing compensation, Chapter 5, is then evaluated in this FASR LR estimation scenario.

In Section 6.3, an experiment investigating the detectability of vocal ageing by human listeners is presented. Using a subset of the TCDSA database speakers as subjects, a listening test was designed to assess the ability of naïve listeners to detect the presence of ageing in the voice of the same speaker at different ages. The results offer an insight into the perception of the ageing voice. A subset of the listening test results were previously presented in [111].

6.1 An overview of forensic speaker recognition and ageing

The nature of the forensic scenario means that there is always non-contemporaneity between the evidential and suspected speaker recordings [122, 172]. For example, in a typical scenario of recovering voice evidence from telephone surveillance, some time will pass before a suspect is identified and recorded. The duration of this time-lapse varies case-by-case. Police reports generally indicate the extent of this delay, and thus this information is available to the analyst [165]. According to J.P. French, chairman of J.P. French Associates forensic speech and acoustics laboratory [103], the typical time-lapse for cases in the UK is 2-5 months (pp 20 in [165]). Eriksson notes that time-lapses of a year or more are not unusual [59]. Cases with significantly longer time lapses do occur occasionally [165], an extreme example being the Yorkshire Ripper Hoaxer trial [63], where the interval between the evidential and suspected speaker recordings was 26 years. Thus, the effect of these very long time-lapses, where ageing is likely to play a strong role, is of relevance to both automatic, auditory-perceptual and acoustic-phonetic forensic speaker recognition.

Other scenarios in the forensic domain where the effect of ageing is of relevance include: speaker profiling cases where the reference recording is not recent [165]; the composition of

a reference population speaker set (Section 2.5.5) for FASR where there is a large time-lapse between evidential and suspected speaker recordings; and the selection of foils for an earwitness line-up [59] where there is a long delay between the crime-scene and line-up events.

Hollien and Schwartz [92, 94] studied the effect of non-contemporary speech samples on auditory-perceptual speaker recognition. Using recordings from a set of speakers at intervals ranging from 4 weeks up to 20 years, the ability of lay listeners to identify same-speaker pairs was evaluated. The percentage of correct decisions decreased from 95% in the contemporary condition to around 80% at 4 weeks. At intervals of 8 weeks, 32 weeks and 6 years the correct decision rate remained stable, before dropping significantly at a 20 year interval to 33%. Repeating the experiment with a set of trained phoneticians as listeners, the correct decision rate dropped to only 74% after 20 years. Hollien and Schwartz conclude from this result that "noncontemporary speech samples can be used effectively in nearly all types of speaker identification". This point will be revisited in Section 6.4.

Another scenario where auditory-perceptual speaker recognition is affected by the passing of time is an earwitness line-up. In this case, the witness must make a comparison between a voice they heard at a crime-scene and a selection of voices in a line-up [59, 93]. The ability of a listener to remember a voice has been shown to decay progressively with time [59]. In an early study by McGehee [138], the correct recognition rate of listeners dropped progressively to 80% after one week, 69% after two weeks, 35% after three months and to 13% after 6 months. Subsequent studies, including that by Papcun et al. [149] supported this degradation in listener identification rate with time. A recent study simulating realistic earwitness scenarios, with a two week delay between hearing the voice and listening to a line-up, found that the correct recognition rate of both adults and children was below chance level [147]. A shortcoming with these studies is that they do not recreate the stress experienced by the witness in a real-life event [59], which may impair, or possibly improve [93], the ability of the witness to recognise the voice of the perpetrator. In any case however, it seems that auditory memory, rather than vocal ageing of the suspect, is the limiting factor in an earwitness lineup.

One of the earliest forensically-motivated longitudinal studies of the voice, by Endres et al. [58], analysed formants and fundamental frequency measurements of four males and two females over a 13-15 year period at five year intervals. It was found that the frequency of the point of concentration of formants decreased with age for all speakers. The mean fundamental frequency and its variability was also shown to decrease for all speakers. The motive for this study was to assess how ageing would effect speaker recognition based on 'voiceprints'. Although the visual examination of voiceprints (spectrograms) to identify speakers is now widely discredited [59], the findings of Endres et al. are of relevance to the current acoustic-phonetic approach, whereby fundamental and formant frequency measurement play an integral part. Thus, other (non-forensically-motivated) studies on the acoustic correlates of the ageing voice, e.g. [160], along with the experimentally-measured acoustic correlates of ageing on the TCDSA database, Section 3.4, are of relevance to the forensic acoustic-phonetic approach.

French et al. [63] document the forensic speaker recognition procedure undertaken in the 'Yorkshire Ripper' Hoaxer investigation. The case involved the comparison of two recordings: that of an hoax call made to a police station by a man purporting to be the Yorkshire Ripper murderer, and a recording of a suspect, John Humble, who in police custody 26 years later, read the transcript of the original hoax call. French and Harrison, the forensic investigators assigned to the case, carried out an auditory-perceptual and acoustic-phonetic analysis of the speech on the recordings. Based on the voice quality, rhythm, intonation and segmental formant and fundamental frequency measurements, the investigators concluded that there was a "very high degree of correspondence" between the samples. Humble admitted to making the hoax call before the forensic analysis was completed. French et al. extend the findings of this case to conclude that, "forensic speaker comparison work on widely non-contemporaneous samples is not only possible but justified". Again, this point will be revisited in the discussion in Section 6.2.2.3.

Rhodes [165, 166] uses formant frequency estimates from four male Up Series [2] subjects to estimate likelihood ratio (LR) scores at increasing ageing intervals of 7, 14, 21 and 28 years. The LR estimates generally decrease with ageing interval. In same-year comparisons, all LR estimates correctly supported the same-speaker hypothesis (e.g. H0, Section 2.5.3), whereas after 21-28 years the majority of estimates incorrectly supported the different-speaker hypothesis (e.g. H1, Section 2.5.3).

Künzel [122] presents a forensic automatic speaker recognition (FASR) evaluation of 10 males, with recordings 11 years apart. The estimated LRs at an 11 year time-lapse all correctly show support for the same-speaker hypothesis (H0). The LRs for only two speakers drop below a value of 10^2, 'moderate support' on a standard verbal scale, after the 11 year interval, and for only one of these speakers, the decrease with respect to their contemporary LR is significant. Künzel's findings will be discussed in Section 6.2.2.3, in relation to the experiments in this Chapter.

Rhodes [165, 166] investigates the effect of ageing on FASR with six male Up Series [2] subjects. Five suspected speaker models are trained for each speaker, using recordings from ages 21-49, at seven year intervals. For each model, the remaining different-year recordings for the speaker are treated as evidential recordings, and used to estimate a LR (this experimental design is adopted in Section 6.2.2). A general trend of decreasing LR with age interval is observed, with some speakers affected more than others. For most of the LR estimates at an age interval of 21-28 years the estimate is below a 'moderate support' value of 10^2 (Table 2.1) or incorrectly shows support for the different-speaker (H1) hypothesis. Rhodes' experiments will be further discussed in Section 6.2.2.3, in relation to the experiments in this Chapter.

6.2 Forensic automatic speaker recognition of ageing males

In this Section, the male subjects from the TCDSA database are used to evaluate the effect of ageing on likelihood ratio (LR) estimation with a forensic automatic speaker recognition

(FASR) system. The experiments presented here expand on previous research by Künzel [122] and Rhodes [165, 166].

A GMM-UBM based FASR system, described in Section 2.5.5, is used in this Chapter. In Section 6.2.1, an evaluation of the effect of increasing time-lapse on LR estimation is presented, using all TCDSA males as subjects. In Section 6.2.1, a detailed evaluation of the effect of ageing on LR estimation for Irish-accented males, representing a more realistic forensic scenario, is presented. In Section 6.2.2, eigenageing compensation, as introduced in Chapter 5, is applied in the forensic scenario.

6.2.1 Likelihood ratio estimation of ageing males

An experiment was designed to evaluate the effect of increasing age interval between suspected speaker and evidential recordings on FASR, using the 15 males from the TCDSA database as subjects. The experimental design was structured in the same way as the initial evaluation of ageing on GMM-UBM speaker verification, Section 3.8.2, considering both forwards (testing sample is older than the training sample) and backwards (testing sample is younger than the training sample) scenarios.

As mentioned previously, the backwards direction is of more relevance to forensics, as it replicates the scenario where a model is trained using a custody recording, and the LR is estimated from an evidential recording from some date previously. The forwards direction is of more direct relevance to speaker verification. However, a situation may arise in a forensic comparison where the evidential recording is more suitable for generating a model than the suspected speaker recording. In this case, a comparison would occur in the forwards direction.

6.2.1.1 Feature extraction and GMM-UBM system configuration

The GMM-UBM configuration used is that in [106, 111] and in Chapter 5, and the FASR setup is as described in Section 2.5.5. The UBM-screened version of the TCDSA database, Figure 3.13 was used in this Section.

All speech was downsampled to 8 kHz after silence removal and pre-emphasis. MFCC features of length 12 were extracted over 20 ms windows with 50% overlap and a Mel filterbank of 26 bands. Mean and variance normalization were applied to the features followed by RASTA filtering. Finally, delta coefficients were appended.

A 512-component UBM was trained with 2.5 hours of speech from the (male-only) TCDSA-FD database, distributed evenly across age ranges, i.e. 30 minutes from each of the five age ranges: 25-35, 36-45, 46-55, 56-65 and 66+.

Z-norm was applied to all test scores, using statistics calculated from a set of 25 speakers from the TCDSA-FD database (separate from UBM set), distributed evenly across age ranges i.e. five speakers from each of the five age ranges: 25-35, 36-45, 46-55, 56-65 and 66+.

6.2.1.2 FASR evaluation

- 'Suspected speaker' GMMs were trained for each of the 15 TCDSA males' youngest and oldest recordings by adaptation from the UBM, given one minute of speech.
- Each suspected speaker GMM was tested with the remainder of its training recording, in 30 second segments. After applying Z-norm, the scores were used to estimate a single-Gaussian within-source (WS) distribution (Section 2.5.5) for each suspected speaker GMM.
- For each of the 15 TCDSA males, their recordings *not* used as suspected speaker recordings were treated as 'evidential recordings'. A 30 second segment from each was tested against their oldest and youngest GMM, resulting in a set of forwards and backwards evidential scores for each speaker.
- A set of 50 'reference population speakers' (Section 2.5.5) were taken from the TCDSA-FD database (separate from UBM and Z-norm sets), distributed evenly across age ranges, i.e. ten speakers from each of the five age ranges: 25-35, 36-45, 46-55, 56-65 and 66+. A GMM was trained for each reference population speaker in the same way as the TCDSA GMMs: adapting from the UBM with one minute of speech.
- The set of 50 reference population GMMs was tested with a 30 second segment from each of the 15 TCDSA males' evidential recordings. After applying Z-norm, the resulting set of scores were used to estimate a single-Gaussian between-source (BS) distribution (Section 2.5.5) for each evidential recording.
- Within-source degradation prediction (WDP), within-source minimum variance limiting (WMVL) and outlier removal, as detailed in Section 2.5.8, were applied to the WS distribution of each suspected speaker GMM to account for the limited (one recording per year) TCDSA data.
- LR scores were then obtained for each suspected speaker GMM, for each of their evidential recordings. The evidential recording score obtained by testing a 30 second segment against the suspected speaker GMM is evaluated on the WS distribution of that suspected speaker GMM, yielding the numerator of the LR. The evidential recording score is then evaluated on the BS distribution corresponding to the suspected speaker GMM and evidential recording, yielding the denominator of the LR. This process is illustrated in Figure 2.6.
- A schematic of the full experiment is shown in Figure 6.1.

6.2.1.3 Experimental results

In Figure 6.2, the LRs for each speaker are plotted against the age difference between the suspected speaker recording and the evidential recording, in forwards and backwards directions. For

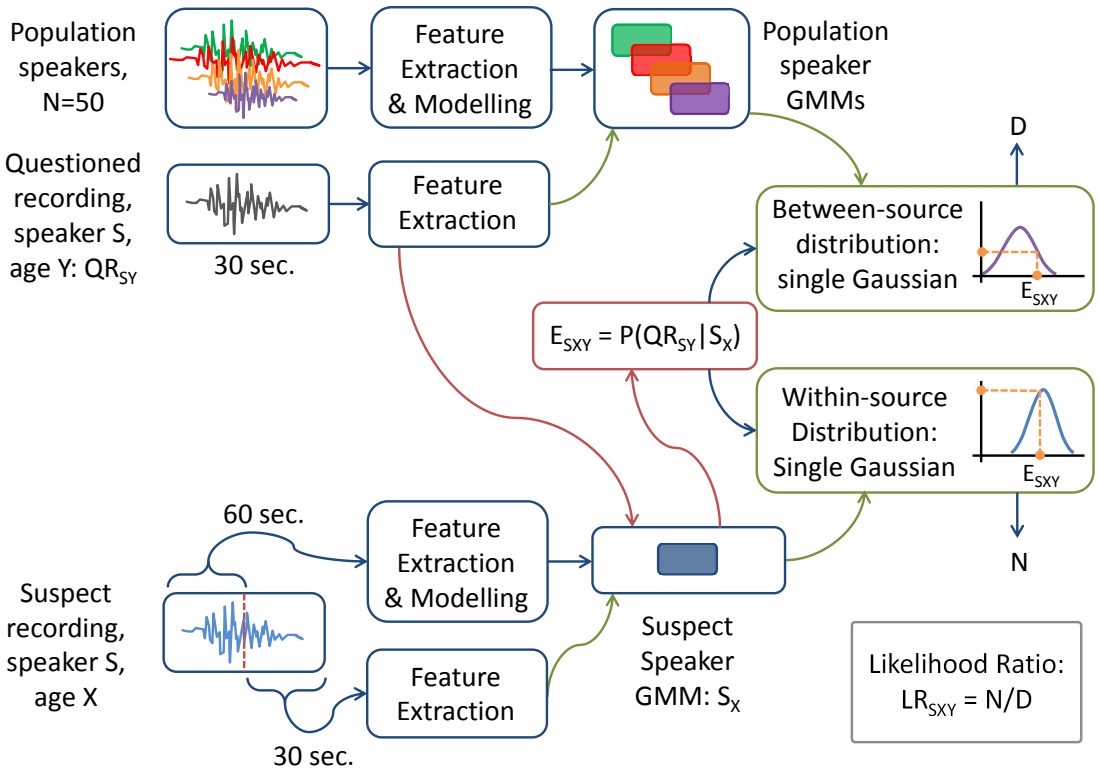

Figure 6.1: A schematic of the experiment described in Section 6.2.1.2. The diagram illustrates the estimation of an LR score for a given suspected speaker and evidential recording. The colour of the arrows correspond to the processing chain for each input. The process is repeated for each of the 15 males in turn, treating their youngest and oldest recordings as suspected speaker recordings and the remainder of their recordings in each case as evidential recordings.

scaling purposes, a log_{10} compression is applied to the LRs. A linear fit has been superimposed on the scores of each speaker.

With only one exception, Magee, the LR scores of all speakers degrade at a relatively consistent rate in both forwards and backwards directions. For the majority of speakers, the LR scores fall below 10^2 within an age difference of 20-30 years, and fall below 10^0 at their maximum range. An LR value of 10^2 corresponds to the minimum value for 'moderate support' on a widely used verbal LR scale [172] (Section 2.5.4). An LR value of 10^0 corresponds to the point at which the same-speaker and different speaker hypotheses (H0 and H1) have equal support. Thus, LR values below a value of 10^0 in this experiment give erroneous support to the different-speaker hypothesis. The results suggest that after 30 years (but likely sooner), the LR of a typical male speaker degrades to a point where its evidential value is significantly weakened.

These results are in contrast to Künzel's [122], who reported that for 9 out of 10 test subjects,

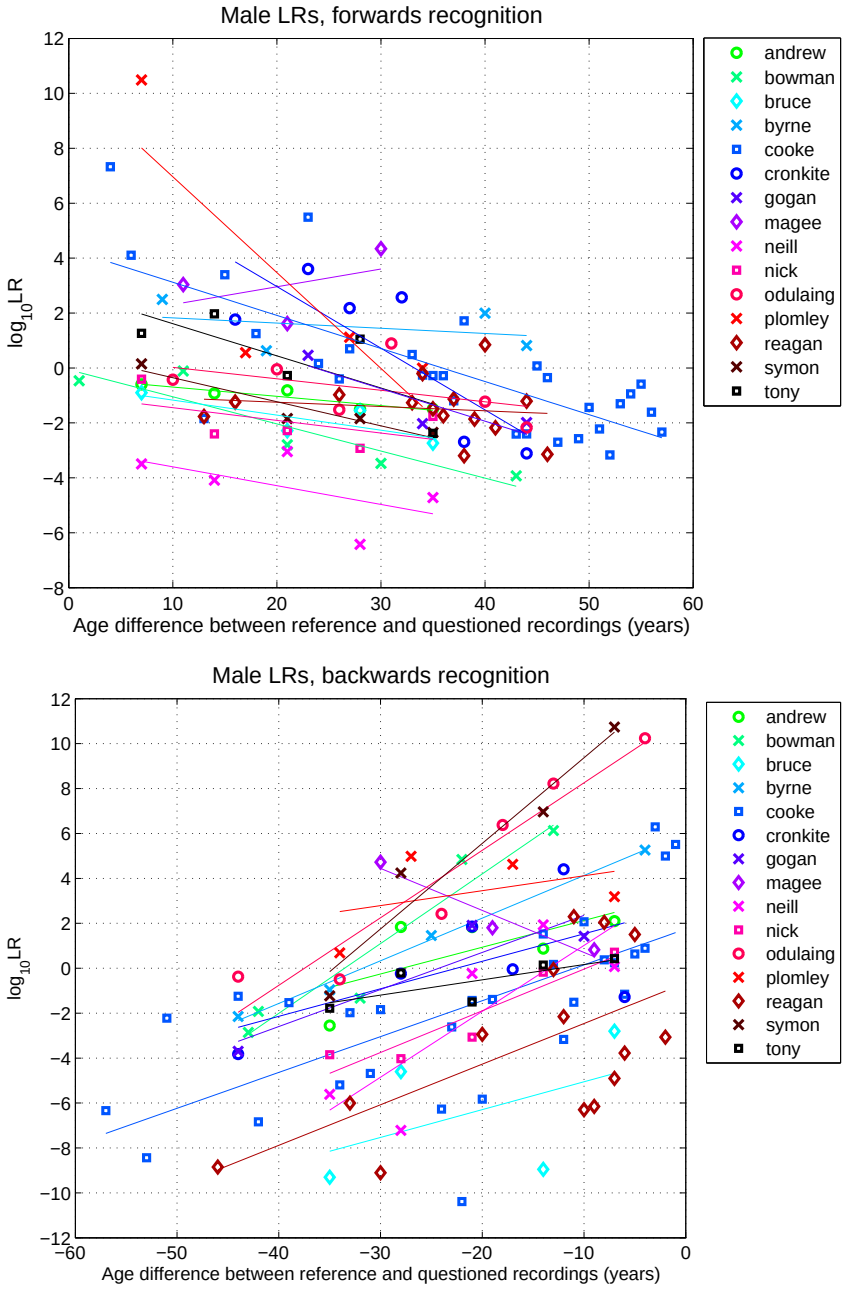

Figure 6.2: $\log_{10}$LR scores against the age difference between suspected speaker and evidential recordings, for all males. Top: forwards direction, Bottom: backwards direction. Note: the y-axis limits in the bottom plot are greater than in the top plot.

same-speaker LRs estimated at an 11 year age difference were not significantly different to same-speaker LRs estimated from contemporary samples. Differences in test recordings and design are the likely reason for this discrepancy.

There is much inter-speaker variability observed in the LRs, e.g. after less than 10 years,

a number of speakers have values below 10^0 in both directions. Aside from the influence of recording conditions, there are several reasons for this variability: the UBM, reference population and normalization sets were composed of Irish-accented speech only.

The issue of accent mismatch between the suspected speaker and the reference population is investigated by Hughes and Foulkes in [97], who show that the LRs from both same-speaker and different-speaker comparisons are higher in the case of a mismatched reference population accent than in the case of a match reference population accent. This factor may contribute to the relatively high LRs of the British-accented Plomley and Cooke in the forwards direction, and Symon in the backwards direction.

Conversely, the use of an Irish-accented UBM likely contributes to lower LRs for some speakers with mismatched accents. The speakers with the lowest LR in both directions are either British or American accented. With an accent mismatch, the adaptation of the UBM given the training sample can place a stronger reliance on the training sample than for a matched accent (due to the 'relevance factor'). This may result in a 'poorer' speaker modelling for some training samples, and consequently, lower LRs given the evidential recordings.

In addition to the issue of accent mismatch, there is also the issue of age mismatch to consider: in the case of an age difference between suspected speaker and evidential recordings, a choice must be made whether to select speakers of similar age to the suspected speaker or to the evidential recording. In this experiment, speakers balanced across a range of ages were chosen. This choice may affect the LR estimation both at extremes of age difference and for very young or very old speakers. This raises broader questions about how to select a reference population given the suspected speaker and evidential recordings, e.g. if a speaker has been geographically mobile or has developed heath problems, should this be reflected in some way in the reference population? This is very much an open issue in forensic speaker recognition.

In Section 6.2.2, an FASR experiment using a subset of Irish-accented speakers is presented, with the aim of reducing a number of the variabilities that influenced the experiment in this Section. By restricting the subjects to be Irish-accented, the accent mismatch with the TCDSA-FD Irish-accented development data is reduced. An age-specific reference population set is introduced, to reduce age-mismatch. All combinations of training and testing ages are considered, removing the reliance on one model per speaker to observe the long-term trend. Finally, all Irish-accented speakers are sourced from the same broadcaster, and thus there is less inter-session variability between their recordings than between the full set of TCDSA males.

6.2.2 Detailed likelihood-ratio estimation of ageing Irish males

The results of the initial forensic investigation in Section 6.2.1 inform the setup of a more detailed experiment: to restrict the variability due to accent, age and environment, in the experiments in this Section, all speech data used is Irish-accented, and the data for the subject speakers is exclusively studio recordings from the same broadcaster. Thus, the TCDSA database subjects

used are the five Irish-accented males, Bowman, Byrne, Gogan, Magee and Odulaing.

The GMM-UBM and FASR system was as in Section 6.2.1.2, with the exception of the between-source (BS) distribution estimation. In Section 6.2.1.2, forwards and backwards recognition was evaluated using a suspected speaker model trained for each speaker at their youngest and oldest ages. Here, this evaluation setup is extended by training a suspected speaker model for each speaker at every available age, and considering the remaining set of their recordings in each case as evidential recordings.

6.2.2.1 FASR evaluation

- Suspected speaker GMMs were trained for each recording from each of the five Irish-accented TCDSA males, by adaptation from the UBM, given one minute of speech.
- For each suspected speaker GMM, the WS distribution was estimated in the same manner as Section 6.2.1.2.
- An age-dependent reference population was chosen for each suspected speaker model: from a set of 50 age-balanced TCDSA-FD males (ten speakers from each of the five age ranges: 25-35, 36-45, 46-55, 56-65 and 66+), the 30 speakers closest in age to the suspected speaker age were selected as reference population speakers. This resulted in a mean absolute age difference of 10.2 years between each suspected speaker model and their reference population speakers. A population size of 30 has been shown to be sufficiently large to produce stable LRs for same-speaker and different-speaker comparisons [99]. The BATVOX software tool [4] (a standard FASR solution for forensic practitioners) uses 35 population speakers as a default setting [195]. To minimise the age-mismatch factor, the slightly reduced number of 30 speakers was used here.
- In Section 6.2.1.2, a single Gaussian was used to model the BS distribution. In this experiment, a GMM of 4 components is used to model the BS distribution. A 32-component Gaussian was used for BS distribution modelling in the FASR system proposed by Gonzalez-Rodriguez et al. [74]. Owing to the limited number of reference population speaker scores (30), the number of components was reduced to 4 in this implementation. The set of GMMs trained with the suspected speaker model's age-dependent reference population are tested with a 30 second segment of an evidential recording. After a Z-norm is applied, these scores are used to estimate the 4-Gaussian BS distribution.
- In addition, four age-dependent imposter recordings were chosen for each suspected speaker model: one recording from each of the other Irish-accented TCDSA males that is closest in age to the evidential recording speaker age.
- LR scores were then obtained for each suspected speaker model and evidential recording combination in the same manner as Section 6.2.1.2. For each LR estimation, the process is repeated with the suspected speaker model's closest-in-age imposters, resulting in a set

of corresponding imposter LR scores.

- A schematic of the full experimental design is given in Figure 6.3.

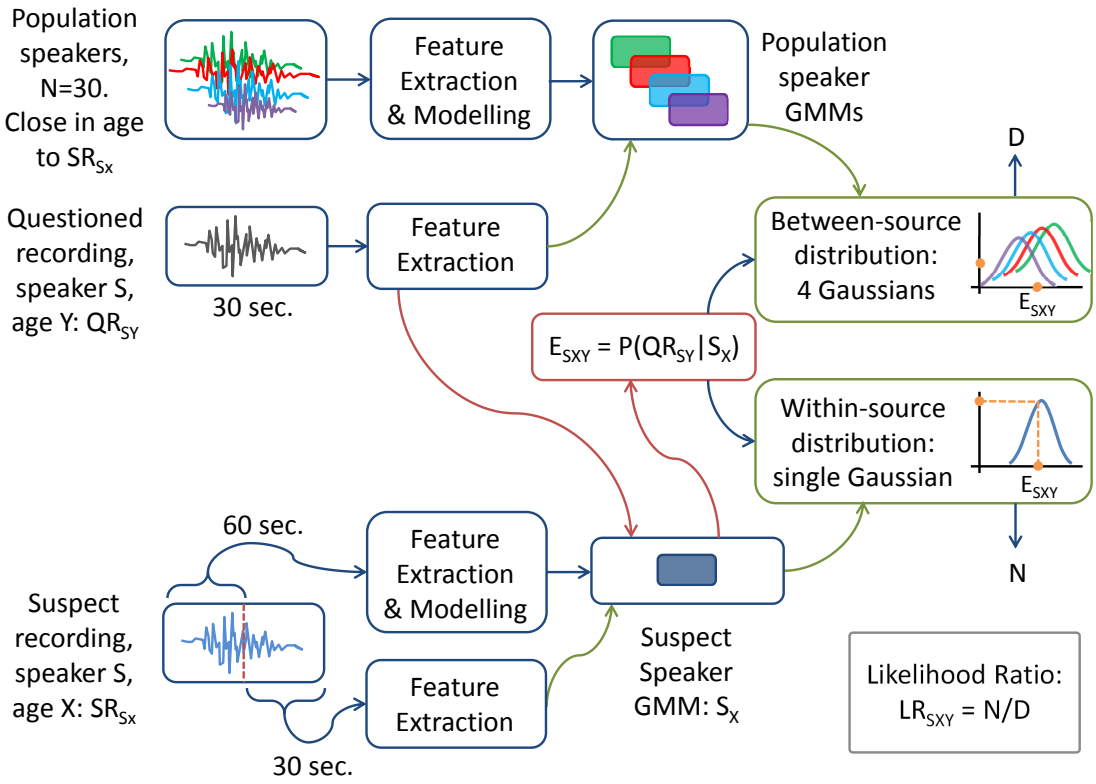

Figure 6.3: A schematic of the experiment described in Section 6.2.2.1. The diagram illustrates the estimation of an LR score for a given suspected speaker and evidential recording. The colour of the arrows correspond to the processing chain for each input. The process is repeated for each of the 5 Irish males, treating each of their recordings in turn as suspected speaker recordings and the remainder of their recordings in each case as evidential recordings.

6.2.2.2 Experimental results

An example of an LR estimation for the speaker Bowman, with a suspected speaker model at age 62 and four evidential recordings at increasing age difference (in a backwards direction), is shown in Figure 6.4. H0 and H1 denote the same-speaker and different-speaker hypotheses, which are modelled by the WS and BS distributions respectively. As the age difference between the suspected speaker recording and the evidential recording increases, the evidential recording log-likelihood ratio decreases. As a result, the numerator of the LR, determined from the WS distribution, decreases, and the denominator of the LR, determined from the BS distribution,

increases. Hence the LR decreases with age difference. At an age difference of 32 years, the $\log_{10}$ LR is negative (-3.35), and thus erroneously supports the different-speaker hypothesis (H1). The imposter LR, while always correctly supporting the different-speaker hypothesis (its $\log_{10}$ value is always negative), eventually exceeds the same-speaker LR. This is likely due to the closeness in age of the imposter to the evidential recording compared with the large age mismatch between the suspected speaker and evidential recording.

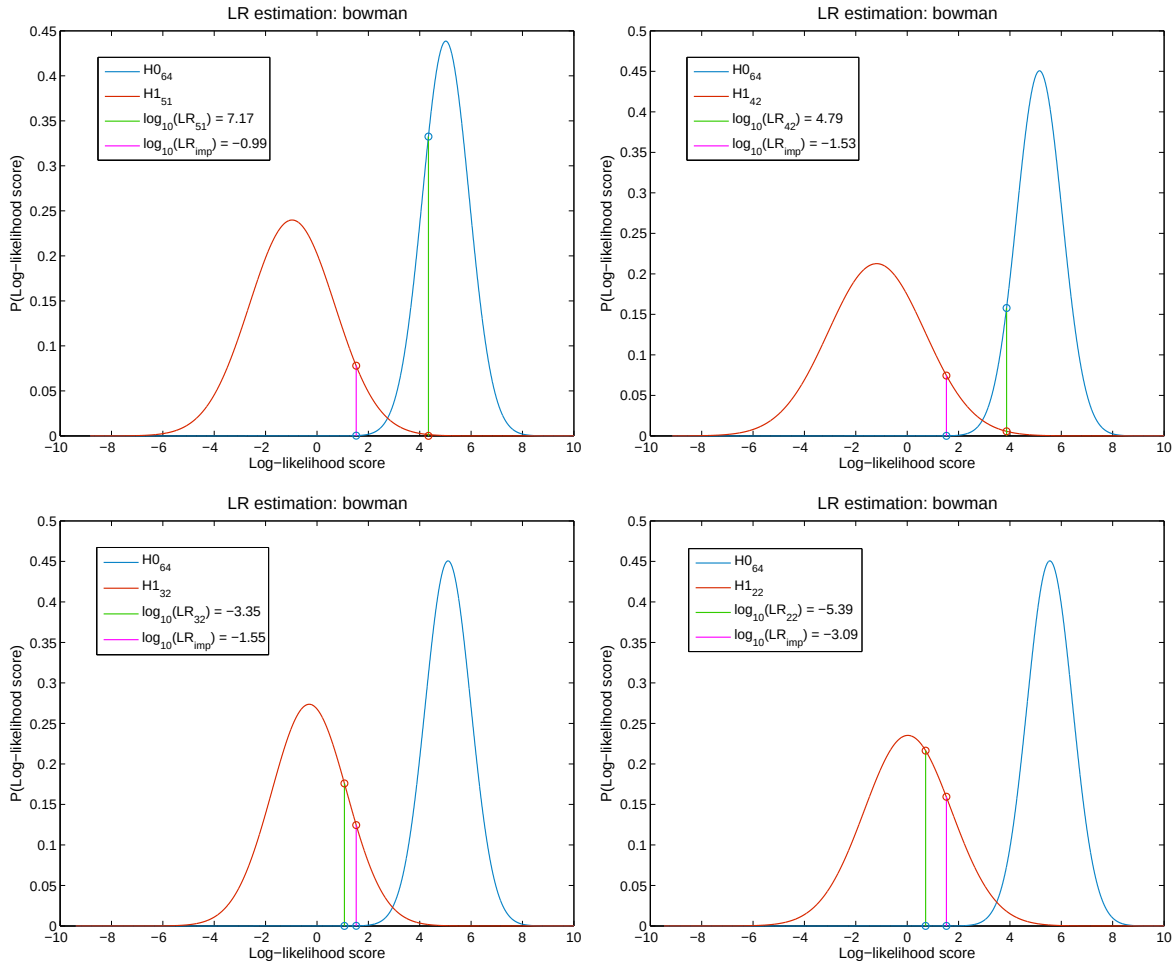

Figure 6.4: Example of ageing LR estimation at increasing time intervals. Speaker: Bowman, backwards direction. H0_x denotes the within-source distribution estimated with a recording of the suspected speaker at age x. H1_y is modelled with the BS distribution estimated from the age y evidential recording and the suspected speaker model's set of age-dependent reference population speakers. LR_y is the likelihood ratio of the age y evidential recording given the H1 and H0 distributions. LR_{imp} is the likelihood ratio of a sample close-in-age imposter recording given the H1 and H0 distributions.

In Figure 6.5, the full set of LR estimates for each speaker are displayed. Each square in each plot indicates a separate LR estimation, with the age of the suspected speaker model on the y-axis and the age of the evidential recording on the x-axis. The squares are colour-coded:

blue-green shades indicate correct support for the same-speaker hypothesis, i.e. ($\log_{10} > 0$), and grey shades indicate erroneous support for the different-speaker hypothesis, i.e. ($\log_{10} < 0$). No same-year comparisons were carried out, and thus the main diagonal is blank. All LR scores to the right of the main diagonal are in a forwards recognition direction, and all to the left are in a backwards recognition direction.

6.2.2.3 Discussion

As the age difference between suspected speaker and evidential recordings increase, there is a general tendency for LRs to decrease. At an age difference of approximately 10 years, represented by the LRs adjacent to the main diagonal, the majority of LRs are positive, correctly supporting the same-speaker hypothesis. There are a number of exceptions however, particularly with the speakers Gogan and Magee. At increasing age differences, i.e. moving outward from the main diagonal, the majority of LR scores decrease progressively. In nearly all cases, erroneous (negative) LR values are produced at an age difference of between 10 and 30 years. Considering that the score in question is the log_{10} of the LR, these decreases are significant.

For example, taking bowman with age 51 suspected speaker recording. Given an evidential recording age 42 the log_{10} LR is 7.78, corresponding to 'very strong' evidence' on the verbal LR scale (Table 2.1). With an age 32 evidential recording the LR is reduced to 2.73, corresponding to 'moderately strong evidence' in support of the same-speaker hypothesis. By age 22, the LR is -2.01. Being negative, the LR represents 'moderately strong evidence' *against* the same-speaker hypothesis. This example demonstrates the significant effect of ageing on the evidential value.

Exceptions to the trend of decreasing LR with age difference are Magee, suspected speaker recording ages 42, 63 and 72 and Byrne, suspected speaker recording ages 61 and 75. It is probable that a degree of mismatch with the reference population, along with stability of certain voice features result in this trend. As mentioned previously, accent mismatch between reference population and the evidential recording has been shown to inflate LR estimates [97]. Although Irish-accented, there are likely other features of the speech of Magee and Byrne that set them apart from the reference population. Referring back to the acoustic measurements of Byrne in Chapter 3, measurements of standard deviation of F0, local jitter and shimmer and articulation rate are remarkably stable relative to other speakers over the course of his recordings. This likely contributes to the stability of his MFCC feature distribution. The same measurements are not as stable for Magee, and therefore it is likely a stylistic 'long-term' voice feature that sets him apart from the reference population and preserves the stability of his feature distribution. In the initial FASR evaluation, Figure 6.2, Magee is the exception to the decreasing trend and Byrne's scores, particularly in a forwards direction, do not strongly decrease.

There is also a general pattern of lower LRs for younger suspected speaker recordings, i.e. the top 2-3 rows in each plot, than for older reference recordings, i.e. the bottom 2-3 rows. The cases where there is a low LR across all age differences for a given suspected speaker

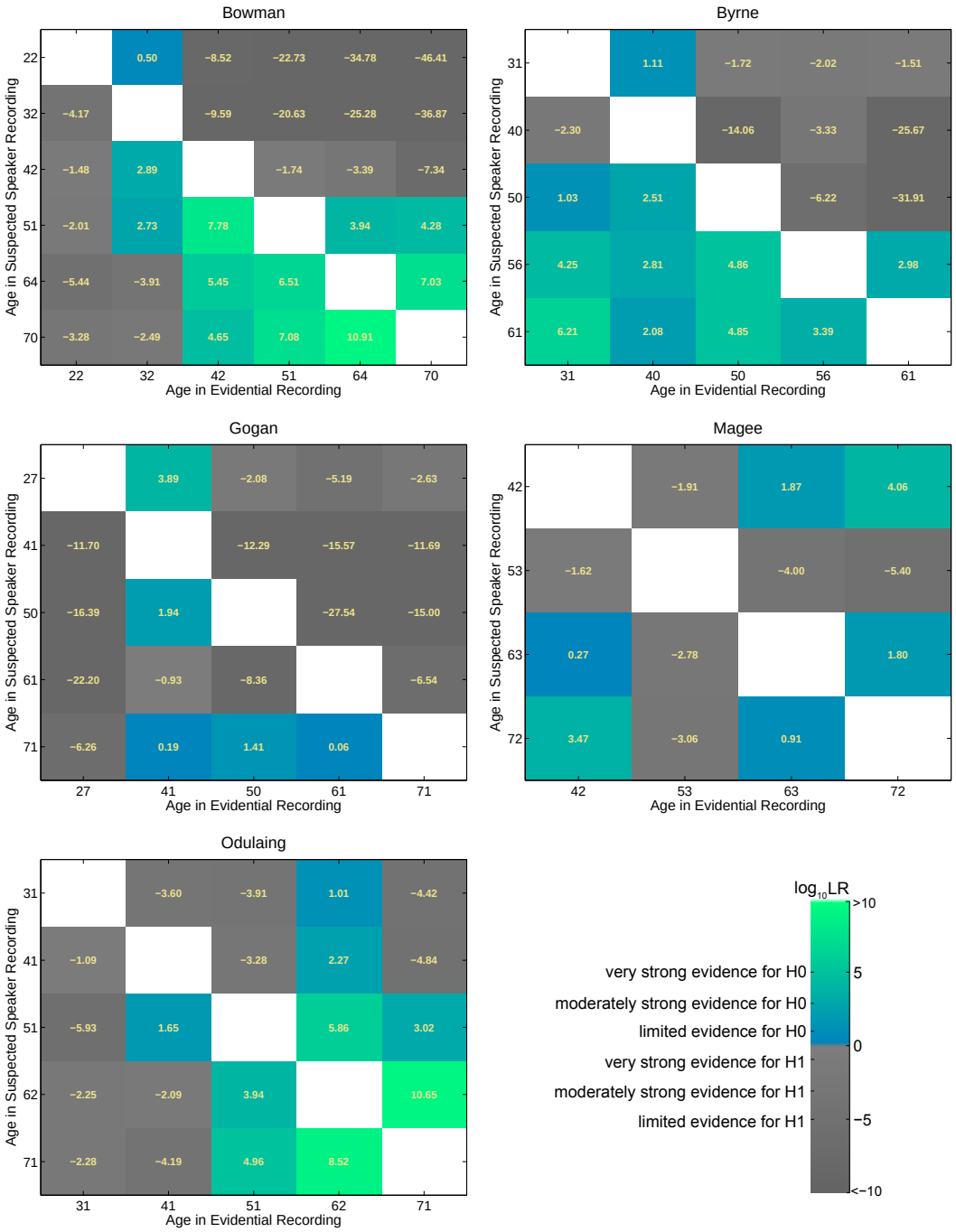

Figure 6.5: LRs for each Irish male speaker for every combination of suspected speaker model and evidential recording ages (apart from same-age comparisions). The y-axis is the speaker age in the suspected speaker recording, i.e. the training age. The x-axis is the speaker age in the evidential recording, i.e. the testing age. Each square represents a separate suspected speaker and evidential recording comparison, and contains the resulting ($log10$) LR. Errors (negative LRs) are coloured from light grey to dark grey with increasing magnitude. Correct (positive) LRs are coloured from blue to green with increasing magnitude. LR scores to the right of the white diagonal represent forwards recognition, and those to its left represent backwards recognition.

recording are as a result of a difference between the training and testing feature distributions. The issue of recording quality is minimised in this experiment, due to the exclusive use of studio recordings from the same broadcaster. Therefore it may be a mismatch in the speaking style and linguistic content of a speaker's young and old recordings that gives rise to these low LRs. It can be observed by listening to the speaker's youngest recordings, those from the beginning of their broadcasting careers, that their speech is more constrained and formal. In their older recordings, with their professional roles established, their speaking style sounds more relaxed and personal. A cultural shift towards a less formal style in radio broadcasting probably contributed to this.

The results in Figure 6.5 are consistent with the only comparable study, by Rhodes [165], in which six male Up-series speakers are used as subjects in an LR evaluation designed in the same way as the present experiment. The FASR system used was BATVOX 3.1 [4], and the reference population speakers were drawn from a British-accented database. Rhodes found that at an age difference of 21 years, the LRs for most same-speaker comparisons were significantly degraded. There was a similar inter-speaker variability to that observed here; in some cases, the onset of age-related LR degradation occurs after 7-14 years. In the case of two speakers with relatively stable ageing LRs, Rhodes suggests that their voice characteristics are responsible for inflated LRs; the two speakers concerned (Bruce and Nick, present in the TCDSA database), have the longest and shortest estimated vocal tract lengths respectively, and are also have strong accent-related features; Tony has a strong Cockney accent and Nick has a Yorkshire accent with American influence. This may render them atypical of the reference population and increase their LRs. A similar source of atypicality may be behind the relative stability of Byrne and Magee recordings in this experiment.

The results here are extend those of Rhodes, in that an accent-tailored UBM and Z-norm speaker set, and an age-tailored reference population is used. Additionally, the suspected speaker models are tested with imposters. Results of imposter testing are presented in Section 6.2.2.4, by means of Tippet plots.

The results demonstrate that age has a significant effect on LR estimation. As would be expected, the effect is varied between speakers. For the majority, errors in LR estimation are introduced after 10-30 years age difference. In some cases, there is a sharp transition from a high positive LR to a low negative LR over a 10 year interval. Since the recording conditions are constrained, variability can be attributes mainly to the speakers themselves. The relative stability or instability of a particular speaker's acoustic features, and a mismatch between speaker populations, in terms of linguistic or stylistic content, are responsible for the observed variability of results.

Referring back to the 'Yorkshire Ripper' Hoaxer investigation [63], it now seems clear that the categorical statements made there regarding forensic speaker identification being unaffected by ageing do not hold. Similarly, the statements made by Hollien and Schwartz [92, 94] that "noncontemporary speech samples can be used effectively in nearly all types of speaker identification" and the conclusion by Künzel [122] are at odds with the findings in this Chapter.

The aforementioned studies demonstrate that in a forensic comparison case given the same texts [63,94,122], in controlled conditions [94,122], with analysis by phoneticians [94] and forensic experts [63], *subjective* conclusions can be reached that are unaffected by ageing.

However, in the realistic scenario where non-matched texts are to be compared in uncontrolled conditions with an automatic system, the conclusions here are in line with Rhodes [165], in that an ageing significantly affects LR estimation, leading to errors after 10-30 years.

6.2.2.4 Tippet plots

Tippet plots are a standard means of observing the performance of FASR systems, in a similar manner to DET plots in speaker verification (Section 2.5.6). For a given system, a tippet plot indicates the probability of a particular LR for both same-speaker and different-speaker comparisons. Thus, the discrimination of the system between same-speaker and different-speaker comparisons can be observed.

Tippet curves for the experiment in this Section are shown for increasing age differences in Figure 6.6. The LR scores from the same-speaker comparisons in Figure 6.5 were used to create the 'target' curves. The imposter LR scores from the close-in-age different-speaker comparisons were used to generate the 'imposter' curves.

As per the discussion relating to Figure 6.5, as the age difference increases from 10 years to 30 years, the target curves move in a negative direction, resulting in a higher probability that the LR values are below 10^0. At an age difference of 30 years, the target and imposter curves overlap, indicating that there is a loss of discrimination between same and different-speaker comparisons.

At an age difference of 10 years, the probability of a target score with a LR above 10^0 is approximately 55%. In other words, there is a 45% chance of an erroneous LR. The high probability of error after 10 years underlines the significance of ageing on LR estimation. At 20 years, the chance of error rises to approximately 55%, and at 30 years it is approximately 65%.

A calibration (Section 2.5) of the system could be applied to shift the LRs to a desirable operating range. For example, considering an age difference of 10 years in Figure 6.6, a simple linear shift of the LR scores by $10^{2.5}$, would result in a probability of approximately 70% of a same-speaker LR being greater than 10^0 and a probability of approximately 15% of a different speaker LR being greater than 10^0.

6.2.3 Eigenageing compensation for ageing LR estimation

In Section 6.2.2, the significant effect of ageing on LR estimation was presented. In this Section, the application of eigenageing compensation, introduced in Chapter 5, to LR estimation is presented.

Eigenageing compensation is more suitable in the forensic scenario than the alternative stacked classifier approach proposed in Chapter 4. The primary problem with using a stacked

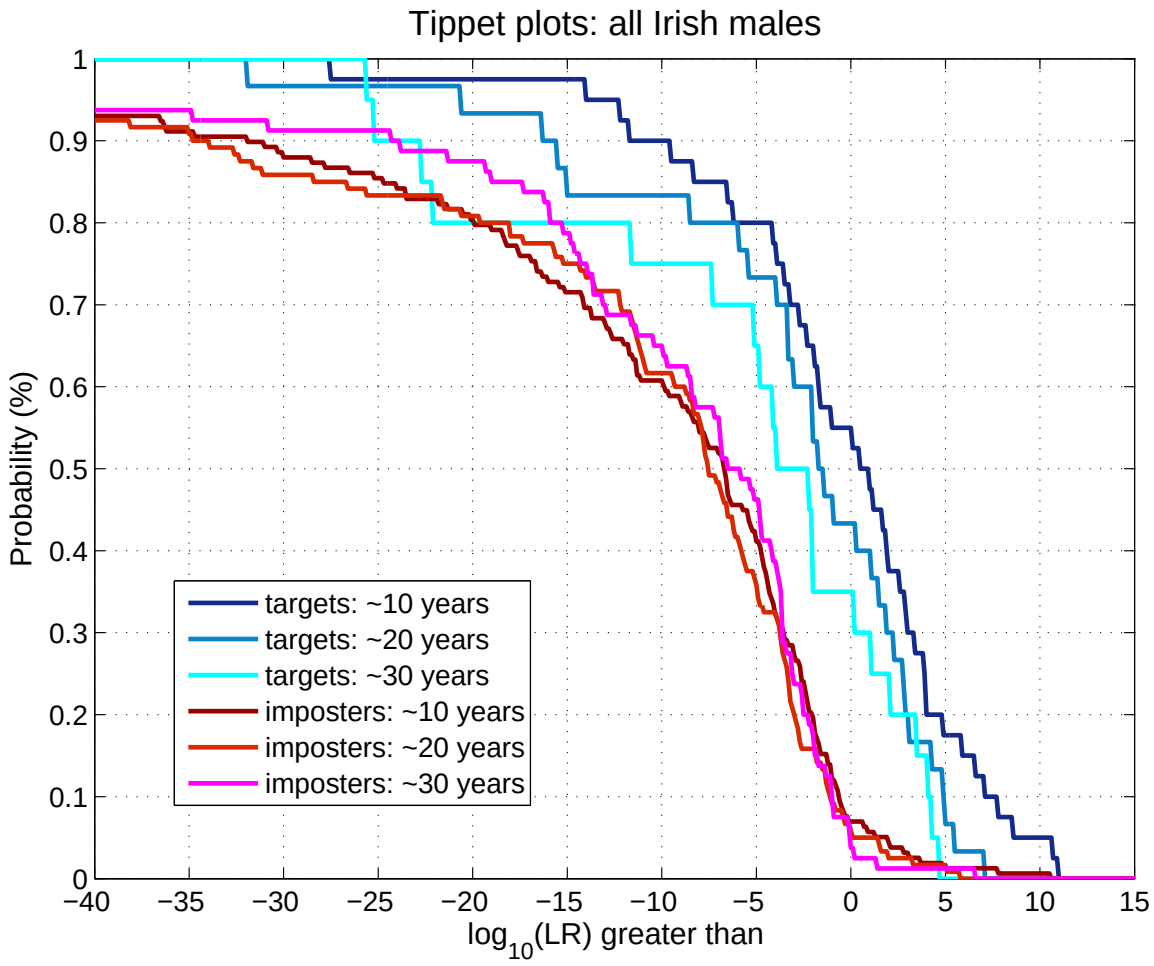

Figure 6.6: Tippet plots for age differences of approximately 10, 20 and 30 years for the pooled LRs of the Irish males and their close-in-age imposters.

classifier approach in the forensic scenario is that it relies on knowledge of the test speaker age. It cannot be assumed that the age of a speaker in an evidential recording is known however. Automatic age estimation, discussed in Section 5.3, has not reached the level where an estimate of age could be provided for the system to a desirable level of accuracy. Secondly, eigenageing compensation has greater scope for dealing with variability in the ageing process, observed in individual speaker score trends throughout this thesis, than the stacked classifier approach. The fact that eigenageing operates on a model (GMM supervector) level, and the stacked classifier operates on a score (log-likelihood ratio) level, mean that there are far greater number of dimensions available to eigenageing to project away ageing effects than is possible for the stacked classifier.

6.2.3.1 Experimental design

The previous experiment with Irish-accented males was used to evaluate the effect of eigenageing compensation in LR estimation. Preserving all parameters of the experiment in Section 6.2.2.1, eigenageing compensation was applied to the output of the GMM-UBM system given an evidential recording, as detailed in Section 5.1.3.

For each suspected speaker, an ageing subspace was created from the recordings of all other TCDSA males, as detailed in Section 5.1.2. This subspace was suitable for the forwards comparison direction. However, since a backwards comparison direction was also considered in this experiment, an additional backwards subspace was created.

This was done in the same manner as the forwards subspace, with a difference at the supervector subtraction stage. In the forwards direction the youngest supervector was subtracted from the set of all other supervectors. In the backwards direction the *oldest* supervector was subtracted from the set of all other supervectors. The second and third oldest supervectors were then subtracted, completing the three-stage difference operation.

In comparisons where the speaker model was younger than the evidential recording, the forwards ageing subspace was used. Where the speaker model was older than the evidential recording, the backwards ageing subspace was used.

6.2.3.2 Experimental results

In Figure 6.7, the LR estimates after applying eigenageing compensation are presented. This result set can be directly compared to Figure 6.5, which shows the LRs for the same comparisons before eigenageing compensation.

There is a general increase in the LR values across all comparisons after applying eigenageing compensation. Many originally positive LR values are increased by orders of magnitude after eigenageing compensation. A number of negative LR values become positive after eigenageing compensation. Most highly negative LR values are increased, although they still remain negative.

6.2.3.3 Discussion

Taking the application of eigenageing compensation to Byrne, suspected speaker recording age 31, as an example: the LR for an evidential recording at age 40 increases from 1.11 to 2.01, representing an increase in the verbal strength of evidence from 'moderate' to 'moderately strong' support for the same-speaker hypothesis. The LR for an evidential recording at age 40 increases from -1.72 to 0.56, moving the balance of the evidence from 'moderate' support against the same-speaker hypothesis to 'limited' support for the same-speaker hypothesis. The LR for an evidential recording at age 61 increases from -2.02 to -0.66. While the LR remains (erroneously) negative, the support against the same-speaker hypothesis is reduced from 'moderately strong' to 'limited'.

The effectiveness of eigenageing compensation on LR estimation cannot be based solely on

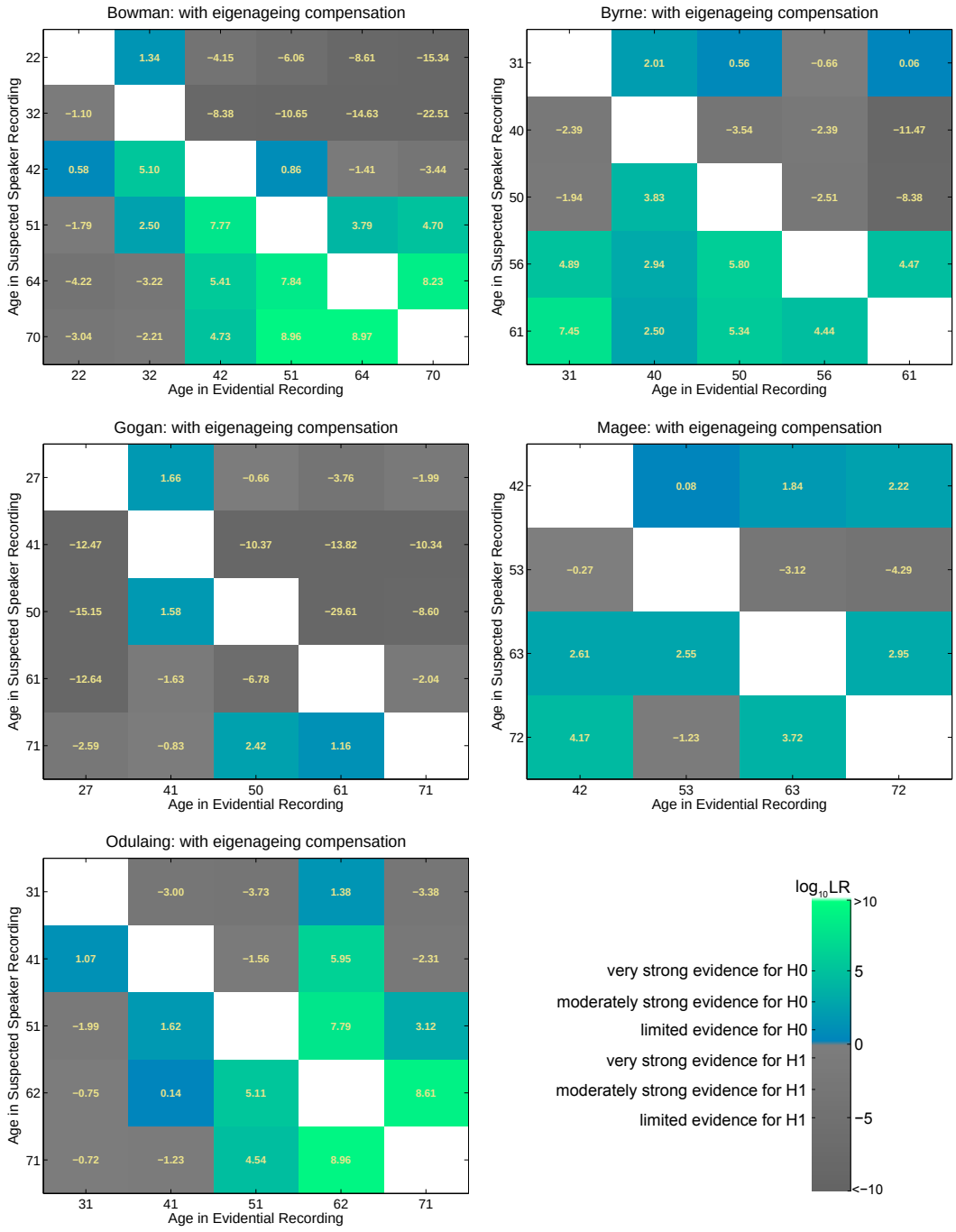

Figure 6.7: LRs for each comparison considered in Figure 6.5, after the application of eigenageing compensation

same-speaker LRs. If the LRs of different-speaker comparisons are increased in a similar manner to the same-speaker LRs then there is no advantage to the technique. In fact, it will probably be a disadvantage, as it may result in different-speaker comparisons resulting in positive LRs. To observe the same and different speaker LRs before and after compensation, a set of Tippet curves for each condition, at different age intervals, are plotted in Figures 6.8, 6.9 and 6.10.

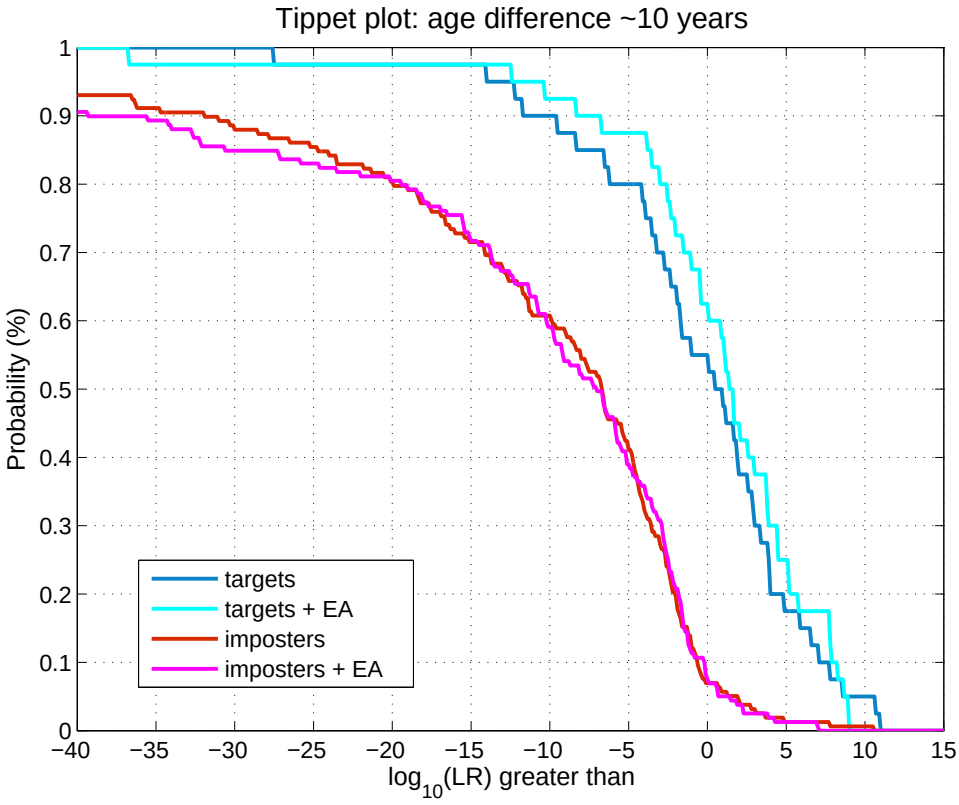

Figure 6.8: Tippet curves for an age difference of approximately 10 years, before and after the application of eigenageing compensation

In all cases, the application of eigenageing compensation has a greater effect on the target (same-speaker) LRs than the imposter (different-speaker) LRs. Following eigenageing compensation: at an age difference of 10 years, the probability of a negative LR is decreased from approximately 45% to 40%. At 20 years, the probability of a negative LR is decreased from approximately 55% to 50%. At 30 years the probability of a negative LR is slightly increased, however the probability of a $\log_{10}$ LR less than -1 is decreased from approximately 65% to 50%. As discussed in Section 6.2.2.2, a calibration could be applied to shift the score to a different operating range.

At all three age differences, the probability of a positive *imposter* LR is between 5-10%, and is not increased with the application of eigenageing compensation.

Thus, based on this experiment, eigenageing compensation increases the strength of evidence of same-speaker comparisons where there is an age difference between reference and evidential recordings, and importantly, does not compromise the system by erroneously increasing the strength of evidence of different-speaker comparisons.

In this experiment, all TCDSA males were used for subspace estimation. Ideally, a set of speakers matched to the suspected speaker would be used for subspace estimation, in the same

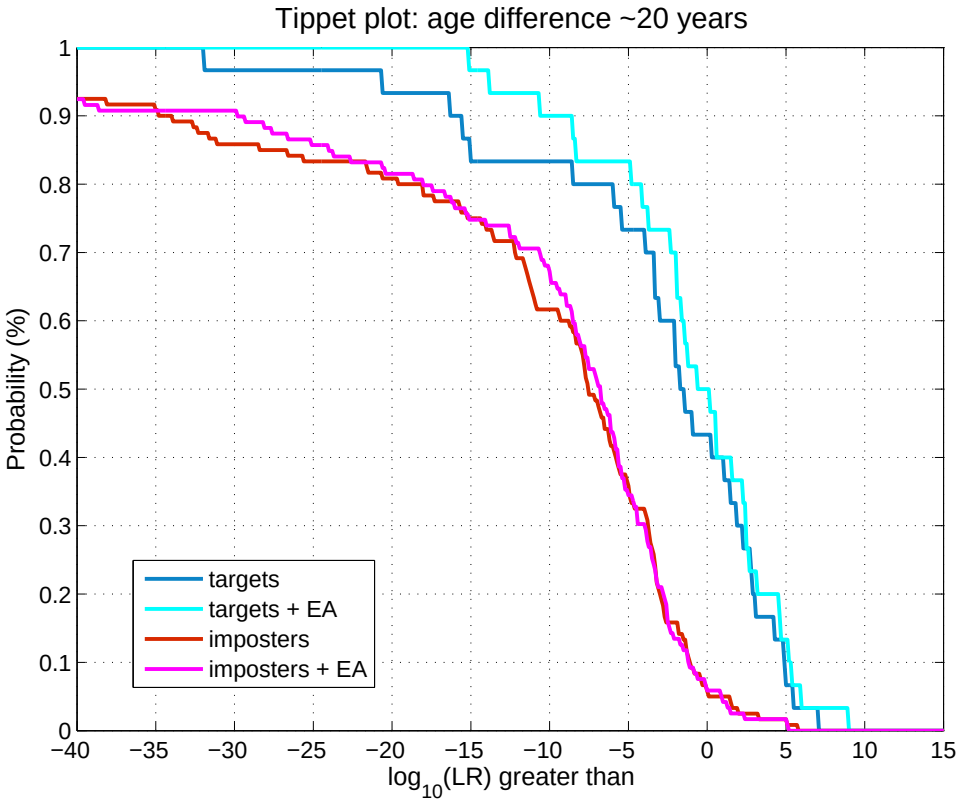

Figure 6.9: Tippet curves for an age difference of approximately 20 years, before and after the application of eigenageing compensation

was as a reference population is selected. Additionally, the subspace could be tailored to a particular ageing pattern by incorporating prior knowledge about the speaker, e.g. if they were a smoker, or were geographically mobile, then this could be reflected in the choice of subspace speakers.

6.3 Auditory detectability of ageing

In the previous Sections of this Chapter, the effect of vocal ageing on forensic automatic speaker recognition has been the focus. In this Section, an investigation into the auditory detectability of vocal ageing is presented.

The purpose of this investigation is to assess the extent to which naïve listeners can detect ageing change in recordings of the same speaker at different ages. These human decisions can then be compared to the output of the automatic system given the same samples, offering insight into the effect of vocal ageing on the recognition on individual speakers.

This is of relevance to aspects of forensics that rely on subjective listening, including speaker profiling, selection of features for acoustic-phonetic analysis, and auditory-perceptual analysis

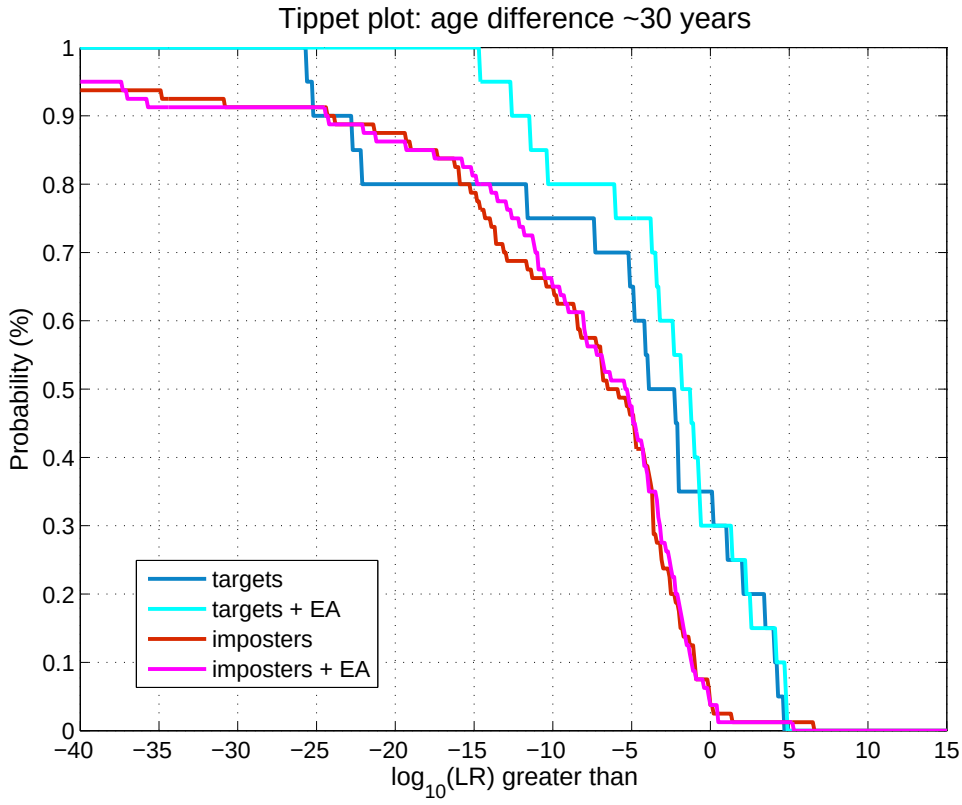

Figure 6.10: Tippet curves for an age difference of approximately 30 years, before and after the application of eigenageing compensation

- any of which may be adversely affected when dealing with non-contemporary samples. The degree to which listeners can, or can not, detect ageing-related changes in the voice is an indicator of the degree to which ageing will affect their judgement.

By comparing the perceptual cues in a speaker's voice that listeners use to make a judgement on age with the acoustic measurements from that speaker's speech, an objective understanding of the acoustic attributes that make a speaker sound older may be gained.

In addition, identifying the most important perceptual ageing cues may aid realistic automatic synthesis of ageing in the voice or potentially point towards prosodic/long-term features that are ageing-robust.

6.3.1 Listening experiment design

A listening experiment was designed to assess the detectability of ageing in the voice at different age intervals. A two-alternative forced-choice [134] testing format was adopted, whereby a question consisted of a pair of speech samples from the same speaker at different ages, with the listener indicating the sample in which they thought the speaker was older. The test was similar in its aim and its design to that of Künzel [122]. The age-differences between samples

were increased from the 11 year span considered in Künzel's study to approximately 30 years.

From the TCDSA database, 10 male and 10 female British and Irish accented speakers from the BBC, RTÉ and Up Series were chosen as test subjects. Based on their distribution of ages (Figure 3.2), a set of age ranges [1,2,3,4] were defined as ages [28-39, 40-45, 46-54, 55-64]. A sample was extracted for each test speaker from each age range, resulting in an average age-difference of 9.7 years between each sample for each speaker.

For each of the 20 speakers, a set of six comparison tests were created by pairing samples from their different age ranges in the following way: [1-2, 1-3, 1-4, 4-1, 4-2, 4-3], where, for example, 1-2 denotes a sample from age range 1 followed by a sample from age range 2 and 4-3 denotes a sample from age range 4 followed by a sample from age range 3. Thus, for each speaker a set of comparisons at approximately 10, 20 and 30 year age differences, presented in both forward and backwards directions, was considered. For each occurrence of an age range in the set of comparison tests, a different 5 second sample was used. This ensured that no listener heard the same sample twice. Thus, there were 12 different 5 second samples used to compile the six comparison tests. Each pair of samples were separated by a 0.5 second beep, with 0.25 seconds of silence either side.

A simple web application was designed to administer the listener test (accessible at [1]). Of a total of 120 possible comparison tests (six for each of 20 speakers), a listener was presented with 48 tests in a random order. Each comparison test was comprised of an audio player object and a two-part question. The audio player object allowed multiple plays of each recording, and listeners were instructed that they could listen to a comparison more than once if necessary. The number of recording plays for a given question was stored. An example on a question is illustrated in Figure 6.11.

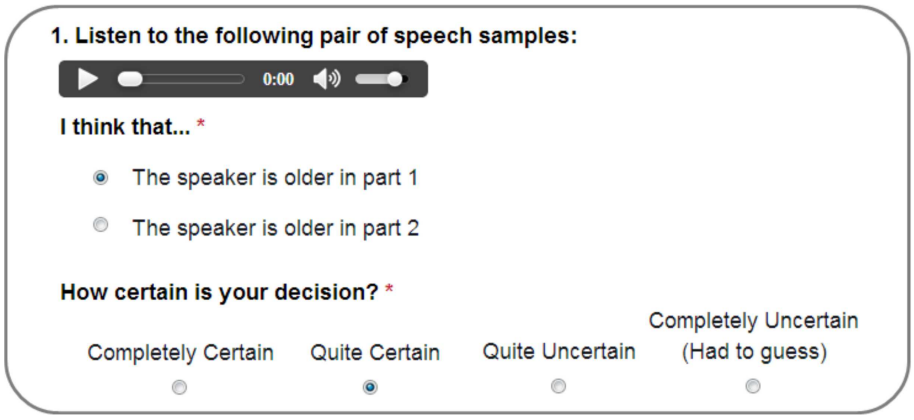

Figure 6.11: A screen capture of an example question from the listening experiment.

In each question, listeners were asked whether the speaker was older in part 1 or part 2 of the comparison test. They were also asked to give the certainty of their decision on a 4-point scale: "completely certain", "quite certain", "quite uncertain", "completely uncertain (had to guess)".

Having responded to the 48 comparison tests, listeners were asked what they found to be the most important feature of a voice when making an age comparison, and whether they found the task easier for male or female voices.

In total, 36 listeners (25 males, 11 females) completed the experiment [1]. 27 listeners were native English speakers, 9 listeners were non-native (fluent) English speakers. The mean age of the respondents was 30.7 years. The youngest and oldest respondents ages 17 and 55 respectively, and the majority (23) were aged between 20 and 40. The majority of listeners were University students or academics. None had a history of hearing or speech impairment.

For the test itself, 30 listeners used headphones and 6 used their computer's speakers. The average test duration was 19.4 minutes.

6.3.2 Listening experiment results

The set of responses from all listeners were analysed to determine the overall detectability of ageing at different age intervals. Using a signal detection theory approach [122, 134], results are expressed graphically with receiver operating characteristic (ROC) curves for each of the age differences. Treating the younger-older comparison tests (e.g. comparison tests [1-2, 1-3, 1-4], described in Section 6.3.1) as forwards examples and older-younger comparison tests (e.g. [4-1, 4-2, 4-3]) as backwards examples, the hit rate was calculated as the proportion of forwards examples answered correctly and the false alarm rate as the proportion of backwards examples answered incorrectly. By considering the listener responses on the 4-point certainty scale, a set of hit rates and corresponding false alarm rates were determined for a range of certainty values. The resulting ROC curves for approximate age differences of 10, 20, 30 and 10-30 years are plotted in Figure 6.12.

The discriminability index d' [134] can be considered as the difference between the 'noise-only' and 'noise+signal' conditions, the 'signal' is the effect of ageing in this scenario. d' is given by:

$$d' = z\left(HR\right) - z\left(FA\right) \tag{6.1}$$

where HR is the hit rate, FA is the false acceptance rate and $z\left(\cdot\right)$ is the inverse cumulative normal distribution. d' values for each age difference are included in Figure 6.12. The higher the d', the higher the sensitivity of the listeners to the presence of ageing. The discrimination performance expected by chance is the main diagonal in Figure 6.12. Also included in Figure 6.12 is the percentage of correct decisions, a (this value corresponds to the area under the ROC curve) and n, the number of responses used to generate the d' and a statistics.

It is clear that the effect of ageing is detected with increasing precision as the age difference increases; the discriminability index d' increases from 0.74 at a mean age difference of 10 years to 2.14 at 30 years. The increase in the corresponding percentage of correct decisions is from 64% to 86%. Based on expectations from the physiological changes (Section 3.3) and associated

[1] 5 of these listeners partially completed the experiment; this will be discussed where relevant.

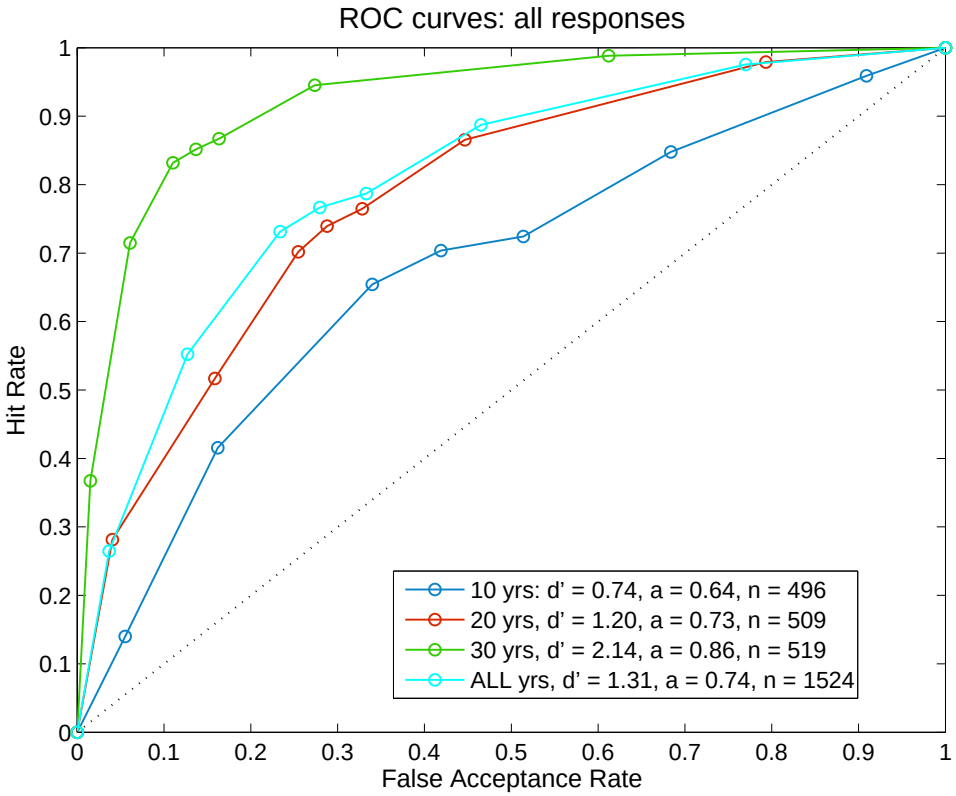

Figure 6.12: ROC curves obtained from the responses of all listeners, given all speakers, at mean age differences of 10, 20 and 30 years, and for the pooled set of all responses ('ALL'). d' is the discriminability index, a is the percentange of correct decisions, n is the number of responses used to calculate the d' and a statistics.

acoustic changes (Section 3.4), this increase is expected. The comparable study by Künzel [122] tested detectability at an 11 year age difference only, and reported a d' value of 0.26 and a 57.4% correct decision rate over this period.

A breakdown of correct listener decisions for male and female speakers at increasing age differences in forwards (younger-older comparison) and backwards (older-younger comparison) directions is provided in Table 6.1. In addition to the trend observed already observed in Figure 6.12, that ageing detectability increases with age interval, two other observations can be made.

Firstly, there is a significant difference between listener performance given male and female speakers. At all age differences and directions, there is a greater percentage of correct decisions given female speakers than male speakers. The combined percentage correct for females is 8-14% greater than males at all age intervals. This gender-dependent finding goes towards explaining the increased detectability in this experiment compared with that of Künzel [122], who used only male speakers in his experiments. The male percentage correct value for both directions combined is 61.9%, which is comparable to the 57.4% in Künzel's paper. Other potential reasons for the increased detectability observed here are discussed in Section 6.3.3.

Secondly, there is generally a greater percentage of correct decisions in the forwards direction compared with the backwards direction, particularly for female speakers and at smaller age differences. The same effect of 'ageing direction' on performance was observed in Künzel's experiment [122]. This finding is discussed further in Section 6.3.3.

		Correct Decisions (%)		
	age difference	forwards	backwards	combined
males	10 years	65.9	58.3	61.9
	20 years	68.9	70.9	70.0
	30 years	80.5	82.5	81.5
females	10 years	81.4	57.4	70.2
	20 years	85.1	72.0	78.2
	30 years	96.1	94.1	95.1
all	10 years	70.4	58.1	64.1
	20 years	74.0	71.2	72.5
	30 years	85.2	86.3	85.7

Table 6.1: Percentage of correct decisions for male and female speakers, for different mean age differences, for forwards (younger-older) and backwards (older-younger) comparisons

6.3.2.1 Effect of speaker gender

To further analyse the difference in detectability for male and female speakers, ROC curves for each gender at increasing age differences are provided in Figure 6.13. The increased performance given female speakers holds at all age differences. The difference is particularly noticeable at a 20 year age difference, where there is a doubling of d' and a 15% increase in correct decisions between females and males.

The ROC curves in Figure 6.13 summarise the performance for the pooled set of scores for all males and females. In Figure 6.14, the correct decision rate for each individual speaker, across all age differences is presented. Speakers are ordered by gender and according to increasing correct decision rate. The 95% confidence intervals around the mean of male and female correct decision rates are superimposed. Although there is some overlap between the male and female scores, the difference between their distributions is statistically significant; $p = 0.0001$ in a two-tailed t-test.

There is clear variability between individual speakers. Aside from the two lowest scoring females, Lawlor and Nibhriain, this variability is more apparent in males compared with females. The lowest scoring speaker, bowman, is 10% lower than the lowest scoring female, Lawlor. The highest scoring female, Finucane, is over 10% higher than the highest scoring male, neill.

The greater detectability of females is also reflected in the number of sample plays needed for

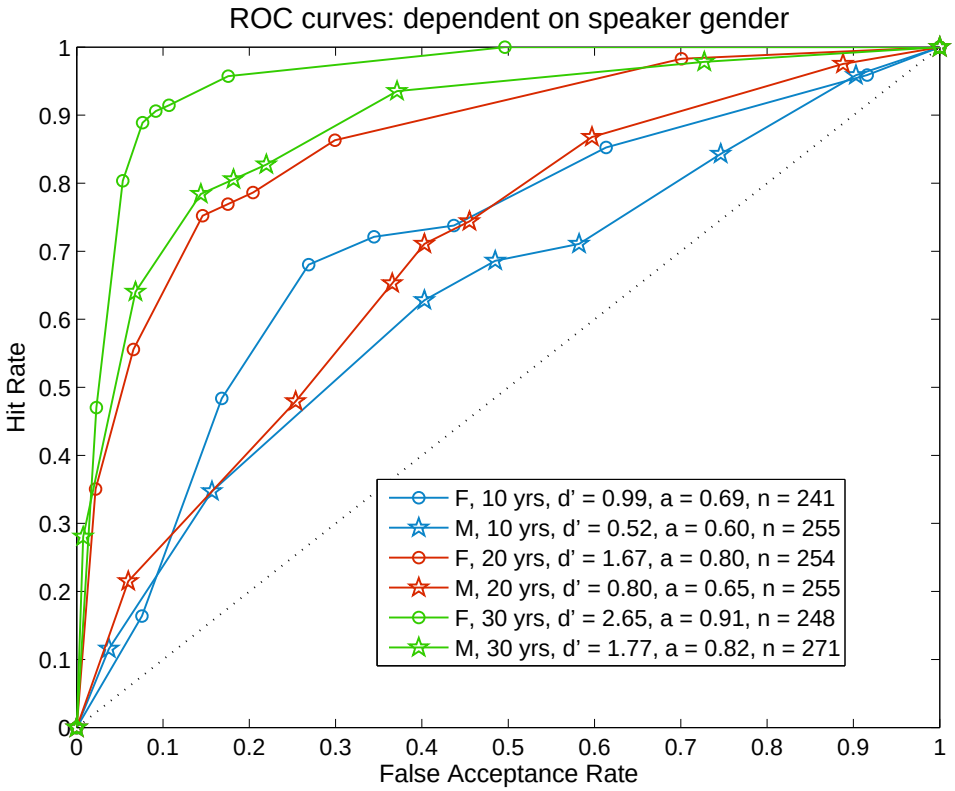

Figure 6.13: ROC curves obtained from the responses of all listeners given male and female speakers at mean age differences of 10, 30 and 30 years. d' is the discriminability index, a is the percentange of correct decisions, n is the number of responses used to calculate the d' and a statistics.

a listener to come to a decision; 18% of male samples were played more than once, corresponding to an average of 1.25 plays per sample. 14% of female samples were played more than once, corresponding to an average of 1.18 plays per sample.

The majority of respondents (23) considered that the task was easier for female speakers. 5 respondents thought the task was easier for males, and the remaining 8 either thought it was equally difficult for both cases or offered no opinion.

As observed in the acoustic analysis of ageing speakers in Chapter 3, males and females displayed different fundamental frequency (F0) trends with ageing; female F0 consistently decreased, while male F0 decreased then increased, and was not very consistent between speakers. Considering that F0 was the most commonly reported feature used by listeners to make a decision (discussed in more detail later in this Section) and that there are differences in both male and female F0 trends and detectability results, there may be a relationship between F0 and detectability.

To observe the correlation between F0 and detectability, the F0 difference for each speaker between their age range 1 and age range 4 recordings is plotted against the percentage correct decisions for the 1-4 and $4-1$ comparisons in Figure 6.15 for males and females. A linear fit

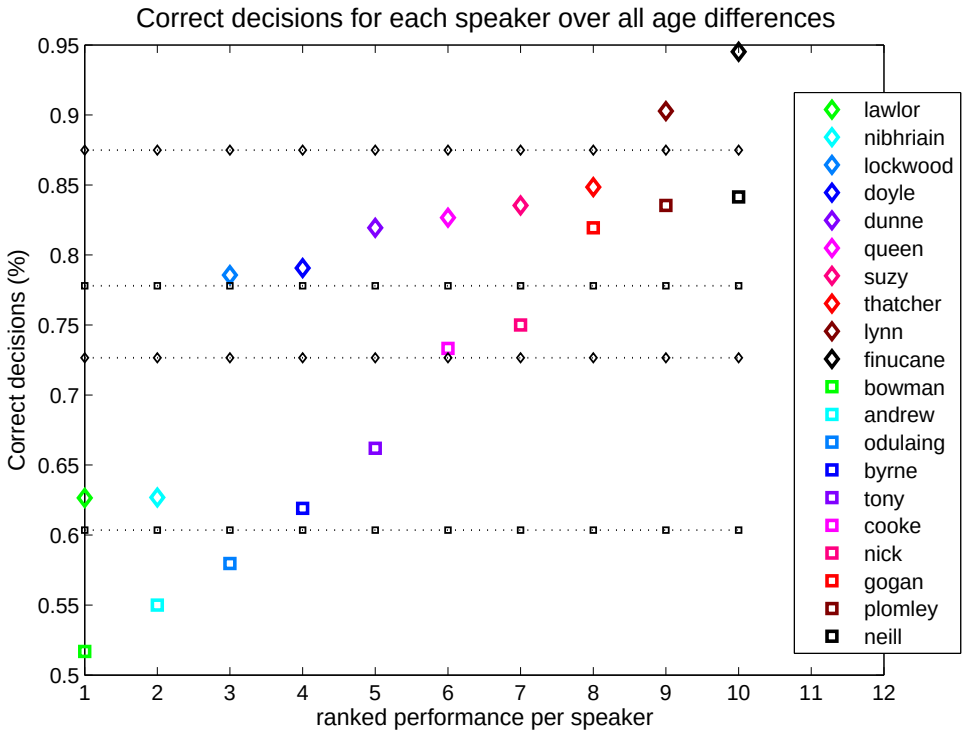

Figure 6.14: The percentage of correct decisions from all listeners given individual speakers. Speakers are sorted by gender and are ranked from lowest to highest correct decision rate. Females are indicated with diamonds and males with squares. The dotted lines (with diamond/square markers corresponding to females/males respectively) denote the 95% confidence intervals around the mean of the male and female score distributions.

and correlation coefficient is detailed on each graph.

For the female case, there is a strong positive relationship between F0 change and the percentage of correct decisions. It is also interesting that the female with the lowest detectability overall, Lawlor, Figure 6.14, has the lowest F0 change in Figure 6.15 and the female with the highest detectability overall, Finucane, Figure 6.14, has the greatest F0 change in Figure 6.15.

There is a much weaker (negative) relationship between correct decisions and F0 in the male case. This suggests that F0 is a weaker cue to ageing in males than in females. Overall, this provides support for the proposition that vocal ageing is more detectable in females than males.

6.3.2.2 Effect of listener gender

Having observed a clear difference in ageing detectability across speaker genders, it is also of interest to analyse the effect of *listener* gender on detectability. In Figure 6.16 ROC curves given male and female listener responses to all speakers at different mean age differences are shown. The performance of female listeners is greater than males at all age differences. The difference in performance between male and female listeners is most apparent at an age difference of 30

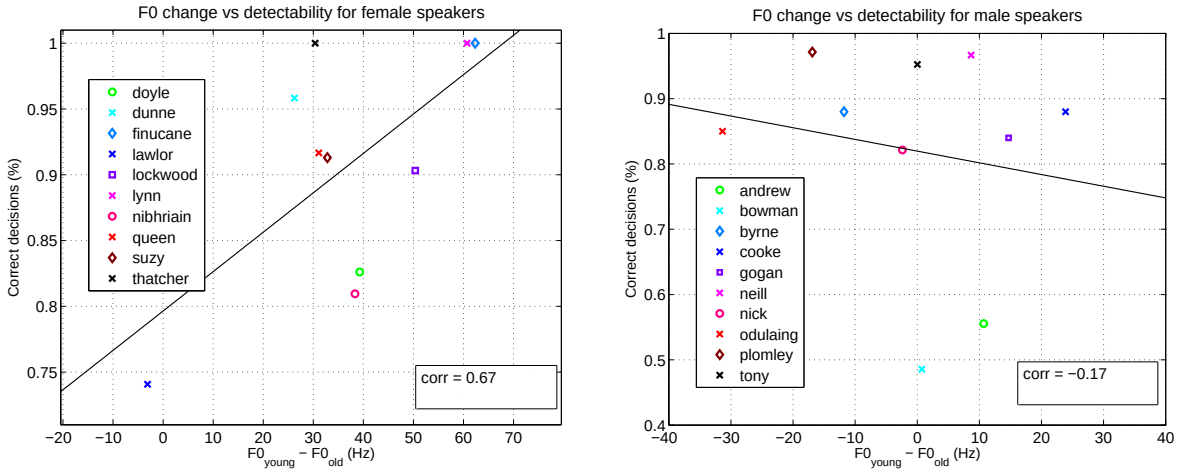

Figure 6.15: The F0 change between young (age range 1) and old (age range 4) samples for each speaker plotted against the percentage of correct decisions from all listeners given that speaker. **Left:** male, **Right:** female

years, where d' is 1.79 for males and 3.33 for females. The corresponding percentage correct rates are 81% and 95%.

In Figure 6.17, the correct decision rate for each individual listener, across all age differences is presented. Listeners are sorted by gender and ranked according to increasing correct decision rate. The five listeners who did not fully complete the test are not included. The 95% confidence intervals around the mean of male and female correct decision rates are superimposed. The difference between male and female score distributions is significant ($p = 0.04$)

There is greater variability in the male listener performance, with both the best and worst performing listeners being male. The sample size is limited, particularly for females. However, from this set of respondents, females are significantly better at detecting ageing overall. The breakdown of performance of male and female listeners given male and female speaker groups was also analysed, and the trends observed here hold: female listeners perform better than male listeners given both male and female speakers, and both female and male listeners perform better given female speakers than male speakers.

6.3.2.3 Other effects

To assess any difference between the performance of native and non-native English speakers, the responses of native and non-native listeners given all speakers at all time differences were compared. Not considering the five listeners that partially completed the test, there were 10 non-native listeners and 21 natives. The average percentage correct decisions was slightly higher for the non-native listeners (77.7%) compared with native listeners (74,2%). This difference was not significant however (p=0.4).

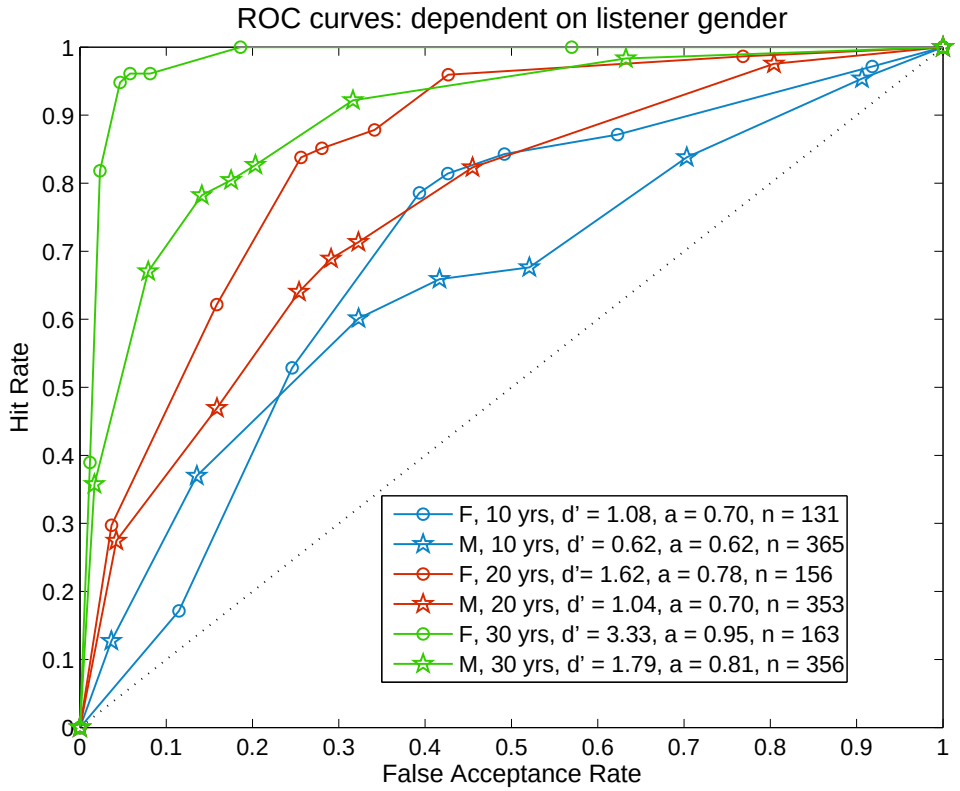

Figure 6.16: ROC curves obtained from the responses of male and female listeners all male and female speakers at mean age differences of 10, 30 and 30 years. d' is the discriminability index, a is the percentange of correct decisions, n is the number of responses used to calculate the d' and a statistics.

Respondents were also asked to provide their opinion on what they considered the most important feature of a voice in making their decision. The responses to this question are summarised below.

- **Fundamental Frequency**: this was the most commonly reported cue, with 21 respondents listing it as an important factor in making their decision. It was referred to as fundamental frequency, pitch or depth by various respondents, with some mentioning specifically that they expected the older voice to contain more 'bass' or 'depth' relative to the younger voice.

- **Voice Quality**: this was the second most reported cue, with 11 listeners mentioning attributes of the voice that fall broadly under 'quality'. Features mentioned as cues to an older voice were: shaky, croaky, creaky, rough, gruff, hoarse and gravelly, and those mentioned as cues to a younger voice were: lilting, bright, light and energetic.

- **Speed**: 7 listeners reported that the speaking speed was a cue to ageing, both in terms of decreasing speed with age and of increased pauses with age.

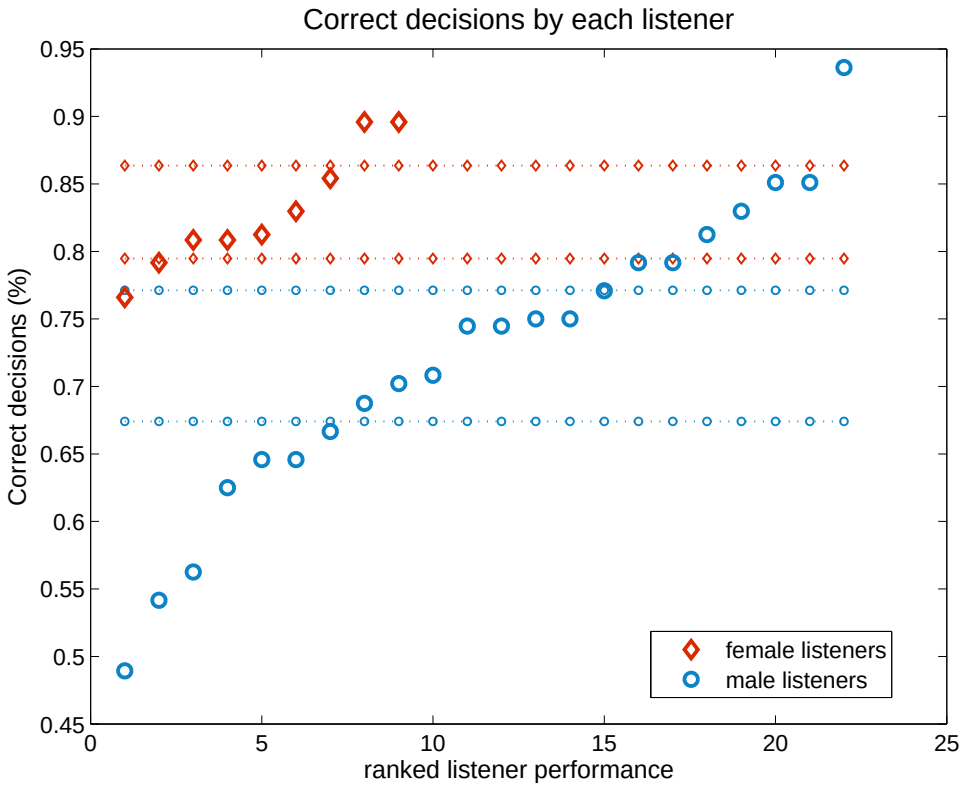

Figure 6.17: The percentage of correct decisions from male and female listeners given all speakers. Listeners are sorted by gender and are ranked from lowest to highest correct decision rate. Females are indicated with diamonds and males with squares. The dotted lines (with diamond/square markers corresponding to females/males respectively) denote the 95% confidence intervals around the mean of the male and female score distributions.

- **Audio Quality**: 6 respondents listed the quality of the audio as a cue. It was mentioned that in the absence of other cues, a decision was made based on which recording sounded older. Other listeners expressed that the audio may have been a misleading cue, and it could have unconsciously influenced their decision making.
- **Familiarity**: 5 respondents mentioned that their familiarity with the voices in the test may have made the their decisions easier.
- **Subject**: 3 respondents mentioned that the subject matter of the samples, in terms of the language, content and 'dictation' was used as a cue to date the era of the recording, and hence make a judgement on ageing.

6.3.3 Listening experiment discussion

The listening experiment confirms that the detectability of ageing in the voice increases with age interval. Considering the subset of the results comparable with a study by Künzel [122],

the detectability of male speakers at a 10 year age difference, the percentage correct rates are similar: 57.4% compared with 61.9% obtained here.

The comments left by the respondents point towards potential reasons for the slightly higher detectability in this experiment; some felt that the acoustic quality, familiarity with the speakers, and the subject matter had an influence on their responses. Listeners were also allowed multiple plays of a sample. In Künzel's study, the acoustic quality was constrained, as the recording environment was consistent within the young and old conditions. The subject matter and language was also constrained, as the stimuli were the same passages of read text in each sample. Only one listening of the samples was permitted. However, the relatively small difference between the comparable percentage correct rates in the two studies indicate that the effect of these factors is small. In addition, Künzel also found that a set of listeners that had familiarity with the speakers did not perform any better than non-familiar listeners, suggesting that the familiarity of listeners with the speakers in this experiment may not be a strong influence.

There a several potential reasons for the difference in performance between younger-older and older-younger comparisons. Referring again to Künzel's study, it is suggested that this effect may be attributed to the fact that ageing has a 'natural' order, and thus the task is easier when the stimuli are presented in this order, i.e. a younger-older comparison. A second explanation put forward is that older voices are more likely to exhibit idiosyncrasies, and thus constitute a 'stronger' stimulus. In the younger-older ordering, with the older stimulus presented last, the listener may be able to make a correct decision even if their memory of the younger stimulus is weakened.

The first suggestion of the natural, chronological order of ageing, is reasonable. However, one can only speculate on whether this factor is at play. The second suggestion relating to the relative strength of the stimuli does not hold based on the experiments here; comparing the forwards and backwards correct decision rate in Table 6.1, the discrepancy between comparison directions for all speakers is smallest at a 30 year age difference in which there is one 'strong' aged stimulus and one young stimulus. The positioning effect is strongest at a 10 year age difference, where the correct decision rate for the comparison pair 1-2 is 74%, compared with 58% for the 4-3 comparison pair. The fact the detectability is higher between the youngest two samples suggests that ageing 'idiosyncrasies' are not responsible for the position effect.

An additional explanation may be the fact that listeners tend to underestimate the age of older voices and overestimate the age of younger voices [181]. Thus, in the older-younger comparison condition, assuming the listener first makes a judgement on the age of the older voice that is likely underestimated, the second younger sample may be more likely to sound older. The idea of a natural ageing order may contribute to this effect occurring only in the forwards direction.

The position effect is an interesting side-effect from this experiment, and would need a larger study to draw more informed opinions on its origin.

Speaker gender was found to be a strong factor in ageing detectability, with females being

significantly more detectable than males. A link between the ageing fundamental frequency (F0) trends of females and the detectability of ageing in their voices was observed. Such a link does not seem to be present in males. In her Thesis, Schötz [181] also reports that F0 is more important to the human perception of female speaker age than male speaker age.

The effect of listener gender was not as strong as speaker gender, but was still found to be significant. In Künzel's study, a group of adolescent girls performed marginally better than a group of boys of the same age. However, this effect was not found to be significant. In age perception experiments in Schötz's Thesis [181], there is no significant effect attributed to listener gender. It would be necessary to increase the number of listeners, and control for effects like listener age, before drawing conclusions about general ageing detection ability differences between genders. Having said that, the number of listeners who fully completed this experiment (31) is in the range of the usual number of responses for such a study, e.g. Schötz [181] analysed human age perception based on the responses of 31 and 29 listeners in two separate experiments.

6.4 Discussion

The experiments presented in this Chapter were designed with the aim of establishing the influence of ageing on aspects of forensic speaker recognition

In Section 6.2 the effect of ageing on forensic automatic speaker recognition was evaluated. An experiment using the 15 TCDSA males as subjects revealed that ageing has significant, but variable, effect on likelihood ratio (LR) estimation. A second experiment using 5 Irish-accented males was designed to restrict inter-speaker variability and provide a more detailed picture of the effect of ageing on LR estimation. The effect of ageing was again found to be significant, leading to errors in the LR estimate after 10-30 years. Explanations for inter-speaker variability were put forward. Eigenageing compensation, introduced for speaker verification, Chapter 5, was applied to the Irish-accented males experiment. It was shown to improve the strength of evidence in almost all same-speaker comparisons, and reverse errors in the uncompensated system in some cases, while not increasing the probability of erroneous support for different-speaker hypotheses.

In Section 6.3, the auditory detectability of ageing was investigated via a listener test. Based on the responses of 36 listeners, who completed a random set of questions eliciting them to identify the older of two samples of the same speaker, several findings were made. Detectability of ageing increases with age difference between the samples concerned, which would be expected based on physiological and perceptual predictions. The effect of speaker gender was significant, ageing more accurately detected in females. Listener gender was also significant, although less so than speaker gender, with females performing better at the task. Finally the effect of sample order is also significant, with older-younger comparisons proving more difficult than younger-older comparisons.

7

Effect of ageing on i-vector speaker verification

In the previous Chapters, a GMM-UBM speaker verification system has formed the basis for experiments. There have been significant developments in speaker verification research in recent years, with progress driven by the regular NIST speaker recognition evaluations (SREs) [51,78]. The current wave of systems, the majority of which build upon the GMM-UBM framework, incorporate various techniques to improve performance in the presence of inter-session variability. As a result, they have reached a level of performance that significantly outperforms the 'classic' GMM-UBM approach in challenging conditions; for example, a comparison of GMM-UBM with GSV-SVM and JFA is presented in Kinnunen's review paper [118]. A current research trend [178] is the use of an i-vector framework [48] with PLDA (probabilistic linear discriminant analysis) [154].

Optimisation of general verification performance is not the focus of this Thesis; what is of interest however, is to evaluate how these recent developments, aimed at dealing with inter-session variability, behave when faced with ageing variability. To investigate this, experiments using an i-vector system with PLDA modelling, developed at Radboud University Nijmegen (RUN) for the NIST SRE 2012 evaluation [78], are presented in this Chapter. A speaker verification evaluation using the TCDSA database is designed to observe the extent to which this 'state-of-the-art' system is affected by long-term ageing variability. The performance of the GMM-UBM system on the same evaluation is presented for comparison.

7.1 Experimental evaluation

In previous Chapters, a 'forwards' and 'backwards' verification approach has generally been adopted, whereby a speaker's youngest and oldest samples are used in training, and the remainder of their samples are reserved for testing. In this Chapter, the experimental protocol is expanded to use all speaker recordings in the TCDSA database for both training and testing. Thus the number of trials are maximised (where a trial is the comparison of two i-vectors, or in the case of the GMM-UBM system, the comparison of a test feature vector with a speaker GMM and the UBM).

There were a few constraints to this all-vs-all protocol: given the widely variable recording durations within the database, all training and testing samples were set at 30 seconds, and a maximum of five 30 second samples from a recording were used for testing. No same-session trials were considered at the results analysis stage. The full TCDSA database, Figure 3.2, was used, as opposed to the reduced database, Figure 3.13.

7.1.1 i-vector system description

The i-vector system was developed in Radboud University Nijmegen (RUN) for the NIST SRE 2012 evaluation [77]. It consists of a standard i-vector [48] configuration with PLDA modelling [30]. The system was used in an 'off the shelf' manner, in the configuration optimised for the NIST SRE task [177, 178]

All speech data used was at 8 kHz (downsampling was applied where necessary). The speech signal is enhanced by applying a Wiener filtering based module to the magnitude spectrum of the frames, with the noise spectrum estimated using an improved minima controlled recursive averaging (IMCRA) approach [41]. IMCRA operates by averaging previous estimates of the noise power spectra. The front-end consists of 19-dimensional MFCC (plus log energy) extraction over 20 ms windows every 10 ms. Delta and acceleration coefficients, computed over 9 consecutive frames, are then appended. Voice activity detection (VAD) is applied according to a Gaussian modelling of the frame energy [139]. Lastly, feature warping [151] is applied.

Gender-dependent UBMs of 2048 components were trained using segments from the following datasets: NIST SRE 2004-2006, Switchboard cellular phase 1 and 2 and Fisher English [178]. An i-vector extractor matrix T of rank 400 was estimated using the same utterances used to train the UBM. Baum-Welch statistics of 0th, 1st and 2nd order are computed using the UBM, and along with the T matrix, are used to extract i-vectors for the relevant utterances.

To reduce intra-speaker variability and enhance inter-speaker variability, LDA (linear discriminant analysis) is applied to the i-vectors, reducing their dimensionality to 200. Finally, the i-vectors are centred, whitened [88] and length-normalized [67].

The speaker and session dependent i-vector distribution is modelled with PLDA [154]. PLDA development data was drawn from NIST SRE 2006-2012 according to the I4U development lists [178]. A score for each trial is the log-likelihood ratio of the pair of i-vectors originating

from the same speaker versus different speakers. Score calibration was not applied.

7.1.2 GMM-UBM system description

The GMM-UBM system configuration is consistent with that in Chapters 5 and 6. All speech is preprocessed by downsampling to 8 kHz, applying pre-emphasis and removing silences with an energy-based voice activity detector. Feature extraction consists of 12-dimensional MFCC extraction over 20 ms windows every 10 ms. Delta coefficients, computed over 5 consecutive frames are appended. RASTA filtering, and mean and variance normalisation are then applied to the feature vector

A gender-independent UBM of 512 components was trained with the TCDSA-UBM database, Section 3.5.2. Speaker GMMs were trained with mean-only adaptation. A relevance factor of 16 was used, and all UBM components were considered in adaptation and scoring.

A log-likelihood ratio score is computed for each trial from the likelihoods of the test sample given both the speaker GMM and the UBM. Z-norm was applied to each score. The Z-norm statistics for each male speaker GMM were estimated given a set of 25 speakers from the TCDSA-FD database (Section 3.5.3). To calculate female Z-norm statistics, an additional database of 25 Irish-accented female speakers, similar in their profile and age distribution to the male-only TCDSA-FD database, was collected. Further details for the speakers in this database are provided in Section 3.5.4 and Table A.4.

7.1.3 Experimental results

The scores for all trials were analysed on a gender dependent basis, and according to the corresponding age difference between the training and testing recordings. In Figure 7.1, DET plots are shown for both systems for a range of conditions.

The top plot of Figure 7.1 is based on all trials with an absolute age difference of less than or equal to one year. The effect of ageing over a one year period is assumed to be minimal, and thus the resulting EERs are considered as ageing-independent baselines in this experiment. The i-vector system clearly outperforms the GMM-UBM system, with an absolute EER difference between the two systems of approximately 6% and 2% for males and females respectively. The i-vector EERs are slightly higher for females than for males, whereas in the GMM-UBM case, male EER is higher than that of females.

In bottom plot of Figure 7.1, the systems are compared given all male and female trials (apart from same-session trials) at *all* absolute age differences, from zero to 60 years. The i-vector system again outperforms the GMM-UBM system, with an absolute EER difference between the two systems of approximately 3% and 5% for males and females respectively. For both systems, the EER is higher in the case of female speakers than male speakers.

In Figure 7.1 all trials have been considered equally in generating the EERs and plotting the DET curves. However, the distribution of trials is not balanced across speakers. Referring to the

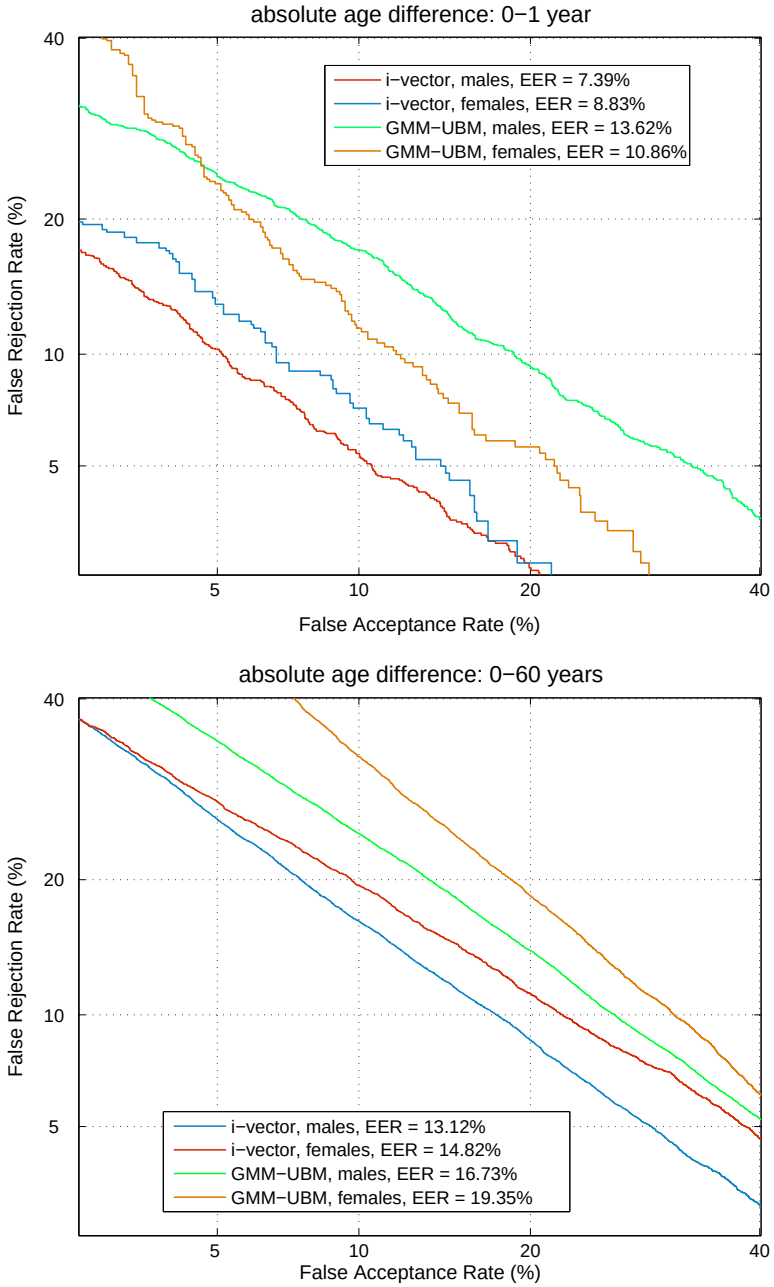

Figure 7.1: DET plots for the i-vector and GMM-UBM systems given male and female speakers. The EER for each condition is shown in the legend. **Top:** all same-year (different-session) trials and all trials with a one year absolute age difference. **Bottom:** all trials at all absolute age differences (0-60 years).

TCDSA database schematic, Figure 3.2, the varying range of recordings per-speaker is evident. The number of recordings per-speaker ranges from 4 to 47, accounting for multiple within-year recordings. To account for this imbalance, the EER can be calculated in a way that balances the contributions of different-speaker trials. In [126], van Leeuwen presents a framework for

pooling the trials from different conditions in an evaluation. The false acceptance rate (FAR) and false rejection rate (FRR) are determined for each condition individually and combined. This essentially has the effect of weighting each trial by the inverse of the number of trials of that condition. In this case, a different 'condition' is a different speaker. In Table 7.1, weighted EERs determined in this way are compared to the original EERs in Figure 7.1 (determined from the pooled set of unweighted trials).

age diff.	method	i-vector EER(%)	GMM-UBM EER(%)	N. targets	N. non-targets
males					
± 0-1	pooled	7.39	13.62	2,247	4,826
	weighted	4.61	6.90		
± 0-60	pooled	13.12	16.73	20,595	104,105
	weighted	11.52	18.21		
females					
± 0-1	pooled	8.83	10.86	441	1,734
	weighted	9.64	9.62		
± 0-60	pooled	14.82	19.35	8,901	33,219
	weighted	16.69	18.59		

Table 7.1: EERs for absolute age difference ranges 0-1 and 0-60 (all trials) for the i-vector and GMM-UBM systems. The pooled EERs are determined from an unweighted set of trials. The weighted EERs are determined by balancing the contribution of individual speakers. The number of target and non-target trials in each case is also provided.

Considering the male EERs in Table 7.1, weighting reduces the i-vector EER at both 0-1 and 0-60 absolute age differences by 2-3%. In the GMM-UBM case, there is a reduction in EER of over 6% at the 0-1 range after applying weighting, and an increase in EER at the 0-60 range. Weighting allows for a more meaningful comparison between the 0-1 and 0-60 age range, as the different contributions of speakers in both cases is accounted for. For both systems, there is a relative increase in male weighted EER of approximately 150% between the ageing independent 0-1 range and the full 0-60 age range.

Considering the weighted female EERs in Table 7.1, the GMM-UBM system performs marginally better than the i-vector system at a 0-1 age range. At the full 0-60 range however, the i-vector weighted EER is 2% lower. In general, the female EERs are higher than the equivalent male EERs, and the GMM-UBM system is closer in performance to the i-vector system. There are a lower number of female trials however, particularly at the 0-1 age range where there are only 441 target trials.

The results in Table 7.1 confirm that the i-vector system performance, although unoptimised for the task, exceeds that of the classic GMM-UBM approach. Both systems are affected by ageing to the same degree however, with an equivalent relative increase in EER observed in

both.

To evaluate the performance of each system at increasing age differences between training and testing samples, a number of error metrics were determined at each of six absolute age difference ranges: 0-10, 11-20, 21-30, 31-40, 41-50 and 51-60 years.

1. **Weighted EER**: the speaker-weighted EER, as evaluated in Table 7.1 was calculated at all age ranges. This is evaluated to indicate the effect of ageing on discrimination ability of the systems.

2. **Half-total error rate (HTER)**: from the trials in the range 0-10, the score threshold at the weighted EER point was extracted. Applying this threshold to the scores at each age range, the HTER (average of FAR and FRR) was determined. By using the 0-10 range as a 'development dataset' to determine the decision threshold, the HTER provides a more realistic measure of the way in which ageing affects the discrimination ability of the systems than the EER.

3. **Calibrated HTER**: to convert the scores into 'calibrated likelihood ratios' [136], linear calibration parameters were extracted from the scores of trials in the range 0-10. Shifting and scaling parameters, w_0 and w_1 respectively, were optimized on this range 0-10 development set with the Bosaris Toolkit [26]. The scores s of all trials at subsequent age ranges were then linearly mapped to calibrated scores s_{cal}:

$$s_{cal} = w_0 + w_1 s \tag{7.1}$$

   The calibrated HTER was determined by applying the score threshold at the weighted EER point of the 0-10 range to the calibrated scores at subsequent age ranges.

4. **Speaker average HTER**: the score threshold at the weighted EER point of the 0-10 range was applied to the scores *of each speaker individually* at each age range. The resulting HTERs were averaged to obtain the speaker average HTER. This measure was evaluated to observe the influence of the imbalance of speaker distribution on the HTER. By using a threshold determined from the 0-10 range, it provides a more realistic speaker-weighted measure than the weighted EER.

5. **Calibrated speaker average HTER**: the score threshold at the weighted EER point of the 0-10 range was applied to the *calibrated* scores of each speaker individually at each age range. The resulting HTERs were averaged to obtain the calibrated speaker average HTER.

In Figure 7.2, these error metrics are provided for both systems, in male and female cases.

The number of trials per range was uneven, with a progressively decreasing amount with increasing range. For reference, figure 7.3 indicates the number of target and non-target trials

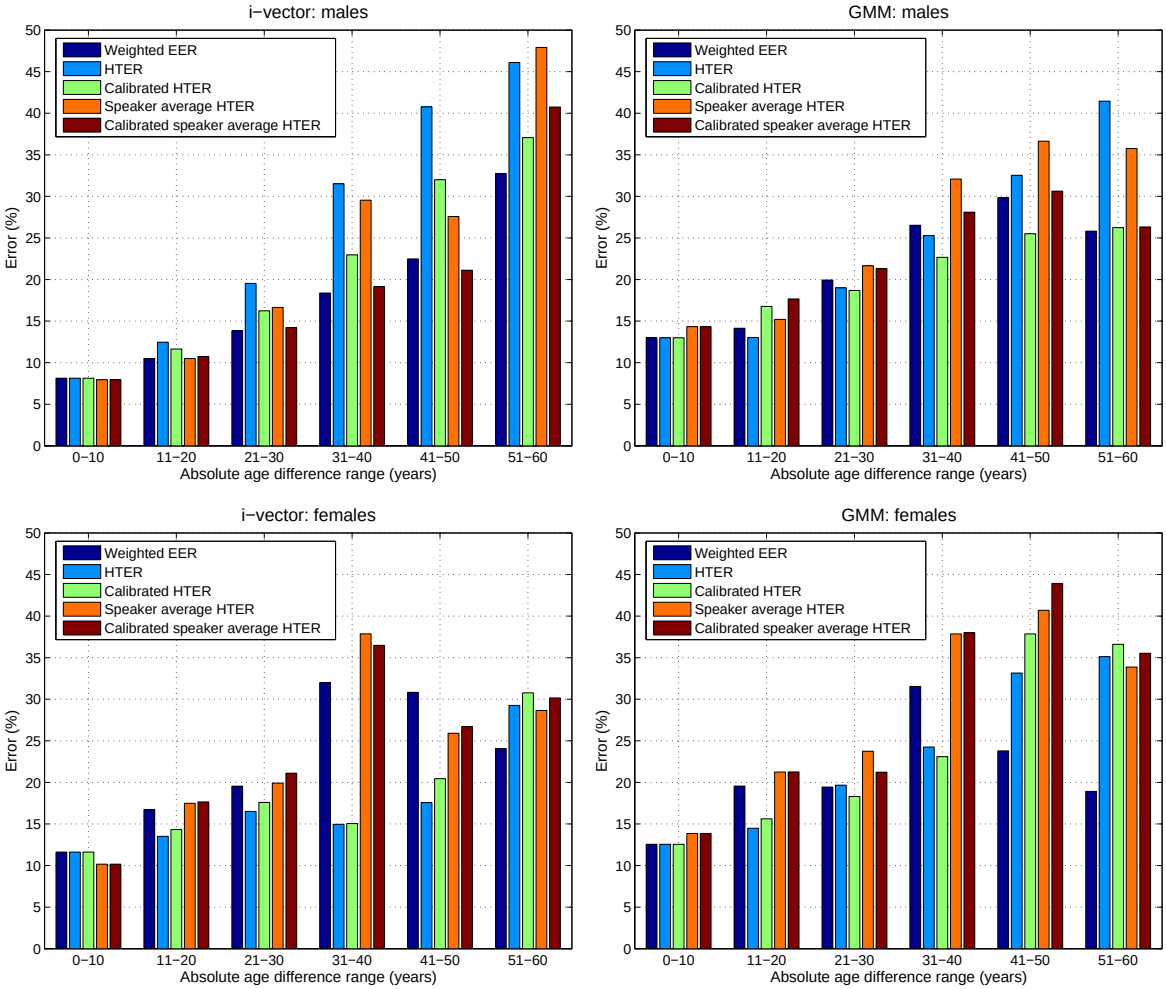

Figure 7.2: i-vector and GMM-UBM system performance given increasing ranges of absolute age difference between training and testing samples. Five error metrics are given for each condition. The EER is weighted per-speaker. For each HTER metric, the 0-10 age range is used to determine the decision threshold and/or calibration paramteers. **Top Left:** i-vector, males, **Top Right:** GMM-UBM, males, **Bottom Left:** i-vector, females, **Bottom Right:** GMM-UBM, females

at each age difference range. The minimum number of trials were at the 51-60 absolute age difference range, where there were approximately 200 male and 400 female target trials.

The plots in Figure 7.2 compare the performance of both systems in terms of their discrimination ability, as well as illustrating the effects of score calibration and of balancing the contribution of different speakers.

The weighted EER values in Figure 7.2 generally follow an increasing trend as the age ranges increase. The fall off in discrimination between targets and non-targets as age difference increases is an indication that their distributions are becoming progressively more overlapped. In the male case, the EER for the i-vector system at the range 0-10 is significantly lower than

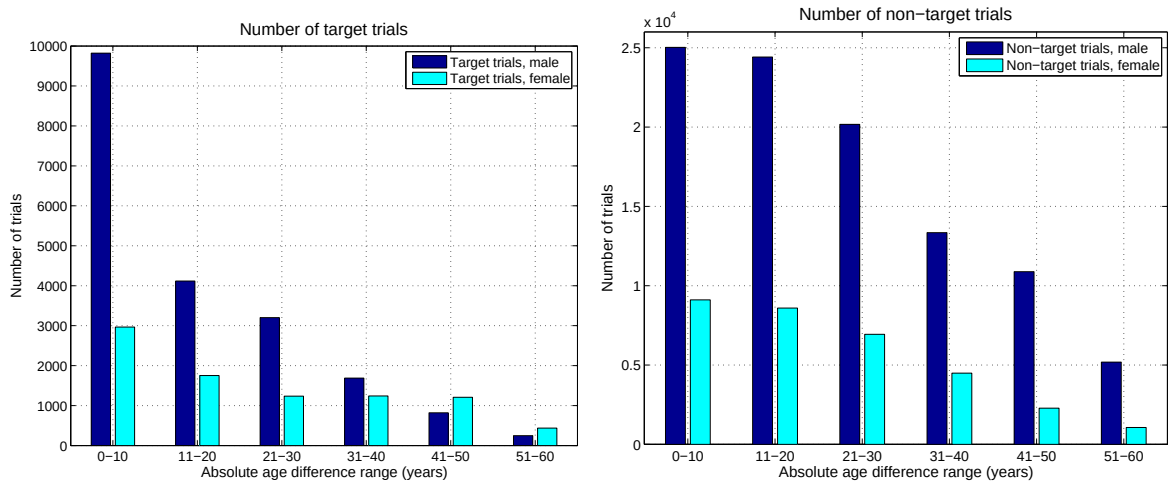

Figure 7.3: The number of trials for each of the age ranges considered in Figure 7.2. **Right:** target trials, **Left:** Non-target trials. Note the different y-axis scales.

the GMM-UBM system, as would be expected based on Figure 7.1. For subsequent age ranges, the EER increases for both systems at approximately the same rate. At the final range 51-60 however, the GMM-UBM EER drops below that of the i-vector system.

In the female case, there is little difference between the EERs of both systems at each age range. For both, there is an increase up to 31-40 range, followed by a decrease to the final age range. For the final four age ranges, the GMM-UBM EER is lower than that of the i-vector system. The reason for the 'spike' in the EER at 31-40 is unclear. At this age range, Figure 7.3, the number of target trials for males and females is similar. The particular distribution of speakers within this age range may be the cause.

The HTER values in Figure 7.2 follow a similar increasing trend to the EER values. In the male case, particularly for the i-vector system, the HTER results in a lower accuracy than the EER, and increases at the same rate. This would be expected, given that the threshold for the HTER is determined from the scores from the 0-10 range. The female case is again more variable, with the HTER lower than the EER in a number of cases. The fact the HTER is significantly lower than the EER at the i-vector age range 21-30 backs up the assumption that it is the particular distribution of speakers at this age range that causes the EER spike (the HTER is not speaker-weighted at the range range 21-30, unlike the EER).

The effect of linear score calibration is evident in the 'Calibrated HTER' values of Figure 7.2. For the male case, there is a reduction in error rate between HTER and the calibrated HTER, and the improvement is more noticeable at longer age ranges. In the female case, calibration generally increases the HTER, although the effect is small. There are significantly less female trials than male trials at age range 0-10, Figure 7.3. The number of trials and also the distribution of individual speakers with the age range may have led to calibration parameter estimates for females that generalized poorly to the other age ranges.

For the male case, the 'Speaker average HTER' values in Figure 7.2, are generally higher than the HTER values, particularly at larger age differences. When calibration is included however, i.e. 'Calibrated speaker average HTER', the error values are generally the lowest across all the HTER metrics, and closest to the weighted EER. Using this metric for comparison, the GMM-UBM and i-vector system performance can be seen to degrade at the same rate as ageing progresses, with i-vector HTER approximately 6% lower than the GMM-UBM system up until the final age range, at which it is approximately 15% higher than the GMM-UBM.

Using the 'Calibrated speaker average HTER' to compare female performance between the two systems, a similar increase in error to the corresponding male rates are observed, albeit with variability around the 31-40 and 41-50 age ranges. The i-vector HTER is approximately 1-5% lower than the GMM-UBM system apart from the age range 41-50, at which it is approximately 15% lower than the GMM-UBM system.

7.2 Discussion

In this Chapter, the effect of ageing on the performance of a state-of-the-art i-vector speaker verification system was presented and compared to the GMM-UBM approach used in previous Chapters. An experimental protocol was designed that maximised the number of trials given the TCDSA database.

For male speakers, the i-vector system significantly outperforms the GMM-UBM approach at 'short' age difference ranges (0-1 and 0-10) and overall (over the complete 0-60 range). This result is expected given the additional levels of inter-session compensation present in the i-vector framework. For female speakers however, there is a less significant difference between the i-vector and GMM-UBM approaches (in terms of weighted EER) at 'short' age difference ranges (0-1 and 0-10) and overall (over the 0-60 range). It is unclear why there is not a more marked improvement with the i-vector system. Overall, there are less than half the number of female target trials than male target trials (8,901 compared to 20,595). The fewer number of trials and the uneven speaker distribution may be contributing factors.

The EER of the i-vector system for males at the 0-1 range (4.61%) is comparable to the performance of the system in the NIST SRE 2012 common conditions 'CC1' (EER = 5.75%) and 'CC3' (EER = 4.83%) [178]. This confirms that although the i-vector system was not optimised for this experiment, e.g. by including data similar to the TCDSA database content in development, it performs effectively.

For females, the EER at the 0-1 range (6.90%) is slightly higher than the common conditions 'CC1' (EER = 4.86%) and 'CC3' (EER = 4.09%) [178]. This slight discrepancy may indicate the i-vector system is not as well suited to the female content of the TCDSA database, resulting in a greater discrepancy between male and female EER at the 0-60 age range (11.52% vs 16.69%).

An important observation in the context of this Thesis is that the relative change in i-vector performance (in terms of EER or HTER) in the presence of ageing is equivalent to that of the

GMM-UBM system. This clearly demonstrates that long-term ageing variability is not reduced by the inter-session compensation applied in the i-vector framework. Vocal ageing presents a different form of variability than standard inter-session variability compensation frameworks are capable of dealing with. This therefore justifies dedicated compensation approaches for ageing variability, such as those proposed in this Thesis. In addition, this supports the use of a GMM-UBM system at the investigative stage, for which the data requirements are significantly less.

The large changes to both EERs and HTERs after balancing the contribution from each speaker shows the importance of condition weighting in results analysis. The large spread in the number of per-speaker recordings, and the small number of speakers make the effect particularly significant here. This is likely to be an unavoidable feature of long-term ageing evaluations, due to the nature of the type of data that is required.

Score calibration was clearly effective in the male case, moving the speaker average HTER very close to the weighted EER. However, the contribution of each speaker was not considered as a parameter of the calibration. This is the likely reason that the calibration was not effective in the case of females - due to different distributions of speakers between the 'training' (range 0-10) and 'testing' (ranges 11-60) trials.

Score calibration offers another possibility: by incorporating *age difference* in the calibration model, ageing could be compensated for at the score level. A similar calibration strategy was adopted by Mandasari et al. [136] to account for variability due to recording duration. This strategy could be applied to ageing variability by incorporating an additional term into the standard linear calibration model (Equation 7.1):

$$s_{cal} = w_0 + w_1 s + w_2 Q\left(a_1, a_2\right) \tag{7.2}$$

where Q is a function applied to the age of the speaker in the training recording a_1, and the age of the speaker in the testing recording a_2. Q could simply be the absolute age difference, or some other combination of a_1 and a_2. The parameter w_2 is optimized at the training stage of the calibration. This idea is in fact very similar to the inclusion of ageing and quality information in the stacked classifier framework, Chapter 4. However, the calibration approach has the advantage in that could be readily integrated into an existing score calibration module. As a practical consideration, particularly from the perspective of the evaluation in this Chapter, it would be necessary to balance the contribution of both different speakers and different age differences in this model.

The results of the evaluation in this Chapter provide benchmark levels of performance on the TCDSA database. Since the database has been made freely available for academic research, this will hopefully stimulate research in the area.

8

Conclusions

The experiments in this Thesis demonstrate that long-term vocal ageing significantly degrades speaker recognition performance. Ageing variability was shown to progressively increase classification error in an automatic speaker verification system. In a forensic automatic speaker recognition (FASR) system, the presence of vocal ageing results in an increase in the proportion of cases where the evidence supports the incorrect hypothesis. In Chapter 7, the effect of ageing on the performance of both i-vector and GMM-UBM systems, using the Trinity College Dublin Speaker Ageing (TCDSA) database, was presented. A similar rate of performance degradation was observed for both systems, establishing that the ageing problem is not solved by standard inter-session variability compensation methods.

The long-term ageing problem is not just of interest for its 'curiosity factor'; there is currently rapid growth in the deployment of commercial biometric systems, including large-scale speaker recognition systems operating in real-world conditions. As the user-bases and operating durations of these systems grow, the impact of ageing is inevitable.

Ageing is a challenging variability to deal with; due to genetic predisposition, hormonal, health and environmental influences, every voice ages in a unique way. The *heterogeneity* of ageing was observed throughout this thesis: In Chapter 3, the acoustic features of ageing speakers displayed widely varying behaviour, in line with expectations from physiology. The speaker recognition scores of individual speakers, presented in Chapters 3 and 6, displayed rates of degradation that varied both between speakers and within a speaker at different ages. From a perceptual perspective, it was observed in Chapter 6 that the effect of ageing is more noticeable in some speakers than others.

Despite this heterogeneity, ageing manifests itself in a speaker recognition system in a way that is still general enough to apply a speaker-independent compensation approach. In Chapter 4, a common decline in the scores of genuine-speakers as the age difference between training and testing increased was observed. The scores of imposters, over the same period, were relatively stable. This trend was exploited by implementing an ageing-dependent decision threshold. Although effective at reducing long-term error rates, it relied on the assumption that the scores of ageing speakers decay linearly. Considering the complexity of the ageing process on the voice (i.e. relative change in different acoustic features, Chapter 3), modelling the effect at the *score-level*, to a high level of accuracy, would require much larger amounts of data. A more feasible approach, based on the ageing-related changes in the components of a speaker model, was introduced in Chapter 5. The *model-level* eigenageing compensation method succeeded in reducing the impact of ageing on speaker verification error, and demonstrated that there are common directions of change in the models of ageing speakers. This promising approach has the potential to enable a speaker-independent solution to speaker-dependent ageing variability. The notion of an ageing model also presents some other applications, which are commented on in Section 8.4.

8.1 Forensics

Forensic speaker recognition is of particular relevance where vocal ageing is concerned; due to the nature of the domain, the speech samples of interest are always non-contemporary. While a time-lapse of a few months is the norm, time-lapses in the range of years can, and do, occur [63]. Forensic speaker recognition is undergoing a 'paradigm shift' [75] towards emulating other forensic disciplines, such as DNA, in quantifying the value of the evidence. This relies on transparent, reliable and testable means of comparing speech samples. Thus, a likelihood-ratio framework for evaluating evidential strength is being increasingly used in practice for semi-automatic (i.e. acoustic-phonetic) and fully-automatic (i.e. FASR) approaches.

The observation of the large ageing-related shift in the acoustic features of the voice (Chapter 3) has clear implications for acoustic-phonetic LR estimation, which relies on measures such as fundamental frequency. In evaluating the effect of ageing on an FASR framework (Chapter 6), the shift in the value of the evidence was significant. After a 10 year age difference, a number of the LR estimates had shifted from correctly supporting the same-speaker hypothesis to incorrectly supporting the different-speaker hypothesis. Considering that forensic practitioners are likely to use commercially available systems, like BATVOX [4], as a 'black box', this is certainly a cause for concern. The recordings of Irish male speakers used in this evaluation are of significantly higher quality than what would be expected in a real forensic environment. In addition, they all are healthy, professional speakers who have been geographically stable. Since the impact of ageing is significant in this optimistic scenario, then it is fair to assume it will be at least as severe in a real case.

In Chapter 6, eigenageing compensation was shown to reduce the negative impact of ageing on FASR by increasing the strength of the evidence in cases where the same-speaker hypothesis was true. It did so without increasing the proportion of cases where the evidence incorrectly supports the different-speaker hypotheses, demonstrating its suitability for forensics.

There remains a question of reference population selection for forensics; both in terms of the speakers used for estimating a between-source distribution and estimating a vocal ageing subspace. While ideally they should be tailored to the speaker, compromises have to be made in reality due to a non-exhaustive supply of data. There is the additional problem surrounding age; if there is an age difference between the evidential and suspected-speaker samples, and a corresponding difference in vocal qualities (and potentially accent), it is unclear on which sample to base the reference population, or the effect of this choice on the strength of the evidence.

The listener test in Chapter 6 revealed that vocal ageing is strongly perceptible in the voices of some speakers, particularly females. Thus, aspects of forensics that rely on subjective listening, including speaker profiling, selection of features for acoustic-phonetic analysis, and auditory-perceptual analysis, may be adversely affected when dealing with non-contemporary samples.

8.2 TCDSA database

The TCDSA database compiled for this Thesis is, at the time of writing, the largest longitudinal speech database in the public domain (in terms of the number of subjects, and the quantity and longitudinal range of the recordings). Acquiring long-term data free of non-ageing-related variability is not a realistic aim. Thus, the primary goal in compiling the database was to ensure that ageing was the *dominant* source of variability.

Collectively, the experiments throughout this Thesis provide strong evidence that this aim was met: A trend of progressive degradation in the verification scores of genuine-speakers, and a contrasting stability in the scores of imposters, was observed in the first speaker verification experiments in Chapter 3. The presence of a progressive variability in the data, affecting genuine-speakers only, aligns with the expected effect of cumulative vocal ageing changes.

In Chapter 5, a model of vocal ageing change was determined via a principal components analysis (PCA) of the TCDSA data. The successful application of the resulting model to age estimation demonstrated that ageing is a prominent variability in the data.

In Chapter 7, an i-vector speaker verification system achieved a lower level of classification error than a GMM-UBM system when tested with the TCDSA data. However, the performance of both systems degraded at the same rate in the presence of ageing. This experiment provides an interpretation of the levels or variability entangled in the data; 'normal' non-ageing-related variability, which is dealt with by the inter-session compensation modules of the i-vector framework, is superimposed on longer-term ageing variability, which demands a specific compensation approach (such as the stacked classifier and eigenageing compensation proposals).

Thus, the TCDSA database, which has been made freely available for research, is a suitable resource for studying vocal ageing.

8.3 Male and Female ageing trends

A recurring observation throughout this Thesis has been the difference between male and female vocal ageing. In general, the patterns of vocal change that accompany ageing are gender-dependent, and are more consistent between females than between males. The analysis of the acoustic correlates of ageing in Chapter 3 demonstrated this experimentally. Female fundamental frequency (F0), in particular, followed a trend that was clearly different from males, and that was consistent between speakers. Gender-dependent ageing change, from a physiological perspective, has been well-documented, e.g. [12]. Gender-dependent ageing change in acoustic features of the voice have also been highlighted previously, e.g. [132]. However, gender-dependent ageing change in a speaker modelling and recognition context has not previously been demonstrated.

In the first evaluation of the GMM-UBM system with ageing data, Chapter 3, female verification scores degraded with ageing at a faster rate than males. In Chapter 5, in the creation of a vocal ageing subspace, the results of PCA demonstrated that the dominant directions of change in the models of female speakers were of significantly higher magnitude than in the male case. For the i-vector system evaluation in Chapter 5, as the age difference between training and testing samples grow, the error rate (calibrated speaker average half-total error rate (HTER)) is higher in absolute terms, and more rapidly increasing, in females than in males. Collectively, these observations point towards ageing change in females that is more consistent between speakers, and possibly of greater magnitude, than in a comparable group of male speakers.

This point raises the issue of gender-dependency in speaker recognition systems in general; a gender-dependent modelling approach is usually taken in speaker verification systems. However, performance is often worse with female speakers, e.g. this is observed in an evaluation of both NIST SRE 2008 and 2010 tasks [44]. Ageing may inflate gender-dependent performance differences, especially considering the consistency in female F0 decrease for example. In addition to an ageing compensation approach that is gender-dependent, a front-end that is gender-aware may ultimately be necessary to overcome this male/female performance discrepancy.

The listener test in Chapter 6 demonstrated that vocal ageing is significantly more detectable in the voices of female speakers than male speakers. In addition, it was shown that the detectability of ageing in female voices is correlated with the relative decrease in fundamental frequency. Thus, F0 has an important influence on the way humans perceive ageing in female voices.

8.4 Future work

The research described in this Thesis presents a variety of opportunities for further study:

TCDSA database expansion

As discussed in Section 8.2, the method of TCDSA database collection, primarily via the archives of national broadcasters, was demonstrated to provide a suitable resource for the study of vocal ageing. With a collaborative effort from other researchers, this process could be repeated for national and regional broadcasters in other English-speaking countries, which could quickly expand the size of the database. Any study approaching vocal ageing would benefit from a larger set of ageing speakers.

Features

In this Thesis, only the short-term spectral features, MFCCs, were considered. It would be of interest to evaluate the effect of ageing on other features. It is likely that 'high-level' features, capturing long-term prosodic and lexical information, are more robust to ageing effects. In terms of short-term spectral features, an alternative filterbank to the Mel scale may prove more resilient to ageing, particularly in females. Mason and Thompson [137] demonstrated that a linear filterbank improves female speaker recognition performance. A first step of investigation could be to evaluate if this proposal improves female speaker recognition performance in the presence of ageing.

Stacked Classifier

This approach could be implemented as part of a linear calibration model, as proposed in Chapter 7. It could be included in addition to other 'quality' factors, such as duration or indeed, any of the quality measures evaluated in Chapter 4. As discussed, ageing is not a linear process, and thus it is likely that such a model would need to be non-linear to achieve maximum discrimination potential. An increase in complexity however, brings with it the requirement for more data.

Eigenageing compensation

As with a score-level approach, eigenageing compensation would greatly benefit from additional data to model the ageing subspace. A restricted age range (20-60) was considered in the subspace modelling process. The end goal would be to extend this model to the whole age range. A first step towards this aim could be to model a subspace for several, overlapping age regions. Given an automatic age estimation front-end (with a minimum accuracy subject to the range of subspace age regions), an age appropriate subspace would be selected. It remains to be seen how effectively the rapid changes in childhood and adolescent speech could be captured. The speech of older speakers may also present problems due to the prevalence of pathological voice conditions in old age.

Eigenageing compensation presents a promising approach for automatic age estimation,

which needs to be more fully evaluated, using male and female speakers.

The model of ageing change generated in eigenageing compensation could also be applied to aid automatic synthesis of ageing in a speaker's voice. This could be applied to long-term realistic speech synthesis as part of a speech-generating device for individuals with severe speech impairment.

Forensics

The issue of reference population selection is of concern for FASR, especially when dealing with non-contemporary samples. Future study should evaluate the effect of age mismatch between the reference population, suspect-speaker and evidential recording on the likelihood ratio. As ultimately, there will be a compromise between the ideal and actual reference populations due to data requirements, perhaps alternative means of between-source distribution estimation should be considered.

Eigenageing compensation should be tested fully within a forensic scenario, including a determination of the effect of the subspace speaker composition.

Perceptual

The link between F0 and perception of ageing prompts an investigation into other objective measures of a speech signal that are correlated with the *perceived* age of a speaker. This could enable an objective assessment of the quality of vocal ageing synthesis or give impartial feedback to individuals undergoing voice therapy.

A

Speaker Ageing Data

A.1 Trinity College Dublin Speaker Ageing (TCDSA) database contents

Table A.1: Biographical information of TCDSA database speakers. In the **Source** field, 'B' = BBC, 'R' = RTÉ, 'U' = The Up Series, 'Y' = YouTube, and 'M' = The Miller Center [153]. In **Accent** field, 'RP' is received pronunciation and 'Dublin' indicates a mainstream Dublin accent. In the **Notes** field, professions for which speaking plays a central role are listed; these are 'professional speakers'. Any health-related information is provided in this field. Note that the 'smoker' label indicates a speaker is known to be a long-term smoker. The absence of this label does not rule out the possibility that a speaker smokes.

Speaker	Source	Gender	Accent	Age Range	Notes
Andrew	U	m	British RP	21 - 56	
Black	Y	f	British, Liverpool	24 - 62	Presenter
Bowman	R	m	Irish, Dublin	21 - 64	Presenter
Bruce	U	m	British RP	21 - 56	Teacher
Byrne	R	m	Irish, Dublin	31 - 75	Presenter
Cooke	B	m	British RP	39 - 96	Presenter, mobile U.S. accent influence
Cronkite	Y	m	U.S., midlands	47 - 91	Presenter

Continued on next page

Table A.1 – *Continued from previous page*

Speaker	Source	Gender	Accent	Age Range	Notes
Doyle	R	f	Irish, Dublin	28 - 57	Presenter
Dunne	R	f	Irish, Dublin	26 - 54	Presenter
Finucane	R	f	Irish, Dublin	29 - 59	Presenter, smoker
Gogan	R	m	Irish, Dublin	27 - 71	Presenter
Lawlor	R	f	Irish, Dublin	23 - 48	Presenter
Lockwood	B	f	British RP	35 - 64	Actress, smoker
Lynn	U	f	British, Cockney	21 - 56	smoker
Magee	R	m	Irish, Dublin	42 - 72	Presenter
Neill	U	m	British RP	21 - 56	mobile, ongoing mental health issues
Ní Bhriain	R	f	Irish, Dublin	19 - 58	Presenter
Nick	U	m	British, Yorshire	21 - 56	Teacher, mobile, U.S accent influence
O'Dulaing	R	m	Irish, Cork	31 - 75	Presenter
Plomley	B	m	British RP	37 - 71	Presenter
Queen E. II	B+Y	f	British RP	26 - 84	Public Figure. YouTube samples age 80+ only
Reagan	M	m	U.S., midlands	37 - 83	Politician
Suzy	U	f	British RP	21 - 56	
Symon	U	m	British, London East-End	21 - 56	Afro-Caribbean accent influence
Thatcher	B+Y	f	British RP	34 - 81	Politician, onset of dementia in late 70s
Tony	U	m	British, Cockney	21 - 56	Actor (part-time)

A.2 TCDSA-UBM database contents

Table A.2: UBM database speakers

Gender	Age	Accent	Gender	Age	Accent
f	18	American, neutral	m	19	Irish, regional
f	19	English, neutral	m	19	American, neutral
f	19	Irish, neutral	m	19	American, neutral
f	20	English, neutral	m	20	Irish, neutral

Continued on next page

Table A.2 – *Continued from previous page*

sex	Age	Accent	sex	Age	Accent
f	20	English, neutral	m	20	English, regional
f	20	American, neutral	m	20	American, regional
f	22	English, regional	m	21	American, neutral
f	22	American, neutral	m	22	English, neutral
f	22	American, neutral	m	22	American, neutral
f	23	American, neutral	m	23	American, neutral
f	23	English-American, neutral	m	24	American, neutral
f	25	Irish, regional	m	25	American, neutral
f	25	American, neutral	m	25	American, neutral
f	25	American, neutral	m	26	American, neutral
f	26	American, neutral	m	26	American, neutral
f	27	American, neutral	m	27	English, regional
f	29	Irish, neutral	m	27	American, neutral
f	29	English, neutral	m	28	Irish, neutral
f	30	American, neutral	m	31	Irish, neutral
f	31	Irish, neutral	m	31	Irish, regional
f	40	Irish, neutral	m	37	English, neutral
f	40	American, neutral	m	37	Irish, neutral
f	40	American, neutral	m	39	American, neutral
f	41	Irish, neutral	m	40	Irish, regional
f	42	American, neutral	m	40	English, neutral
f	43	English, neutral	m	41	American, neutral
f	44	American, neutral	m	41	American, neutral
f	45	American, neutral	m	42	American, neutral
f	45	English, neutral	m	43	American, neutral
f	45	Irish, regional	m	43	American, neutral
f	48	American, neutral	m	44	American, neutral
f	48	American, neutral	m	45	English, neutral
f	49	American, regional	m	46	English, neutral
f	50	American, neutral	m	48	American, neutral
f	50	American, neutral	m	48	Irish, neutral
f	50	Irish, neutral	m	49	English, neutral
f	52	American, regional	m	50	American, neutral
f	54	American, neutral	m	53	English, regional
f	55	Irish, neutral	m	54	American, southern

Continued on next page

Table A.2 – *Continued from previous page*

sex	Age	Accent	sex	Age	Accent
f	57	English, neutral	m	54	American, regional
f	65	Irish, regional	m	62	American, neutral
f	65	American, regional	m	62	American, southern
f	68	American, neutral	m	65	English, neutral
f	69	Irish, neutral	m	65	American, regional
f	69	American, neutral	m	67	American, neutral
f	70	English, neutral	m	68	American, neutral
f	70	English, neutral	m	70	English, neutral
f	72	Irish, regional	m	70	English, neutral
f	73	English, neutral	m	70	American, neutral
f	74	English, rp	m	71	Irish,neutral
f	75	American, neutral	m	75	American, neutral
f	75	American, neutral	m	80	Irish,neutral
f	76	English, neutral	m	81	Irish,neutral
f	80	Irish, neutral	m	82	English, neutral
f	81	English, neutral	m	84	English, neutral
f	81	American, regional	m	84	American, neutral
f	83	American, neutral	m	90	American, neutral
f	87	American, neutral	m	91	American, neutral
f	91	American, neutral	m	93	American, neutral
f	99	American, neutral	m	100	American, regional

A.3 TCDSA Forensic Development (TCDSA-FD) database contents

Table A.3: TCDSA Forensic Development Database Speakers. Where the age of a speaker is unknown, the cell is left blank (in these cases the age range is either known or estimated).

Name	Age range	Age	Accent	Length(s)	Content	Location
Derek Nolan	25-35	29	Irish, neutral	243	speech	hall
David Tobin	25-35	30	Irish, south-west	148	speech	hall
Paul Dillon	25-35	29	Irish, dublin	47	speech	hall
John Lyons	25-35	35	Irish, dublin	300	speech	hall
Pearse Doherty	25-35	34	Irish, north	258	speech	hall
Paul McAuliffe	25-35	32	Irish, dublin	196	speech	outdoors

Continued on next page

Table A.3 – *Continued from previous page*

Name	Age range	Age	Accent	Length(s)	Content	Location
Darragh O'Neill	25-35	25	Irish, dublin	137	speech	hall
Rudhain MacAodhain	25-35	27	Irish, neutral	188	interview	studio
Ian McArdle	25-35	27	Irish, neutral	147	speech	hall
unknown	25-35		Irish, neutral	243	speech	hall
unknown	25-35		Irish, south-dublin	38	speech	studio
unknown	25-35		Irish, neutral	39	speech	studio
unknown	25-35		Irish, dublin	24	speech	studio
unknown	25-35		Irish, dublin	21	speech	studio
unknown	25-35		Irish, south-dublin	30	speech	studio
Niall Breslin	25-35	31	Irish, west	93	interview	studio
Abie P. Bowman	25-35	29	Irish, south-dublin	243	speech	studio
Simon Carswell	25-35	35	Irish, neutral	284	speech	office
Carl O'Brien	25-35	33	Irish, neutral	252	speech	office
Bernard Dunne	25-35	32	Irish, dublin	183	speech	kitchen
Peter Crooks	25-35	32	Irish, south-dublin	139	interview	office
Cillian Murphy	25-35	35	Irish, cork	280	interview	studio
Johnathan R. Meyers	25-35	34	Irish, south-dublin	247	interview	studio
Kenny Egan	25-35	29	Irish, dublin	212	interview	studio
Danny O'Reilly	25-35	25	Irish, south-dublin	26	interview	studio
Brian O'Driscoll	25-35	31	Irish, south-dublin	34	interview	studio
Donal Óg Cusack	25-35	31	Irish, south-west	32	interview	studio
Domhnall Gleeson	25-35	28	Irish, south-dublin	34	interview	studio
unknown	25-35	25	Irish, east	27	speech	studio
Robin Wilson	36-45	37	Irish, neutral	209	interview	office
unknown	36-45		Irish, south-dublin	119	interview	office
unknown	36-45		Irish, neutral	96	interview	office
Gerald Fleming	36-45	45	Irish, east	120	speech	studio
Aodhn Rordin	36-45	36	Irish, neutral	342	speech	hall
Ged Nash	36-45	37	Irish, dublin	308	speech	hall
Joe Tynan	36-45	36	Irish, neutral	212	interview	office
Harry McGee	36-45	36	Irish, south-dublin	316	speech	hall
Dan O'Brien	36-45	40	Irish, south-dublin	249	interview	office
Nevin Maguire	36-45	37	Irish, east	114	speech	kitchen
David McWilliams	36-45	45	Irish, south-dublin	152	speech	hall
Roy Keane	36-45	40	Irish, south-west	248	interview	studio

Continued on next page

Table A.3 – *Continued from previous page*

Name	Age range	Age	Accent	Length(s)	Content	Location
Hector hEochagin	36-45	40	Irish, east	33	speech	studio
Dara O'Briain	36-45	38	Irish, south-dublin	108	interview	studio
Glen Hansard	36-45	42	Irish, dublin	193	interview	studio
Ardal O'Hanlon	36-45	43	Irish, east	349	interview	studio
Brian Dobson	36-45	41	Irish, south-dublin	26	interview	studio
Mark Little	36-45	41	Irish, neutral	31	speech	outdoors
Darren Frehill	36-45	36	Irish, neutral	27	speech	studio
unknown	36-45		Irish, south-west	33	interview	office
Ken Hardy	36-45	40	Irish, neutral	36	speech	office
Niall Crowley	46-55	55	Irish, south-dublin	300	speech	hall
Eamonn Gilmore	46-55	54	Irish, west	301	speech	hall
Michael O'Leary	46-55	49	Irish, west	139	speech	hall
Kevin Humphreys	46-55	54	Irish, dublin	162	speech	hall
Michel Martin	46-55	52	Irish, south-west	214	speech	hall
Dominic Hannigan	36-45	46	Irish, neutral	282	speech	hall
Ray Darcy	36-45	47	Irish, east	225	interview	studio
Colm Tobn	46-55	55	Irish, neutral	327	interview	studio
Bono	46-55	52	Irish, neutral	318	interview	office
unknown	46-55		Irish, neutral	233	interview	studio
San Gallagher	46-55	49	Irish, east	226	interview	hall
unknown	46-55		Irish, neutral	120	speech	office
Jim Corr	46-55	48	Irish, east	324	speech	office
Danny McCoy	46-55	50	Irish, neutral	205	speech	hall
John Kelly	46-55	46	Irish, north	328	speech	hall
George Lee	46-55	48	Irish, neutral	74	interview	studio
unknown	46-55		Irish, neutral	88	interview	office
Eamonn Ryan	46-55	48	Irish, neutral	33	speech	office
John Curran	46-55	50	Irish, dublin	31	interview	office
Morgan Kelly	46-55	46	Irish, south-dublin	32	interview	studio
Aidan Nulty	46-55	50	Irish, dublin	33	speech	studio
Eamonn Donaghy	46-55	55	Irish, north	33	speech	studio
John Gormley	56-65	61	Irish, neutral	306	speech	hall
Ruairi Quinn	56-65	63	Irish, neutral	234	speech	hall
unknown	56-65		Irish, neutral	48	interview	office
Eamonn Maloney	56-65	58	Irish, north	298	speech	hall

Continued on next page

Table A.3 – *Continued from previous page*

Name	Age range	Age	Accent	Length(s)	Content	Location
Robert Dowds	56-65	58	Irish, dublin	337	speech	hall
Pat Kenny	56-65	64	Irish, south-dublin	325	interview	studio
John McCarthy	56-65	61	Irish, south-west	363	speech	office
Brendan Ryan	56-65	60	Irish, dublin	313	speech	hall
Patsy McGarry	56-65	58	Irish, west	261	interview	office
Denis Doherty	56-65	62	Irish, north	86	speech	office
John Ellis	56-65	60	Irish, neutral	46	interview	office
Bertie Ahern	56-65	57	Irish, neutral	319	interview	office
Gabriel Byrne	56-65	58	Irish, dublin	269	interview	office
Mick Lally	56-65	65	Irish, west	189	interview	office
Brendan Gleeson	56-65	56	Irish, dublin	338	interview	office
Eric Byrne	56-65	65	Irish, dublin	31	speech	hall
Christy Moore	56-65	64	Irish, east	34	interview	studio
Brendan Howlin	56-65	56	Irish, east	32	speech	office
Colm Keena	56-65	56	Irish, dublin	34	interview	office
Eamonn Lawlor	56-65	60	Irish, neutral	34	interview	hall
Garret Fitzgerald	66+	81	Irish, neutral	309	speech	hall
George Hook	66+	71	Irish, neutral	278	interview	studio
Michael D. Higgins	66+	70	Irish, west	324	speech	hall
Richard Harris	66+	72	Irish, neutral	146	interview	office
Mike Murphy	66+	70	Irish, neutral	131	interview	studio
Eamon Dunphy	66+	66	Irish, dublin	328	interview	office
Seamus Heaney	66+	72	Irish, north	182	interview	office
Bill O'Herlihy	66+	72	Irish, cork	335	interview	studio
Terry Wogan	66+	69	Irish-English	124	speech	studio
Duncan Stewart	66+	66	Irish, west	238	speech	outdoors
John Giles	66+	70	Irish, dublin	334	interview	studio
Vincent Browne	66+	67	Irish, neutral	290	interview	studio
David Begg	66+	66	Irish, dublin	182	speech	hall
John Montague	66+	70	Irish, neutral	203	speech	hall
Michael Longley	66+	72	Irish, north	118	interview	outdoors
San Kenny	66+	69	Irish, west	83	speech	hall
Joe Costello	66+	66	Irish, west	217	speech	hall
John Suttle	66+	66	Irish, dublin	31	speech	hall
Fergus Finlay	66+	70	Irish, neutral	32	interview	studio

Continued on next page

Table A.3 – *Continued from previous page*

Name	Age range	Age	Accent	Length(s)	Content	Location
David Kelly	66+	71	Irish, neutral	29	interview	outdoors
Bill Cullen	66+	68	Irish, dublin	34	interview	studio
Martin Donlan	66+	66	Irish, west	31	interview	studio

A.4 Additional Female Irish-accented Speakers

Table A.4: Additional Female Irish-accented Speakers

Name	Age range	Accent	Length (s)	Content	Location
f01	20-40	Dublin mainstream	31	spontaneous	studio
f02	20-40	Dublin south	31	spontaneous	studio
f03	20-40	Dublin south	32	spontaneous	studio
f04	20-40	Dublin south	31	spontaneous	studio
f05	20-40	Dublin mainstream	32	spontaneous	studio
f06	20-40	Dublin south	32	spontaneous	studio
f07	20-40	Dublin mainstream	31	spontaneous	studio
f08	20-40	Dublin mainstream	29	spontaneous	school
f09	20-40	Dublin mainstream	29	spontaneous	studio
f10	20-40	Dublin mainstream	32	spontaneous	studio
f11	41-55	Clare	31	speech	studio
f12	41-55	Donegal	32	speech	room
f13	41-55	Dublin mainstream	31	spontaneous	hall
f14	41-55	Dublin mainstream	31	speech	studio
f15	41-55	Dublin mainstream	31	weather	studio
f16	41-55	Dublin south	29	spontaneous	outside
f17	41-55	Dublin south	32	speech	studio
f18	41-55	Dublin mainstream	23	spontaneous	studio
f19	41-55	Dublin mainstream	30	spontaneous	studio
f20	56+	Dublin mainstream	31	spontaneous	studio
f21	56+	Louth	32	spontaneous	studio
f22	56+	Dublin mainstream	33	spontaneous	hall
f23	56+	Dublin mainstream	29	spontaneous	studio
f24	56+	Dublin mainstream	33	spontaneous	studio
f25	56+	Belfast	32	speech	outside

Bibliography

[1] http://edu.surveygizmo.com/s3/1172145/Age-Comparison-Remote-Firefox. Age Comparison Experiment - accessible as of 24/03/2013.

[2] The Up series, 1977-2012. Directed by Michael Apted, produced by Granada Television.

[3] Daubert v Merrell Dow Pharmaceuticals Inc., 1993.

[4] Agnitio Corporation. BATVOX forensic speaker recogntion tool, http://www.agnitio-corp.com/producto.php?id_producto=2, 2013.

[5] C. G. G. Aitken and F. Taroni. *Statistics and the evaluation of evidence for forensic scientists*. John Wiley & Sons, London, 2004.

[6] A. Alexander. *Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions*. PhD thesis, 2005.

[7] ANSI. Methods for calculation of the speech intelligibility index (s3.5 - 1997), 2007.

[8] L. C. Arviso and M. M. Johns III. Challenges and Opportunities in Management of the Aging Voice. Technical report, AAO-HNS (American Academy of Otolaryngology - Head and Neck Surgery) Annual Meeting, 2010.

[9] R. Auckenthaler. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[10] M. H. Bahari, M. McLaren, H. van hamme, and D. A. van Leeuwen. Age estimation from telephone speech using i-vectors. In *InterSpeech 2012*, 2012.

[11] K. Bartkova, D. L. Gac, D. Charlet, and D. Jouvet. Prosodic parameter for speaker identification. In *7th International Conference on Spoken Language Processing*, pages 1197–1200, 2002.

[12] J. M. Beck. Organic Variation of the Vocal Apparatus. In *The Handbook of Phonetic Sciences*, pages 153–201. Blackwell Publishing Ltd., 2010.

[13] M. Ben, R. Blouet, and F. Bimbot. A Monte Carlo method for score normalization in Automatic Speaker Verification using Kullback-Leibler distances. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 689–692, 2002.

[14] S. Bengio and J. Marithoz. A Statistical Significance Test for Person Authentication. In *Odyssey 2006*, pages 279–284, 2006.

[15] B. J. Benjamin. Speech Production of Normally Aging Adults. *Seminars in Speech and Language*, 18:135–141, 1997.

[16] M. S. Benninger and J. Abitbol. Voice: Dysphonia and the aging voice. In *Geriatric Care Otolaryngology*. American Academy of Otolaryngology - Head and Neck Surgery Foundation, 2006.

[17] J. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical report, International Computer Science Institute, 1988.

[18] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, M. Chagnolleau, S. Meignier, T. Merlin, O. Garcia, P. Delacretaz, and Reynolds. A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[19] P. Boersma and D. Weenink. Praat: doing phonetics by computer [Computer program] Version 5.3.49, `http://www.praat.org`, 2013.

[20] F. Botti, A. Alexander, and A. Drygajlo. An interpretation framework for the evaluation of evidence in forensic automatic speaker recognition with limited suspect data. In *Odyssey 2004*, pages 63–68, 2004.

[21] P. Bourdieu. The economics of linguistic exchanges. *Social Science Information*, 16:645–688, 1977.

[22] D. Bowie. *The effect of geographical mobility on the retention of a local dialect.* PhD thesis, 2000. University of Pennsylvania.

[23] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely. Greybeard - Voice and Aging. In *Seventh Conference on International Language Resources and Evaluation (LREC '10)*, 2010.

[24] M. Brückl and W. Sendlmeier. Aging female voices: An acoustic and perceptive analysis. In *VOQUAL '03*, pages 163–168.

[25] N. Brümmer. Spescom DataVoice NIST 2004 system description. In *NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, 2004.

[26] N. Brümmer. Bosaris toolkit: Matlab code for calibrating, fusing and evaluating scores from (automatic) binary classifiers, 2011. available at: `https://sites.google.com/site/bosaristoolkit/home`.

[27] N. Brümmer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2072–2084, 2007.

[28] N. Brümmer and J. du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2?3):230–275, 2006.

[29] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):1979–1986, 2007.

[30] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer. Discriminatively trained probabilistic linear discriminant analysis for speaker verification. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[31] S. Calinon, F. Guenter, and A. Billard. On learning, representing, and generalizing a task in a humanoid robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 37(2):286–298, 2007.

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

[33] J. P. Campbell and A. Higgins. YOHO Speaker Verification, 1994. Linguistic Data Consortium.

[34] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf. Forensic speaker recognition. *IEEE Signal Processing Magazine*, 26(2):95–103, 2009.

[35] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *Signal Processing Letters, IEEE*, 13(5):308–311, 2006.

[36] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In *ICASSP*, pages 97–100, 2006.

[37] C. Champod and E. W. Evett. Commentary on a. p. a. broeders (1999) some observations on the use of probability scales in forensic identification, forensic linguistics 6(2): 22841. *The International Journal of Speech, Language and the Law*, 7(2):239–243, 2000.

[38] C. Champod and D. Meuwly. The inference of identity in forensic speaker recognition. *Speech Communication*, 31(2-3):193–203, 2000.

[39] C. Chang and C. Lin. *LIBSVM: a library for support vector machines*. 2001.

[40] Y. Chang Huai, L. Kong Aik, and L. Haizhou. GMM-SVM Kernel with a Bhattacharyya-based Distance for Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1300–1312, 2010.

[41] I. Cohen. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, 11(5):466–475, 2003.

[42] R. Cole, M. Noel, and V. Noel. The CSLU Speaker Recognition Corpus. In *International Conference on Spoken Language Processing*, pages 3167–3170, 1998.

[43] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[44] S. Cumani, O. Glembek, N. Brummer, E. de Villiers, and P. Laface. Gender independent discriminative speaker recognition in i-vector space. In *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pages 4361–4364, 2012.

[45] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4):357–366, 1980.

[46] N. De Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 4(2):385–390, 2009.

[47] W. Decoster and F. Debruyne. Longitudinal Voice Changes: Facts and Interpretation. *Journal of Voice*, 14(2):184–193, 2000.

[48] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech & Language Processing*, 19(4):788–798, 2011.

[49] G. Dobry, R. M. Hecht, M. Avigal, and Y. Zigel. Supervector dimension reduction for efficient speaker age estimation based on the acoustic speech signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1975–1985, 2011.

[50] G. Doddington. The Effect of Target/Non-Target Age Difference on Speaker Recognition Performance. In *Odyssey 2012*, 2012.

[51] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds. The NIST speaker recognition evaluation: Overview, methodology, systems, results, perspective. *Speech Communication*, 31(2?3):225–254, 2000.

[52] A. Drygajlo. *Automatic Speaker Recogition for Forensic Case Assessment and Interpretation*.

[53] A. Drygajlo. Forensic Automatic Speaker Recognition. *Signal Processing Magazine, IEEE*, 24(2):132–135, 2007.

[54] A. Drygajlo and W. Li. Client-specific a-stack model for adult face verification across aging. *Signal, Image and Video Processing*, pages 1–11, 2011.

[55] A. Drygajlo, W. Li, and H. Qiu. Adult Face Recognition in Score-Age-Quality Classification Space. In C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, and M. Fairhurst, editors, *Biometrics and ID Management*, volume 6583 of *Lecture Notes in Computer Science*, pages 205–216. Springer Berlin / Heidelberg, 2011.

[56] A. Drygajlo, W. Li, and K. Zhu. Q-stack Aging Model for Face Verification. In *EUSIPCO 2009*, Glasgow, Scotland, 2009.

[57] L. W. Ellisen. Smoking and emphysema: the stress connection. *Nature Medicine*, 16(7):754–755, 2010.

[58] W. Endres, W. Bambach, and G. Flsser. Voice Spectrograms as a Function of Age, Voice Disguise, and Voice Imitation. *The Journal of the Acoustical Society of America*, 49(6B):1842–1848, 1971.

[59] A. Eriksson. Tutorial on forensic speech science. part 1. forensic phonetics. In *InterSpeech 2005*, 2005.

[60] I. W. Evett, G. Jackson, J. A. Lambert, and S. McCrossan. The impact of the principles of evidence interpretation on the structure and content of statements. *Science & Justice*, 40(4):233–239, 2000.

[61] K. R. Farrell. Adaptation of Data Fusion-based Speaker Verification Models. In *IEEE International Symposium on Circuits and Systems, ISCAS, 2002*, volume 2, pages II–851–II–854 vol.2, 2002.

[62] J. Fierrez-Aguilar, L. M. Munoz-Serrano, F. Alonso-Fernandez, and J. Ortega-Garcia. On the effects of image quality degradation on minutiae- and ridge-based automatic fingerprint recognition. In *IEEE International Carnahan Conference on Security Technology*, pages 79–82, 2005.

[63] J. P. F. French, P. Harrison, and J. Windsor-Lewis. R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial. *The International Journal of Speech, Language and the Law*, 13(2):256–273, 2006.

[64] M. Frigge, D. C. Hoaglin, and B. Iglewicz. Some Implementations of the Boxplot. *The American Statistician*, 43(1):50–54, 1989.

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

[66] J. Galbally, J. Fierrez, and J. Ortega-Garcia. Bayesian hill-climbing attack and its application to signature verification. In *Advances in Biometrics*, volume 4642 of *Lecture Notes in Computer Science*, chapter 41, pages 386–395. Springer Berlin Heidelberg, 2007.

[67] D. Garcia-Romero and C. Y. Espy-Wilson. Analysis of i-vector length normalization in speaker recognition systems. In *InterSpeech 2011*, 2011.

[68] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia. Using Quality Measures for Multilevel Speaker Recognition. *Computer Speech & Language*, 20(2-3):192–209, 2005.

[69] J. S. Garofolo. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993. Linguistic Data Consortium.

[70] O. Glembek, L. Burget, N. Dehak, N. Brümmer, and P. Kenny. Comparison of scoring methods used in speaker recognition with joint factor analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4057–4060, 2009.

[71] K. W. Godin and J. H. Hansen. Session variability contrasts in the marp corpus. In *InterSpeech 2010*, pages 298–301, Makuhari, Japan, 2010.

[72] E. Gold and P. French. International practices in forensic speaker comparison. *The International Journal of Speech, Language and the Law*, 18(2):293–307, 2011.

[73] E. Gold and V. Hughes. Issues and opportunities for the application of the numerical likelihood ratio framework to forensic speaker comparison. In *IAFPA 2012*, 2012.

[74] J. Gonzalez-Rodriguez, A. Drygajlo, D. Ramos-Castro, M. Garcia-Gomar, and J. Ortega-Garcia. Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. *Computer Speech & Language*, 20(2-3):331–355, 2006.

[75] J. Gonzalez-Rodriguez and D. Ramos. *Forensic Automatic Speaker Classification in the "Coming Paradigm Shift"*, volume 4343 of *Lecture Notes in Computer Science*, chapter 11, pages 205–217. Springer Berlin Heidelberg, 2007.

[76] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia. Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(7):2104–2115, 2007.

[77] C. Greenberg. The nist year 2012 speaker recognition evaluation plan. Technical report, 2012. Available online: `http://www.nist.gov/itl/iad/mig/upload/NIST_SRE12_evalplan-v17-r1.pdf`.

[78] C. Greenberg, V. Stanford, A. Martin, M. Yadagiri, G. Doddington, J. Godfrey, and J. Hernandez-Cordero. The 2012 nist speaker recognition evaluation. In *InterSpeech 2013*, Portland, Oregon.

[79] P. Grother and E. Tabassi. Performance of Biometric Quality Measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.

[80] J. D. Harnsberger, W. S. Brown Jr, R. Shrivastav, and H. Rothman. Noise and Tremor in the Perception of Vocal Aging in Males. *Journal of Voice*, 24(5):523–530, 2010.

[81] J. D. Harnsberger, R. Shrivastav, and W. Brown. Modeling Perceived Vocal Age in American English. In *InterSpeech 2010*, Makuhari, Japan, 2010.

[82] A. Harriero, D. Ramos, J. Gonzalez-Rodriguez, and J. Fierrez. Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification. In *Third International Conference on Advances in Biometrics*, pages 434–442. Springer-Verlag, 2009.

[83] J. Harrington. An acoustic analysis of 'happy-tensing' in the Queen's annual Christmas broadcasts. *Journal of Phonetics*, 34:439–457, 2006.

[84] J. Harrington, S. Palethorpe, and C. Watson. Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. In *InterSpeech 2007*, Antwerp, Belgium, 2007.

[85] J. Harrington, S. Palethorpe, and C. I. Watson. Does the Queen speak the Queen's English? *Nature*, 408(6815):927–928, 2000.

[86] T. Hasan and J. H. Hansen. A Study on Universal Background Model Training in Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1890–1899, 2011.

[87] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. v. Leeuwen. Duration mismatch compensation for i-vector based speaker recognition systems. In *ICASSP 2013*, 2013.

[88] A. O. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for svm-based speaker recognition. In *Proc. of ICSLP*, 2006.

[89] V. Hautamäki, M. Tuononen, T. Niemi-Laitinen, and P. Fränti. Improving speaker verification by periodicity based voice activity detection. In *12th Internat. Conf. on Speech and Computer (SPECOM 2007)*, pages 645–650, 2007.

[90] H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *The Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

[91] H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.

[92] H. Hollien. Aural-perceptual speaker identification: problems with noncontemporary samples. *The International Journal of Speech, Language and the Law*, 7(2), 2000.

[93] H. Hollien. On earwitness lineups. *Investigative Sciences Journal*, 4(1), 2012.

[94] H. Hollien and R. Schwartz. Speaker Identification Utilizing Noncontemporary Speech. *Journal of Forensic Sciences*, 46(1):63–67, 2001.

[95] J. Holmes and W. Holmes. *Speech Synthesis and Recognition*. Taylor & Francis, 2nd edition, 2001.

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

[97] V. Hughes and P. Foulkes. Effects of variation on the computation of numerical likelihood ratios for forensic voice comparison. In *IAFPA 2012*, 2012.

[98] A. Hur and J. Weston. A users guide to support vector machines. pages 1–18, 2010.

[99] S. Ishihara and Y. Kinoshita. How many do we need? exploration of the population size effect on the performance of forensic speaker classification. In *InterSpeech 2008*, pages 1941–1944, 2008.

[100] E. Jacewicz, R. A. Fox, and C. O'Neill. Articulation rate across dialect, age, and gender. *Language Variation and Change*, 21(2):233–256, 2009.

[101] G. Jackson, S. Jones, G. Booth, C. Champod, and E. W. Evett. The nature of forensic science opinion – a possible framework to guide thinking and practice in investigations and in court proceedings. *Science & Justice*, 46(1):33–44, 2006.

[102] A. K. Jain, R. P. W. Duin, and M. Jianchang. Statistical pattern recognition: a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(1):4–37, 2000.

[103] J.P. French Associates. Forensic speech and acoustics laboratory, `http://www.jpfrench.com/`, 2013.

[104] F. Juefei-Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen. Investigating age invariant face recognition based on periocular biometrics. In *International Joint Conference on Biometrics (IJCB), 2011*, pages 1–7, 2011.

[105] J. Kahane. Anatomic and physiologic changes in the aging peripheral speech mechanism. *Aging-Communication Processes and Disorders*, 1(1):21–45, 1981.

[106] F. Kelly, N. Brümmer, and N. Harte. Eigenageing Compensation for Speaker Verification. In *InterSpeech 2013*, Lyon, France, 2013.

[107] F. Kelly, A. Drygajlo, and N. Harte. Compensating for Ageing and Quality variation in Speaker Verification. In *InterSpeech 2012*, Portland, Oregon, 2012.

[108] F. Kelly, A. Drygajlo, and N. Harte. Speaker Verification with Long-Term Ageing Data. In *International Conference on Biometrics (ICB) 2012*, New Delhi, India, 2012.

[109] F. Kelly, A. Drygajlo, and N. Harte. Speaker verification in score-ageing-quality classification space. *Computer Speech & Language*, 27(5):1068–1084, 2013.

[110] F. Kelly and N. Harte. Effects of Long-Term Ageing on Speaker Verification. In C. Vielhauer, J. Dittmann, A. Drygajlo, N. Juul, and M. Fairhurst, editors, *Biometrics and ID Management*, volume 6583 of *Lecture Notes in Computer Science*, pages 113–124. Springer Berlin / Heidelberg, 2011.

[111] F. Kelly and N. Harte. Auditory detectability of vocal ageing and its effect on forensic automatic speaker recognition. In *InterSpeech 2013*, Lyon, France, 2013.

[112] P. Kenny. Joint factor analysis of speaker and session variability: theory and algorithms - technical report. Technical report, CRIM-06/08-13, 2006.

[113] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, 2007.

[114] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Speaker and session variability in gmm-based speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1448–1460, 2007.

[115] P. Kenny and P. Dumouchel. Disentangling speaker and channel effects in speaker verification. In *ICASSP*, pages 37–40, 2004.

[116] P. Kenny, M. Mihoubi, and P. Dumouchel. New MAP estimators for speaker recognition. In *Eurospeech*, 2003.

[117] R. D. Kent. Models of speech motor control: Implications from recent developments in neuropsychological and neurobehavioral science. In B. Maassen, R. Kent, H. Peters, P. v. Lieshout, and W. Hulstijn, editors, *Speech motor control in normal and disordered speech*. New York: Oxford University Press, 2004.

[118] T. Kinnunen and H. Li. An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. *Speech Communication*, 52(1):12–40, 2010.

[119] K. Kryszczuk and A. Drygajlo. Improving biometric verification with class-independent quality information. *Signal Processing, IET*, 3(4):310–321, 2009.

[120] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Error handling in multimodal biometric systems using reliability measures. In *12th European Conference on Signal Processing*, 2005.

[121] K. Kryszczuk, J. Richiardi, P. Prodanov, and A. Drygajlo. Reliability-based Decision Fusion in Multimodal Biometric Verification Systems. *EURASIP Journal of Applied Signal Processing*, 2007(1):74–74, 2007.

[122] H. J. Künzel. Non-contemporary speech samples: Auditory detectability of an 11 year delay and its effect on automatic speaker identification. *The International Journal of Speech, Language and the Law*, 14(1):109–136, 2007.

[123] A. Lanitis. A Survey of the Effects of Aging on Biometric Identity Verification. *International Journal of Biometrics*, 2(1):34–52, 2010.

[124] A. D. Lawson, A. R. Stauffer, E. J. Cupples, S. J. Wenndt, W. P. Bray, and G. J. J. The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial findings. In *InterSpeech 2009*, Brighton, U.K., 2009.

[125] A. D. Lawson, A. R. Stauffer, B. Y. Smolenski, B. B. Pokines, M. Leonard, and E. J. Cupples. Long-term Examination of Intra-session and Inter-session Speaker Variability. In *InterSpeech 2009*, Brighton, United Kingdom, 2009.

[126] D. A. v. Leeuwen. A note on performance metrics for speaker recognition using multiple conditions in an evaluation. Technical report, 9 June 2008.

[127] D. A. v. Leeuwen and N. Brümmer. An introduction to application-independent evaluation of speaker recognition systems. *Speaker Classification*, I:330–353, 2007.

[128] W. Li and A. Drygajlo. Global and local feature based multi-classifier a-stack model for aging face identification. In *17th IEEE International Conference on Image Processing (ICIP)*, pages 3797–3800, 2010.

[129] W. Li, A. Drygajlo, and H. Qiu. Aging Face Verification in Score-Age Space using Single Reference Image Template. In *IEEE International Conference on Biometrics: Theory, Applications And Systems (BTAS)*, 2010.

[130] S. E. Linville. Vocal Aging. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 3(3):183–187, 1995.

[131] S. E. Linville. The Sound of Senescence. *Journal of Voice*, 10(2):190–200, 1996.

[132] S. E. Linville. *Vocal Aging*. Canada: Singular, 2001.

[133] S. E. Linville. Voice Disorders of Aging. In R. D. Kent, editor, *The MIT Encyclopedia of Communication Disorders*. MIT press, 2004.

[134] N. Macmillan and D. Creelman. *Detection Theory: A User's Guide*. Lawrence Erlbaum, 2004.

[135] L. Malfait, J. Berger, and M. Kastner. P.563 - the ITU-T Standard for Single-Ended Speech Quality Assessment. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1924–1934, 2006.

[136] M. Mandasari, R. Saeidi, M. McLaren, and D. van Leeuwen. Quality measure functions for calibration of speaker recognition system in various duration conditions. *Audio, Speech, and Language Processing, IEEE Transactions on*, PP(99):1–1, 2013.

[137] J. Mason and J. Thompson. Gender effects in speaker recognition. In *ICSP*, pages 733–736, 1993.

[138] F. McGehee. The reliability of the identification of the human voice. *Journal of General Psychology*, 17:249–271, 1937.

[139] M. McLaren and D. A. van Leeuwen. A simple and effective speech activity detection algorithm for telephone and microphone speech. In *NIST 2011 SRE Workshop*, 2011.

[140] D. Meuwly. Forensic speaker recognition: An evidence odyssey, summary. In *Odyssey 2004*, pages 11–12, 2004.

[141] D. Meuwly and A. Drygajlo. Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modelling (GMM). In *Odyssey 2001*, 2001.

[142] D. Meuwly, M. El-Maliki, and A. Drygajlo. Forensic Speaker Recognition Using Gaussian Mixture Models and a Bayesian Framework. In *COST-250 workshop on speaker recognition by man and machine: directions for forensic applications*, 1998.

[143] V. A. Moreno and A. Drygajlo. A joint factor analysis model for handling mismatched recording conditions in forensic automatic speaker recognition. In *5th IAPR International Conference on Biometrics (ICB) 2012*, pages 484–489, 2012.

[144] P. B. Mueller. The Aging Voice. *Seminars in Speech and Language*, 18(2):159–169, 1997.

[145] NIST. Speech quality assurance (spqa) package, version 2.3, 2000.

[146] F. Nolan. *The phonetic bases of speaker recognition.* Cambridge studies in speech science and communication. Cambridge University Press, Cambridge [Cambridgeshire] ; New York :, 1983.

[147] L. Öhman. *All Ears: Adults and Childrens Earwitness Testimony.* PhD thesis, 2013.

[148] R. F. Orlikoff. The Relationship of Age and Cardiovascular Health to Certain Acoustic Characteristics of Male Voices. *Journal of Speech and Hearing Research*, 33:450–457, 1990.

[149] G. Papcun. Long-term memory for unfamiliar voices. *Journal of the Acoustical Society of America*, 85(2):913–925, 1989.

[150] U. Park, Y. Tong, and A. K. Jain. Age-Invariant Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):947–954, 2010.

[151] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Odyssey 2001*, 2001.

[152] N. Poh and J. Kittler. A unified framework for multimodal biometric fusion incorporating quality measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):3–18, 2012.

[153] Presidential Speech Archive. Miller Center, University of Virginia, `http://www.millercenter.org/scripps/archive/speeches`, 2012.

[154] S. J. D. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *IEEE 11th International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[155] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements. *Objective measures of speech quality.* Prentice Hall signal processing series. Prentice Hall, 1988.

[156] L. R. Rabiner and B. H. Huang. *Fundamentals of Speech Recognition.* Prentice Hall, 1993.

[157] L. A. Ramig. Effects of physiological aging on speaking and reading rates. *Journal of Communication Disorders*, 16(3):217–226, 1983.

[158] D. Ramos and J. Gonzalez-Dominguez. Reliable support: Measuring calibration of likelihood ratios. *Forensic Science International*, (230), 2013.

[159] A. P. Rebera and B. Guihen. Biometrics for an ageing society: societal and ethical factors in biometrics and ageing. In *Proceedings of the International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–4, 2013.

[160] U. Reubold, J. Harrington, and F. Kleber. Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. *Speech Communication*, 52(7-8):638–651, 2010.

[161] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, J. Qin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and X. Bing. The supersid project: exploiting high-level information for high-accuracy speaker recognition. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 4, pages 784–787, 2003.

[162] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *EUROSPEECH 1997*, pages 963–966, 1997.

[163] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[164] R. Rhodes. Changes in the Voice across the Early Adult Lifespan. In *The International Association of Forensic Phonetics and Acoustics (IAFPA) 2011*, 2011.

[165] R. Rhodes. *Assessing the strength of non-contemporaneous forensic speech evidence*. PhD thesis, 2012. The University of York.

[166] R. Rhodes. Changes in the voice across the adult lifespan: formant frequency-based likelihood ratios and ASR performance. In *The International Association of Forensic Phonetics and Acoustics (IAFPA) 2013*, 2013.

[167] J. Richiardi and A. Drygajlo. *Probabilistic models for multi-classifier biometric authentication using quality measures*. PhD thesis, 2007.

[168] J. Richiardi and A. Drygajlo. Evaluation of Speech Quality Measures for the purpose of Speaker Verification. In *Odyssey 2008: The Speaker and Language Recognition Workshop*, 2008.

[169] J. Richiardi, K. Kryszczuk, and A. Drygajlo. Quality Measures in Unimodal and Multimodal Biometric Verification. In *EUSIPCO 2007*, 2007.

[170] J. Richiardi, P. Prodanov, and A. Drygajlo. Speaker Verification with Confidence and Reliability Measures. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2006*, 2006.

[171] P. Rose. Long- and short-term within-speaker differences in the formants of Australian *hello*. *Journal of the International Phonetic Association*, 29:1–31, 1999.

[172] P. Rose. *Forensic Speaker Identification*. Taylor & Francis, 2002.

[173] P. Rose. Technical forensic speaker recognition: Evaluation, types and testing of evidence. *Computer Speech & Language*, 20(2/3):159–191, 2006.

[174] P. Rose and G. S. Morrison. A response to the UK position statement on forensic speaker comparison. *The International Journal of Speech, Language and the Law*, 16(1):139–163, 2009.

[175] A. E. Rosenberg and S. Parthasarathy. Speaker Background Models for Connected Digit Password Speaker Verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1996*, 1996.

[176] Rosistem. Tutorial on Voice Recognition, `http://www.barcode.ro/tutorials/biometrics/img/speech-production.jpg`, 2003.

[177] R. Saedi and D. van Leeuwen. The radboud university nijmegen submission to nist sre 2012. In *NIST Speaker Recognition Evaluation Workshop*, 2012.

[178] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P.-M. Bousquet, E. Khoury, P. L. S. Martinez, J. M. K. Kua, C. You, H. Sun, A. Larcher, P. Rajan, V. Hautamki, C. Hanilci, B. Braithwaite, G.-H. Rosa, S. O. Sadjadi, G. Liu, H. Boril, N. Shokouhi, D. Matrouf, L. El Shafey, P. Mowlaee, J. Epps, T. Thiruvaran, D. Van Leeuwen, B. Ma, H. Li, J.-F. Bonastre, S. Marcel, J. Mason, and E. Ambikairajah. I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification. In *InterSpeech*, 2013.

[179] G. Sankoff. Adolescents, young adults and the critical period: two case studies from "Seven Up". In C. Fought, editor, *Sociolinguistic variation, critical reflections*. Oxford University Press, 2004.

[180] R. T. Sataloff, D. Caputo Rosen, M. Hawkshaw, and J. R. Spiegel. The aging adult voice. *Journal of Voice*, 11(2):156–160, 1997.

[181] S. Schötz. *Perception, Analysis and Synthesis of Speaker Age*. PhD thesis, 2006. Lund University, Sweden.

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

[183] E. E. Shriberg. *Higher-Level Features in Speaker Recognition*, volume 4343, pages 241–259. Springer: Heidelberg / Berlin / New York, 2007.

[184] F. Soong, A. Rosenberg, L. Rabiner, and B. Juang. A vector quantization approach to speaker recognition. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) '85*, volume 10, pages 387–390, 1985.

[185] E. T. Stathopoulos, J. E. Huber, and J. E. Sussman. Changes in Acoustic Characteristics of the Voice across the Life Span: Measures from Individuals 4-93 Years of Age. *Journal of Speech, Language, and Hearing Research*, 54:1011–1021, 2011.

[186] H. G. Sung. *Gaussian Mixture Regression and Classification*. PhD thesis, 2004.

[187] The Oxford Institute of Population Ageing. Research overview, `http://www.ageing.ox.ac.uk/research`, 2012.

[188] K. M. Ting and I. H. Witten. Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*, 10:271–289, 1999.

[189] P. Torre III and J. A. Barlow. Age-related changes in acoustic characteristics of adult speech. *Journal of Communication Disorders*, 42(5):324–333, 2009.

[190] H. Ulatowska. *The Aging Brain: Communication in the Elderly*. San Diego: College-Hill Press, 1985.

[191] U. Uludag and A. K. Jain. Attacks on biometric systems: A case study in fingerprints. In *Proc. SPIE-EI 2004, Security, Seganography and Watermarking of Multimedia Contents VI*, pages 622–633, 2004.

[192] K. Verdolini. Voice Therapy for Neurological Aging-Related Voice Disorders. In R. D. Kent, editor, *The MIT Encyclopedia Of Communication Disorders*. MIT press, 2004.

[193] I. M. Verdonck-de Leeuw and H. F. Mahieu. Vocal Aging and the Impact on Daily Life: A Longitudinal Study. *Journal of Voice*, 18(2):193–202, 2004.

[194] R. Vipperla, S. Renals, and J. Frankel. Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010, 2010.

[195] D. v. d. Vloed, D. Meuwly, R. Haraksim, and J. Vermeulen. Influence of the size of the population dataset on the results produced by the batvox software. In *IAFPA 2011*, 2011.

[196] R. J. Vogt, B. J. Baker, and S. Sridharan. Modelling session variability in text independent speaker verification. In *InterSpeech*, 2005.

[197] V. Wan and S. Renals. Speaker verification using sequence discriminant support vector machines. *IEEE Transactions on Speech and Audio Processing*, 13(2):203–210, 2005.

[198] B. T. Weinert and P. S. Timiras. Invited Review: Theories of aging. *Journal of Applied Physiology*, 95(4):1706–1716, 2003.

[199] J. J. Wolf. Efficient acoustic parameters for speaker recognition. *The Journal of the Acoustical Society of America*, 51(6B):2044–2056, 1972.

[200] D. H. Wolpert. Stacked Generalization. *Neural Networks*, 5(2):241–259, 1992.

[201] S. A. Xue and D. Deliyski. Effects of Aging on Selected Acoustic Voice Parameters: Preliminary Normative Data and Educational Implications. *Educational Gerontology*, 27(2):159–168, 2001.

[202] D. Yuan, L. Liang, Z. Xian-Yu, and Z. Jian. Studies on Model Distance Normalization Approach in Text-independent Speaker Verification. *ACTA AUTOMATICA SINICA*, 35(5), 2009.