Speaker Verification in Score-Ageing-Quality Classification Space

Finnian Kelly[a,1], Andrzej Drygajlo[b] and Naomi Harte[a]

[a] *Department of Electronic and Electrical Engineering,Trinity College Dublin, Ireland*
[b] *Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland*

Abstract

A challenge in automatic speaker verification is to create a system that is robust to the effects of vocal ageing. To observe the ageing effect, a speaker's voice must be analysed over a period of time, over which, variation in the quality of the voice samples is likely to be encountered. Thus, in dealing with the ageing problem, the related issue of quality must also be addressed. We present a solution to speaker verification across ageing by using a stacked classifier framework to combine ageing and quality information with the scores of a baseline classifier. In tandem, the Trinity College Dublin Speaker Ageing database of 18 speakers, each covering a 30-60 year time range, is presented. An evaluation of a baseline Gaussian Mixture Model-Universal Background Model (GMM-UBM) system using this database demonstrates a progressive degradation in genuine speaker verification scores as ageing progresses. Consequently, applying a conventional threshold, determined using scores at the time of enrolment, results in poor long-term performance. The influence of quality on verification scores is investigated via a number of quality measures. Alongside established signal-based measures, a new model-based measure, Wnorm, is proposed, and its utility is demonstrated on the CSLU database. Combining ageing information with quality measures and the scores from the GMM-UBM system, a verification decision boundary is created in score-ageing-quality space. The best performance is achieved by using scores and ageing in conjunction with the new Wnorm quality measure, reducing verification error by 45% relative to the baseline. This work represents the first comprehensive analysis of speaker verification on a longitudinal speaker database and successfully addresses the associated variability from ageing and quality arte-

[1] Corresponding author. Tel: +353 1 896 1580; Email: kellyfp@tcd.ie

facts.

Keywords: speaker verification, ageing, quality measures

1. Introduction

Achieving high accuracy speaker verification in real-world operating conditions is limited by variability between enrolment and verification sessions (Kinnunen and Li, 2010). The voice is subject to many sources of natural variability, from physiological sources like illness or ageing, to cognitive sources such as emotional state or conversational dynamics. When the voice undergoes recording and transmission, further variability, due to background noise or the transmission channel, is introduced. The effect of ageing has received little research attention compared to other sources of inter-session variability.

Vocal ageing is a complex process that leads to change in the pitch and timbre of the voice, and the rate and intensity of speech. Progressive vocal change occurs over the entire lifespan (Stathopoulos et al., 2011), with greatest variability in young (≤ 18) and old (≥ 60) speakers. As a combination of vocal characteristics are used to describe a speaker in a verification system, any long-term application of speaker verification must compensate for the ageing process.

In a standard speaker verification scenario, a user is first enrolled and then attempts to access the system via verification at a later date. As the time lapse between enrolment and verification increases, so does the influence of ageing on the verification decision. With the growth in the reach and scale of biometric systems, this is becoming a more important issue (Lanitis, 2010; Kinnunen and Li, 2010).

In the field of forensic speaker identification (Rose, 2002), a large time lapse between enrolment and identification sessions frequently occurs (e.g. In ‘The Yorkshire Ripper Hoaxer Trial’ (French et al., 2006), there was a gap of 27 years between the ‘scene of the crime’ and the custody recordings). In such cases, ageing undoubtedly must be accounted for. We note that while speaker verification is not suitable for the forensic scenario (where the aim is to quantify the strength of evidence, not to make a decision), the underlying speech features and statistical models are shared across both approaches. Thus a discussion regarding ageing and speaker verification is of direct relevance to forensic speaker identification.

An existing approach for dealing with ageing is model adaptation (Farrell, 2002). This however does not solve the problem, as it relies on regular updates for all users in the database. In a large system, this would be difficult logistically, as well as introducing a potential security weakness. It is also of no use in the forensic domain. A more favorable approach is a system that automatically adapts in response to ageing.

There have been few studies on ageing and speaker verification to date, presumably due to a scarcity of suitable data. Those studies that have been done use cross-sectional data in their experiments (Doddington, 2012; Hansen and Lei, 2009). There are numerous studies on the effect of ageing in other speech technology applications. These again use cross-sectional data, typically separate groups of young, middle-aged and old speakers (Harnsberger et al., 2008; Schötz, 2006; Vipperla et al., 2010; Dobry et al., 2011).

Compensating for ageing in face verification however, has received significantly more attention. Dominant approaches are modelling of ageing changes (Park et al., 2010; Junyan et al., 2006), ageing-dependent verification thresholds (Li et al., 2010) and more recently, age-invariant features (Juefei-Xu et al., 2011). In this domain there are several large, publicly available databases enabling these research efforts, e.g. FG-NET (FG-NET Aging Database, 2010) and MORPH (Ricanek and Tesafaye, 2006). For speaker verification purposes, there are currently no publicly available long-term databases.

In our recent work (Kelly et al., 2012), the Trinity College Dublin Speaker Ageing (TCDSA) database, containing 18 speakers with recordings spanning 30-60 years per speaker, was presented. A speaker verification evaluation of the database demonstrated a progressive degradation in verification score with ageing. Considering the various approaches to dealing with ageing in the face verification domain, an ageing-dependent decision boundary was deemed most applicable to the speaker ageing problem. This approach was implemented by combining the verification score and the time lapse (in years) between training and testing samples, and carrying out verification in a 2-dimensional score-ageing space. A large improvement in verification error rates was achieved.

Compared to other forms of variability, ageing presents unique difficulties. Even in ideal operating conditions, ageing-related change cannot be controlled. Investigation of ageing effects and the development of strategies to compensate for them requires data acquired over very long time intervals. By its nature, such data will contain variability from numerous sources, not ageing alone. Controlling a dataset for ageing is thus not possible, and any study on long-term data must consider the influence of non-ageing-related variability. Another difficulty of ageing influence is its non-uniform nature across a population. The rate and types of changes experienced are influenced by many factors and differ between individuals (Lanitis, 2010). Any modelling of ageing behaviour for one subset of the population may not be directly applicable to another.

In this paper we extend the approach in (Kelly et al., 2012) by incorporating objective measures of recording quality. Firstly, this allows us to reduce non-ageing-related variability and further isolate the ageing effect. Secondly, the reduction in this variability improves the accuracy of the long-term ageing-dependent decision boundary, and results in a drop in overall verification error rates. Along with classical quality measures, we propose and evaluate a new model-based quality measure. We demonstrate the effectiveness of this measure at predicting score degradation due to quality variation, and use it to improve the performance of our long-term ageing system, first presented in Kelly et al. (2012).

In Section 2, the process of vocal ageing is described in more detail. The TCDSA database is detailed in full in Section 3. In Section 4 the Gaussian Mixture Model - Universal Background Model (GMM-UBM) speaker verification system is introduced. The influence of ageing on the performance of the speaker verification system is shown experimentally in Section 5. In Section 6 classical and novel measures of objective recording quality are presented. In Section 7 we introduce a stacked classifier method of combining score, ageing and quality information for classification and present an evaluation of the proposed technique. Finally, conclusions are presented in Section 8.

2. The Ageing Voice

The physiological changes to the vocal system with ageing, and the resulting acoustic changes in the voice have received significant research attention. Physiological studies (Linville, 1995, 2004; Mueller, 1997) have identified changes in all parts of the adult vocal system with ageing: The subglottal system (lungs, diaphragm, trachea) is affected by a stiffening of the thorax and a decreased rate and strength of respiratory muscle contraction. In the larynx, vocal folds thin in males and become thicker in females. The supralaryngeal (oral and nasal cavities) system's primary changes are decreased tongue and facial muscle functionality.

The effect of these physiological changes is manifested in the voice in numerous ways (Linville, 2004; Stathopoulos et al., 2011; Reubold et al., 2010; Rhodes, 2011): A downward shift in fundamental frequency of speech occurs throughout adulthood for both males and females. Around the age of 60 an upward shift occurs in males, continuing into advanced old age. Common for males and females is a change in timbre or tone quality, along with an increase in instability of both pitch and intensity of speech. A decrease in the rate of speech is typically observed. Other changes include increased breathiness and decreased articulatory precision.

Although the latter vocal attributes are typically most noticeable in an elderly voice, progressive changes occur throughout the lifespan (Stathopoulos et al., 2011; Rhodes, 2011). It should be noted that the extent of ageing-related change is variable (Mueller, 1997). This is particularly noticeable in elderly speakers, and is compounded by the increased likelihood of elderly speakers to experience pathological vocal change due to illness, effects of alcohol and tobacco use, or cognitive decline (Mueller, 1997; Benjamin, 1997).

3. Speaker Ageing Data

Obtaining usable data is a major challenge faced when approaching the study of speaker ageing and verification. There are currently no publicly available long-term longitudinal speaker databases that we are aware of. There are a limited number of short-term longitudinal databases available: the CSLU and MARP databases (Cole et al., 1998; Lawson et al., 2009a) contain longitudinal data over a 2-3 year period. The Greybeard Corpus (Brandschain et al., 2010), compiled by NIST for their Speaker Recognition Evaluation (SRE) 2010, contains longitudinal data over a 2-12 year period. The Greybeard Corpus was made available to SRE 2010 participants only however, and the performance results of the SRE systems on this dataset were not released.

An ideal database would contain data over several decades, with regular recordings for each speaker in controlled conditions across the entire time span (Lanitis, 2010). 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 in unrealistic. In this work, we use the Trinity College Dublin Speaker Ageing (TCDSA) database (first presented in (Kelly et al., 2012)), which was constructed with the aim of limiting non-ageing-related variability.

3.1. TCDSA Database

A schematic of the TCDSA database content is given in Figure 1. The database contains 18 speakers (9 male, 9 female). As is evident, the quantity and range of the speech samples varies between speakers. There are between 2 and 35 recordings for each speaker spanning a range of between 30 and 60 years. The recordings vary between 1 and 30 minutes in length approximately. The database contains 31 hours of content in total.

The data originates from the national broadcasters of the U.K. and Ireland, the BBC (British Broadcasting Corporation) and RTÉ (Raidió Teilifís Éireann), as well as publicly available samples, obtained from YouTube and The Miller Center (Presidential Speech Archive, 2012). 10 of the 18 speakers were sourced from the RTÉ archives, 5 from the BBC archives, 1 from the Miller Center and 2 from YouTube. The data of several of the BBC speakers has been supplemented with additional data from YouTube. The content of the recordings is primarily interviews and speeches, thus there is a mix of conversational and read speech. There is a range of Irish, English and American accented speakers in the database (All recordings are in the English language).

As expected for an uncontrolled database, the recording conditions are varied. Even in the case of the RTÉ and BBC broadcast data, the specifics of the recording setup and equipment are unavailable. To limit non-ageing-related variability, the initial 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.

As an objective measure of quality, the likelihood of the recordings given an age-balanced Universal Background Model (UBM) (Reynolds et al., 2000) was calculated (the construction of this UBM is detailed in Section 4.2). This approach has previously been applied for quality measurement (Harriero et al., 2009). The motivation for this measure can be understood by viewing the UBM as a representation of the distribution of speaker features common to all of the test population. In the case of a degraded test recording, its likelihood given the UBM will fall outside the range of scores from typical ‘clean’ recordings. The notion of quality is addressed in depth in Section 6.

The UBM quality screening was applied to the database material by dividing every recording into 30 second segments, extracting features (as described in 4.1) 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 (Frigge et al., 1989). The assumption was that the set of all ‘acceptable’ UBM scores would fall within the limits of this range. Considering that approximately 10% of the total quantity of data was discarded with this method, these limits were considered reasonable. The ‘cleaned-up’ database is depicted in Figure 1.

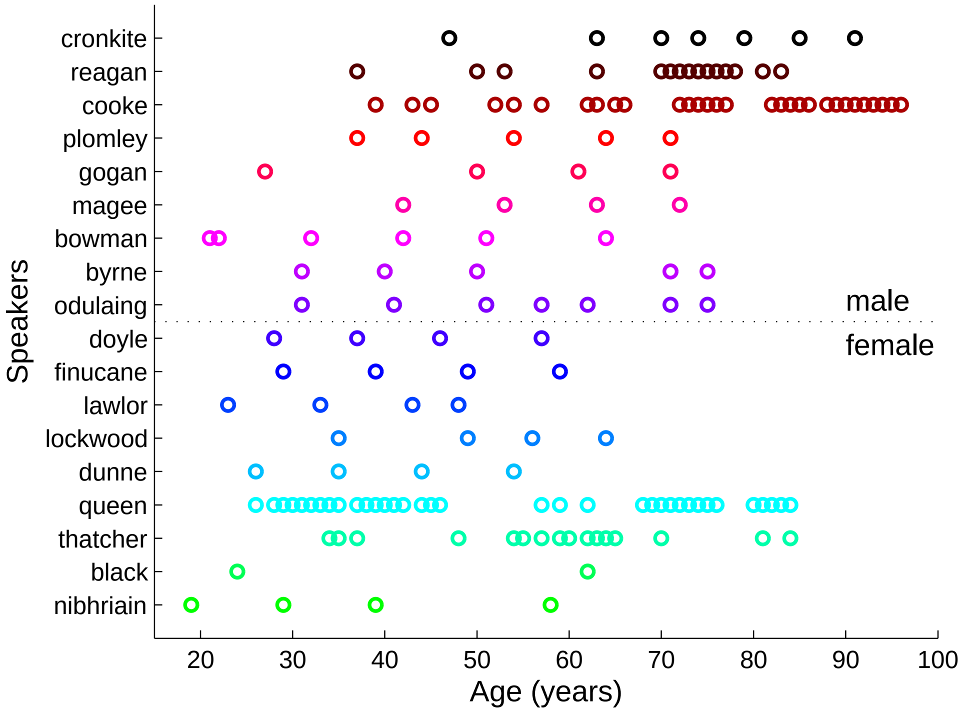


Figure 1: A schematic of the TCDSA database. A circle indicates a year where a recording is available for a speaker, with the speaker surnames given on the y-axis and their age on the x-axis. The speakers are divided by gender, with males and females on the upper and lower halves of the plot respectively.

3.2. Ageing UBM Database

An additional database was compiled to create a suitable UBM for use with the TCDSA database. The University of Florida Vocal Aging Database (UF-VAD) (extemporaneous set) (Harnsberger et al., 2010) was combined with data sourced from Youtube and the Miller Center (Presidential Speech Archive, 2012). The resulting dataset contains 1 hour of speech from 120 speakers (30 seconds each). The speakers are evenly balanced across gender, age and accent, and a variety of recording conditions are present. The ages of the speakers are spread evenly across three age profiles: 35, 36-55 and over 55. English, Irish and American accented speakers are included. The dataset was composed in this way in an effort to ensure it was a close representation of the population of the TCDSA database.

4. GMM-UBM Speaker Verification

The speaker verification system used for the experimental evaluations in this paper was a Gaussian Mixture Model - Universal Background Model (GMM-UBM) system (Reynolds et al., 2000). Recent developments in speaker verification which build on the GMM-UBM approach, such as Factor Analysis and i-vectors (Kinnunen and Li, 2010; McLaren et al., 2012), have proven very effective. These methods require extensive development data (McLaren et al. (2012); Dehak et al. (2011) for example, use multiple development databases stretching into 100s of hours of data). While ideally a state-of-the-art system would be employed for this study, we believe the question over how to identify and use additional data sources, given the ageing database content, speaker accents and age distribution, is not easily answered. The function of this paper is to investigate ageing rather than be verification-performance-driven. Hence an immediate extension of GMM-UBM (to Factor Analysis for example) would introduce additional sources of variability and make it harder to isolate ageing. Given the lack of published work in this domain, an analysis of longitudinal speaker verification using the GMM-UBM system will inform the implementation of more complex systems. Our work to extend the system in this paper is ongoing.

4.1. Pre-processing and Feature Extraction

All recordings were first downsampled to 16 kHz. Energy-based silence removal and pre-emphasis were then applied. 12-dimensional MFCC vectors were extracted over 20 ms windows with 50% overlap using 26 mel filters. RASTA filtering (Hermansky and Morgan, 1994) 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. These preprocessing steps and feature extraction parameters are typical in current speaker verification applications (Kinnunen and Li, 2010).

4.2. GMM-UBM system

In the well-established GMM-UBM (Reynolds et al., 2000) framework, a GMM with a large number of components representing a speaker-independent pool of features - a UBM - is created first. To achieve a speaker-independent representation in the UBM, a large amount of data from a balanced range of speakers is required (Hasan and Hansen, 2011; Rosenberg and Parthasarathy, 1996). As described in Section 3.2, our UBM development database is balanced across sub-populations within the data. To ensure a fully gender-balanced UBM, male and female UBMs of 512 components were trained on a male-female split of the UBM database and concatenated to create a 1024 component (diagonal covariance) UBM. Training consisted of estimating maximum likelihood model parameters with the iterative Expectation-Maximisation (EM) (Bilmes, 1988) algorithm. Speaker-dependent GMMs were then trained via adaptation of the UBM given a set of training features for each speaker (Reynolds et al., 2000). Adapting the speaker models from the UBM ensures that 'gaps' in training data for a speaker are filled by speaker-independent information in the UBM. This improves verification performance, particularly in the case of limited training data. Only the UBM means were adapted. Verification operates by computing the likelihood of a set of test features given both the claimed speaker GMM and the UBM. The log-likelihood ratio (LLR) is the difference between the logs of these two likelihoods. The LLR is compared to a pre-determined threshold and an accept or reject decision is made.

5. Effect of Ageing on Speaker Verification

To observe the extent to which the ageing issue affects speaker verification, an evaluation using the GMM-UBM classifier and the TCDSA database was designed. As discussed in Section 1, the ageing issue is of relevance to any verification scenario where there is a large time lapse between enrolment and verification. It is therefore of interest to investigate the conventional verification scenario, where a verification attempt is made some time after enrolment, and the forensic scenario, where a verification attempt is made using a sample recorded some time before enrolment. We refer to these two scenarios as *forwards* and *backwards* verification respectively.

For a *forwards* verification evaluation of each speaker, a model was trained using their first year of available data. Genuine speaker scores were calculated by testing their subsequent recordings against this model. The set of all other (17) speakers were used as imposters. Imposter scores were calculated by testing all imposter recordings occurring after the date of the training recording against the model.

For a *backwards* verification evaluation of each speaker, a model was trained using their last year of available data. Genuine speaker scores were calculated by testing their previous recordings against this model. The set of all other (17) speakers were used as imposters. Imposter scores were calculated by testing all imposter recordings occurring before the date of the training recording against the model. As well as emulating the forensic scenario, backwards verification provides a form of reverse validation (Lawson et al., 2009b).

This experimental design means that there are some cross-gender trials included. Although these trials will likely produce lower LLR scores than same-gender trials, they were included in order to maximise the number of imposters per test speaker.

In all cases, one minute of data was used for training. Each test evaluated the LLR of 10 segments of 30 seconds duration, all from one session. These scores were averaged to give one LLR score per test year. In Figure 2, LLR scores are plotted against the time between enrolment and verification sessions (i.e. age progression) for the set of all speakers, in forwards and backwards directions. Examples of two individual speaker tendencies are given in Figure 3.

It can be observed from the overall LLR trends in Figure 2 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 (Drygajlo et al., 2009) 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 by Doddington (2012), there is greater separation between genuine speaker and imposter scores as the age difference between them grows.

The individual speaker examples in Figure 3 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 Time Lapse = 0, Figure 3) will result in an increasing verification error rate as time progresses.

6. Effect of Quality on Speaker Verification

As discussed in Section 1, quality variation is an unavoidable attribute of long-term data. To attempt to separate the influence of quality and ageing on the recordings in the TCDSA database and potentially improve classification performance, it is important to be able to quantify the ‘quality’ of a recording, and how it relates to the resulting verification score.

The measurement of quality, in this sense, can be defined as the comparison of a speech sample to a predefined criterion known to influence the performance

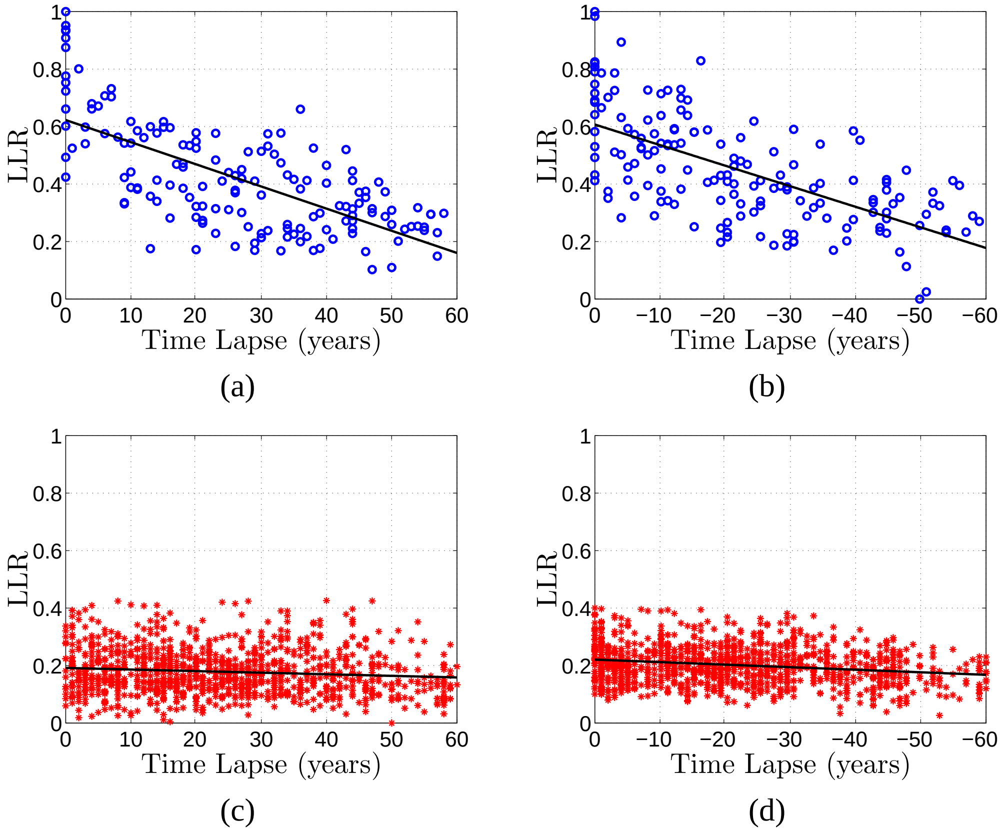


Figure 2: LLR scores against the time lapse between enrolment and verification for all 18 speakers. A line fit of the data is included to demonstrate the trend of the scores over time. (a) Genuine speakers, Forwards verification, (b) Genuine speakers, Backwards verification, (c) Imposters, Forwards verification, (d) Imposters, Backwards verification.

of the speaker verification system (Grother and Tabassi, 2007). A quality measure of a sample should therefore exhibit a relationship with its verification score (Richiardi and Drygajlo, 2008). As noted in (Grother and Tabassi, 2007), a quality measure is not synonymous with the fidelity of a sample as perceived by a human listener, but rather with the utility of the sample in the verification system.

A quality measure fulfilling these criteria can be jointly modelled with the verification score in order to improve classification - Drygajlo et al. (2011) use quality measurements (based on head pose and facial expression) of test images in combination with the verification score to improve face verification performance. In (Kryszczuk et al., 2007), quality measures for voice (signal-to-noise ratio) and

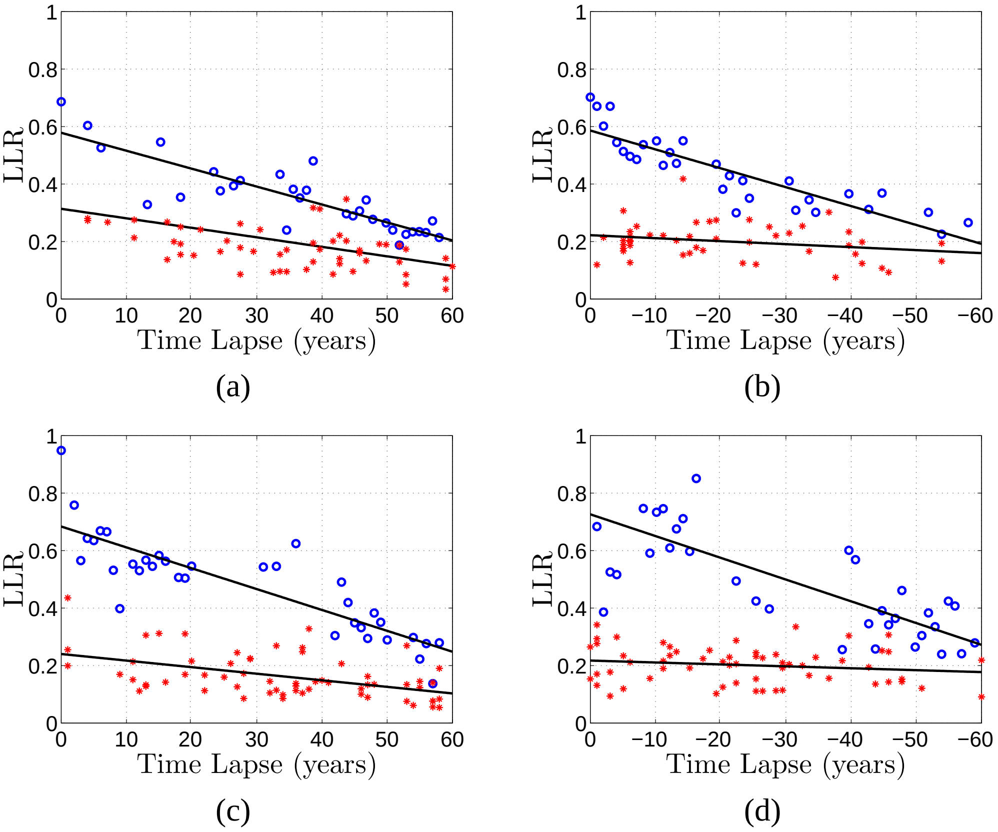


Figure 3: Genuine speaker (blue circles) and imposter (red asterisks) LLR scores against the time lapse between enrolment and verification for two speakers. A line fit of the data is included to demonstrate the trend of the scores over time (a) Cooke (male), Forwards verification, (b) Cooke, Backwards verification, (c) Queen (female), Forwards verification (d) Queen, Backwards verification.

face (distance to average face, image sharpness) are used to optimise multimodal fusion performance. Quality measures can also provide an estimate of the reliability of a verification result - Richiardi et al. (2006) studies this for both speaker and signature verification.

6.1. Quality Measures for Speech

Research into quality measures for speaker verification has been approached previously in Harriero et al. (2009); Richiardi and Drygajlo (2008); Garcia-Romero et al. (2005). Across these studies, reoccurring successful measures of quality have been: signal-to-noise ratio (SNR), P.563 - an International Telecommunication Union (ITU) standard for speech quality assessment (Malfait et al., 2006), and the higher-order statistic measure of kurtosis.

These measures are specific to quality measurement of speech signals, and thus can be referred to as modality-dependent (Richiardi et al., 2007). Modality-independent quality measures have not been thoroughly investigated for GMM-UBM based speaker verification. Richiardi et al. (2007) 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 models (and the UBM) share common covariance matrices. Measures based on covariance matrices are thus not applicable.

Harriero et al. (2009) proposes using the UBM log-likelihood score as a quality measure. While this is a good indicator of quality (used to discard outliers in our set of potential database recordings in Section 3.2), the UBM score is inherently contained in the LLR (the difference between speaker GMM log-likelihood and the UBM log-likelihood - Section 4.2). In addition, in (Harriero et al., 2009), the authors note that the UBM likelihood score reflects speaker-specific traits along with quality variation. Thus, jointly using the UBM score and the LLR to improve classification may be limited. An evaluation in Section 6 demonstrates that this is the case.

We propose a new model-based measure of quality, *Wnorm* (weighted norm), 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 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 Wnorm value.

The proposed approach uses a distance measure between GMMs to normalize verification scores. In this respect, the technique is related to the ‘Dnorm’ (distance normalization) class of normalizing methods (Ben et al., 2002; Yuan et al., 2009). 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. Dnorm is similar to Z-normalization, 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 i.e. the distance between the UBM and a speaker GMM trained from test data.

Since Wnorm is calculated from the mean matrices of a GMM and the UBM, it can be defined as a distance between mean ‘supervectors’ (Kinnunen and Li, 2010). 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 (Campbell et al., 2006) and bhattacharyya distance (Chang Huai et al., 2010). Wnorm is a similar distance measure, but operates in a different setting - as a means of test normalization for a GMM-UBM system.

6.2. Quality Measure Evaluation

Since the TCDSA database is influenced by a combination of ageing and quality actors, a separate dataset was used to evaluate the utility of the proposed Wnorm measure along with the selected classical measures, independently of ageing. The CSLU Speaker Recognition Corpus (Cole et al., 1998) was chosen for this purpose. From a total of 91 speakers, a gender-balanced UBM development set of 24 was selected, leaving 67 speakers for quality measure evaluation. The database consists of both scripted and spontaneous speech utterances over 12 sessions per speaker, spanning a 24 month period. All samples are telephone recordings. The spontaneous speech samples only were used in our experiments. Using the first session for training, and the remaining 11 for testing, a GMM-UBM speaker verification evaluation was carried for each of the 67 test subjects. In addition, for each testing sample, SNR, P.563, kurtosis and Wnorm measures were extracted. A framework for evaluating the utility of quality measures, suggested by NIST (Grother and Tabassi, 2007) was applied. With this approach, a quality measure can be considered useful if as the lowest quality samples are discarded, verification performance improves. All measures are first 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 in the database.

6.2.1. Signal to noise ratio (SNR)

Signal to noise ratio (SNR) has been 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:

$$SNR = 10 \log \frac{E_s}{E_{ns}} \tag{1}$$

Where E_s and E_{ns} are the mean energies of the speech and non-speech frames respectively.

6.2.2. Kurtosis

As clean speech has a distinctive distribution, statistics based on the distribution 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 is given by:

$$Kurtosis = \frac{1}{T} \sum_{t=1}^{T} \sum_{x=1}^{X} \left(\frac{s_{xt} - \mu_t}{\sigma_t} \right)^4 \qquad (2)$$

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.

6.2.3. Wnorm

The proposed Wnorm measure is given by:

$$Wnorm = \sqrt{\sum_{k=1}^{K} \sum_{d=1}^{D} \left(W_k \left(M_{kd}^{S} - M_{kd}^{UBM}\right)\right)^2} \qquad (3)$$

Where M^S and M^{UBM} are matrices of K component means, each of dimension D, of the test sample GMM and the UBM respectively. W is the vector of component weights of length K. A 30 second sample is used to train the GMM in this comparison.

6.2.4. P.563

The final quality measure to consider is P.563 (The ITU-T Standard for Single-Ended Speech Quality Assessment) (Malfait et al., 2006). 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. It was therefore necessary to downsample our data prior to P.563 extraction.

6.3. Experimental Evaluation

To evaluate the utility of each quality measure, we follow the NIST approach outlined by Grother and Tabassi (2007). Following a speaker verification evaluation of the CSLU dataset, samples of lowest quality are progressively removed

(1% at a time) from the evaluation, until 20% of the samples remain. Where a useful quality measure is employed, there should be a decrease in equal error rate (EER) as lower quality samples are discarded.

In Figure 4, EER is plotted against the percentage of samples rejected. As the first 20% of samples are rejected a drop in EER is observed for each of the four measures. As an increasing percentage are rejected however, Wnorm results in the sharpest drop in EER. Kurtosis and SNR continue to bring about a progressive decrease in EER, but at a slower rate. P.563 and the UBM likelihood have a low correlation with EER in comparison. Based on these trends, it would appear the Wnorm is the best indicator of quality on the CSLU database, followed by Kurtosis and SNR. The Wnorm measure is likely to be somewhat speaker-dependent - certain speaker's models will naturally lie further from the UBM than others. In this evaluation however, the extent of this effect is unclear, and would not seem to adversely affect its utility.

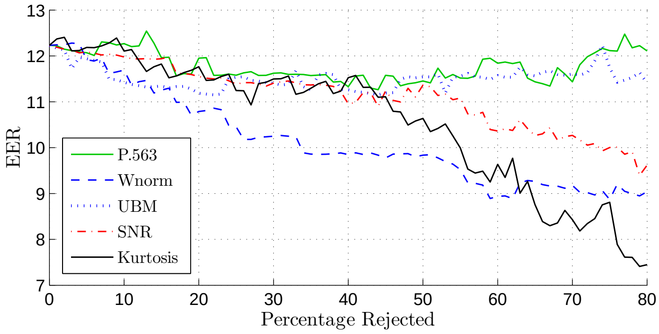


Figure 4: Speaker verification evaluation of the CSLU database: EER vs % of samples rejected, where samples are ordered from worst quality to best according to a range of quality measures.

Due to the fact the TCDSA database cannot be controlled for ageing, a similar EER vs percentage rejected evaluation would not be meaningful. To gauge the potential utility of the quality measures, the correlation between LLR and quality measures for both the TCDSA and CSLU databases can be compared in light of the EER vs percentage rejected evaluation in Figure 4. The linear correlation between the LLR of genuine speakers and the corresponding quality measures is shown in Figure 5. The measures with the strongest positive correlation - Wnorm and Kurtosis - were the most effective in the EER vs percentage rejected experiment, while P.563, which performed poorly in that experiment, has a small negative correlation. A similar correlation analysis on the TCDSA database (for the forwards direction) is given in Table 1. To consider the effect of ageing on the LLR, the correlation between the quality measures and ageing is included. Those measures with a higher correlation with LLR than with ageing can be considered more useful. Wnorm has the highest correlation to LLR relative to ageing. SNR has a greater correlation with ageing than Wnorm, but its correlation with LLR remains stronger than that with ageing. P.563 again displays low correlation. This analysis suggests that Wnorm, SNR and Kurtosis would be useful predictors of LLR on the TCDSA database.

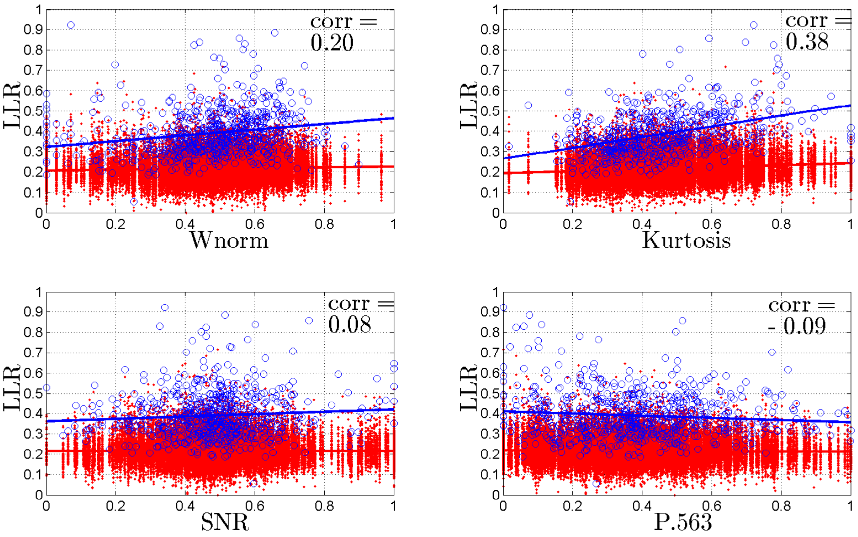


Figure 5: CSLU database: LLR of genuine (blue circles) and imposter (red dots) speakers against a range of quality measures. ‘corr’ is the linear correlation between genuine speaker LLR and quality.

	Wnorm	P.563	Kurtosis	SNR
LLR	0.37	0.10	0.19	0.26
Ageing	-0.11	0.13	-0.10	-0.19

Table 1: Correlation between quality measures, LLR score and ageing progression.

7. Ageing and Quality in Speaker Verification

Stacked generalization (or classification) (Wolpert, 1992; Ting and Witten, 1999), is an approach to classification whereby the output of several lower-level classifiers is used as the training and testing input to a higher-level classifier. This multi-level approach can improve classification performance by effectively combining the lower-level classification evidence.

Kryszczuk and Drygajlo (2007) present a stacked classifier framework as a general solution for uni- and multi-modal biometric verification. It is demonstrated that class-independent quality information can be used, together with the output of a conventional biometric classifier, to form the training and testing material for a higher-level classifier. This framework is presented as a general approach to biometric verification with single or multiple modalities and/or quality measures. An advantage of this approach is the flexibility to add or remove lower-level classifiers depending on available evidence.

This framework was applied successfully in our previous work (Kelly et al., 2012), where ageing information and LLR scores formed the lower-level inputs to a Support Vector Machine (SVM) classifier (Cristianini and Shawe-Taylor, 2000). Here we extend this work by including further class-independent quality measures as lower-level information. As seen in Sections 5 and 6, ageing information and quality measures, which do not carry class-specific information, are correlated with the verification scores of genuine users.

A schematic of the proposed stacked classifier framework is shown in Figure 6. Given a test speech sample, the LLR score is computed as in Section 4. Age progression is given as the time lapse in years between the date of enrolment of the claimed model and the date of testing. Quality measure extraction is provided by one of the four measures detailed in Section 6.1. A Z-normalization (Kinnunen and Li, 2010) is applied to the LLR score, based on the statistics of the training data.

The score, ageing and quality components are concatenated to form a vector input to the higher-level classifier. A scaling of the data is carried out as follows: Ageing progression and Z-normalized LLR are scaled to the range $[0, 1]$ based on the extreme values in the training set. A global scaling of the quality measures to the range $[0, 1]$ is applied based on the extreme values of all recordings in the database. The scaled training vectors are used to train an SVM classifier, which is applied to the scaled test vectors to output the final verification decision.

A speaker verification evaluation was designed to test the performance of the proposed stacked classifier framework with a combination of ageing information

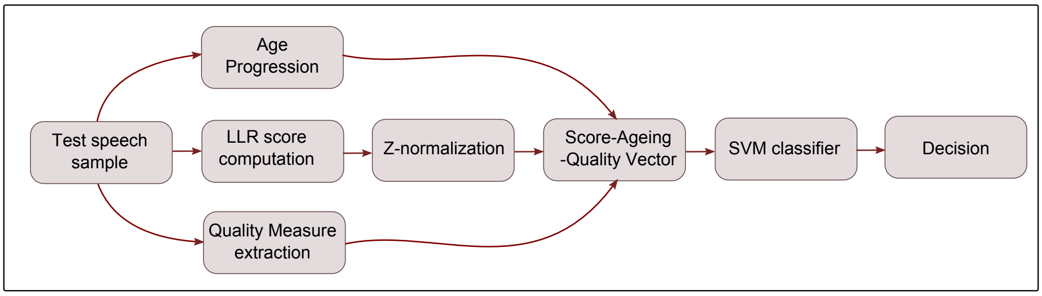


Figure 6: A schematic of the stacked classifier framework: The LLR score from the GMM-UBM system, ageing information and quality measures are combined as an input to an SVM classifer.

and quality measures. As per our previous evaluations, forwards and backwards directions were considered, to emulate both standard long-term speaker verification and forensic identification.

7.1. Baseline GMM-UBM Evaluation

A GMM-UBM evaluation of the TCDSA database, as detailed in Section 5 was first applied. The resulting set of genuine speaker and imposter LLR scores are as shown in Figure 2. The baseline performance of the GMM-UBM system was calculated from these scores. A conventional GMM-UBM system determines a decision threshold at the time of enrolment. A speaker-independent threshold was calculated by pooling the scores from all genuine speakers and imposters at enrolment (i.e. all LLR scores at Time Lapse $= 0$, Figure 2). A threshold was determined such that the HTER (half total error rate) was minimised. The HTER is the average of the FAR (False Acceptance Rate: the percentage of imposters falsely accepted) and the FRR (False Rejection Rate: the percentage of genuine speakers falsely rejected). This threshold was applied to the test data of each speaker individually. Any scores falling below the threshold were rejected and any above were accepted. The resulting average HTER in forwards and backwards directions is given in Table 2. In Figure 7, it can be seen that this threshold provides effective discrimination at the year of enrolment. As the time lapse increases, it provides increasingly poor verification performance.

7.2. Stacked Classifier Evaluation

The stacked classifier framework was evaluated with a leave-one-out training set. Thus for each speaker under test, a training set consisted of data from the other 17 speakers. This set of 17 speakers was divided into two groups with 9 and 8 speakers in each respectively. The speakers for each group were selected randomly, but kept balanced in terms of gender. The training set then consisted of genuine speaker scores from the 17 speakers not under test, along with their imposter scores from the random group of 9 speakers. The testing set consisted of the genuine speaker scores from the test speaker and their imposter scores from the random group of 8 speakers. Given that there was a variable number of data years available per speaker, it was necessary to reduce the number of test years per imposter to a random subset of 5 (in order to avoid biasing towards imposters with more data).

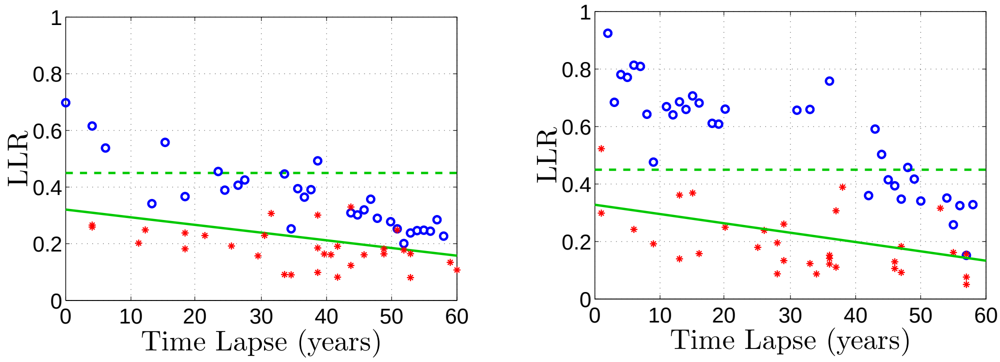
To evaluate the score-ageing-quality stacked classifier for each of the proposed quality measures, each LLR score from the training set was combined with its corresponding ageing information and quality measure. A linear SVM classification boundary was found such that the HTER on this training set was minimised. A linear kernel was chosen for the SVM due to the limited amount of data. The trained boundary was then used to classify the test speaker's data in score-ageing-quality space. To evaluate the performance of the score-ageing and score-quality stacked classifiers independently, the same procedure was followed, with one less dimension in the SVM classification.

A score-ageing boundary applied to the test data of two example speakers is shown in Figure 7. It is clear that the addition of ageing information into the boundary training produces a much more effective decision threshold compared to that of the baseline GMM-UBM. A score-ageing-quality boundary, as trained on the training set and applied to test data is shown in Figure 8 for an example speaker. The addition of quality adds extra discriminatory information to the score-ageing boundary. The average HTERs in forwards and backwards directions for different stacked classifier permutations are given in Table 2.

7.3. Evaluation Results

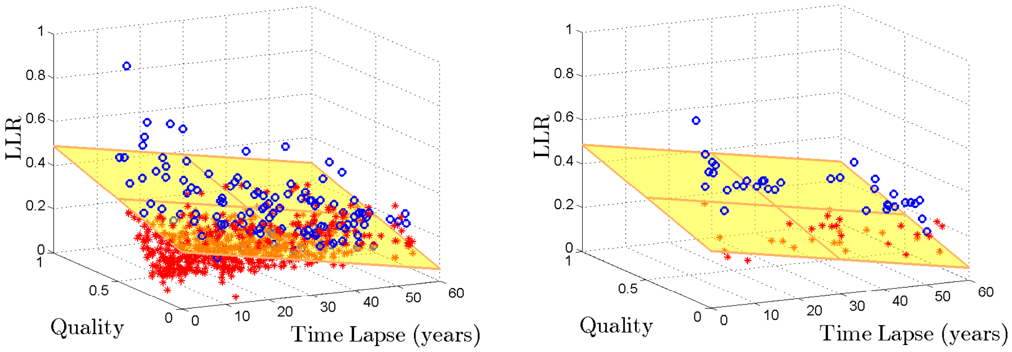
The results presented in Table 2 show that compared with the baseline HTER of 24.98/ 22.77% in forwards and backwards directions, the HTER falls to 15.41/ 16.90% by incorporating ageing information. Incorporating each of the quality measures independently results in a similar reduction. Considering the mean of the forwards and backwards HTER, Wnorm is the most successful, reducing HTER to 17.03%. P.563 and SNR perform marginally worse.

The best overall performance is achieved by compensating for both ageing and quality in the classifier. In the case of each quality measure, the addition of ageing information to the score-quality combination results in a further drop in HTER in both directions. Examining the mean of the forwards and backwards



(a) (b)

Figure 7: Genuine speaker (blue circles) and imposter (red asterisks) LLR scores for two example test speakers. The dotted green line represents the baseline GMM-UBM decision threshold. The solid green line represents the score-ageing stacked classifier boundary. (a) Cooke (male), Forwards Verification, (b) Queen (female), Forwards Verification.



(a) (b)

Figure 8: Training and testing LLR scores of genuine speakers (blue circles) and imposters (red asterisks) for an example speaker. The score-ageing-quality boundary was trained from the pooled LLR scores (forwards direction), ageing information, and Wnorm measures from 17 speakers. The test speaker is Queen (female). (a) Training, (b) Testing

HTER, Wnorm is the best performing measure when combined with ageing, reducing HTER to 13.72%. SNR is the second best performing measure, reducing HTER to 15.30%. Hence it can be seen that overall, although not completely independent, ageing and quality have a separate influence on score variability, and compensating for both results in the best 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 Marithoz (2006) 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. Table 3 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 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.

The P.563 algorithm (Malfait et al., 2006) was originally designed for narrowband speech assessment. The CSLU database is recorded over a telephone and is hence narrowband. The data is the TCDSA Database is wideband, sampled at 16kHz. Extending data down to 50Hz improves the quality of lower frequencies, impacting naturalness, presence and comfort in conversations. The quality factors above 3.4kHz will again yield a fuller and more natural sounding voice, with improvements in intelligibility of fricatives in particular (Skoglund et al., 2008). Ideally a single-ended quality measure designed for wideband speech would be applied to this data, but to date no such measure has been standardised. Applying P.563 on this speech data will consistently underestimate the quality. Assuming that this is the case, in our experiments, both the imposter and genuine speaker measurements are similarly affected. The consequences for the stacked classifier are therefore minimal.

Considering the evaluation of the CSLU database, Figure 4, and the relationship between LLR score and quality measures in Figure 5 and Table 1, it was proposed that Wnorm, SNR and Kurtosis would be successful measures in the stacked classifier evaluation. It seemed that P.563 would perform poorly (due low correlation between LLR score and quality). These predictions were correct in terms of Wnorm and SNR. However, P.563 outperforms Kurtosis, and is not significantly worse than Wnorm and SNR. This highlights that although a simple linear correlation between LLR and quality score can be a useful indicator

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	Wnorm	P.563	Kurtosis	SNR
Forwards	16.98	18.04	20.13	18.00
Backwards	17.08	17.32	19.49	17.45
Mean	17.03	17.68	19.81	17.73
GMM-UBM + Ageing + Quality	Wnorm	P.563	Kurtosis	SNR
Forwards	12.10	15.43	16.01	15.57
Backwards	15.34	16.52	17.54	15.02
Mean	13.72	15.98	16.78	15.30

Table 2: Average HTER (%) for all 18 speakers across their complete timespan (between 30 and 60 years, dependent on speaker), for the Baseline GMM-UBM system and the Stacked Classifier framework. Results for Forwards and Backwards directions, and the results for the mean of both directions are included.

of the utility of a quality measure, it is not a necessity. The discrepancy in the correlation of P.563 and LLR and its performance in the stacked classifier may be attributable to the fact that the pooled data of all speakers is used to generate the correlation value, Table 1, while in the stacked classifier, the training set is speaker-dependent.

8. Conclusions

In this work, we have investigated the effect of ageing and quality on speaker verification, and a stacked classifier framework has been introduced to compensate for these effects. A GMM-UBM speaker verification evaluation of the TCDSA database clearly demonstrated that LLR scores of genuine speakers decrease progressively as ageing progresses. This causes a conventional threshold, determined at the time of enrolment, to perform poorly.

Non-ageing-related variability is unavoidable in data acquired over a long time-period. To investigate the effect of quality variation on the LLR scores, a

	Confidence(%)					
	GMM-UBM + Quality			GMM-UBM + Ageing + Quality		
	P.563	Kurtosis	SNR	P.563	Kurtosis	SNR
Forwards	74.68	74.53	52.89	88.23	85.13	85.06
Backwards	61.09	81.41	58.10	61.20	74.77	-67.00

Table 3: Percentage confidence that the HTERs of Wnorm and each of P.563, Kurtosis and SNR, for the 'GMM-UBM + Quality' and 'GMM-UBM + Ageing + Quality' cases (Table 2) are statistically different. 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.

number of quality measures were extracted from the TCDSA data. Along with established measures of SNR, P.563 and Kurtosis, a new model-based measure of quality, Wnorm, was proposed. With the GMM-UBM evaluation on the TCDSA database providing a baseline for our experiments, it was found that ageing and quality measures combined independently with LLR score in the stacked classifier framework yielded significant improvements over the baseline HTER. The TCDSA database has been constructed specifically so that ageing is the dominant source of variation. The fact that a score-ageing stacked classifier results in a lower HTER than a score-quality stacked classifier confirms that this is the case.

The stacked classifier evaluations are in effect performed over an extremely long time-lapse of 30-60 years. It can be seen in Figure 7 that errors, particularly false acceptances, increase towards the upper end of this time range. In practice, operating ranges of biometric systems and forensic investigations are typically far shorter, in which case the performance of the stacked classifier will be far more robust. As mentioned in the introduction, the only current approach to overcoming vocal ageing is via model adaptation. An unanswered question with this approach is when to update a particular speaker's model. Logistically, it would be desirable to update as seldom as possible. Observing that the scores of genuine speakers and imposters in Figure 7 gradually converge on the score ageing boundary with ageing, a measure of the spread of the scores around the boundary could be used as a 'trigger' to carry out speaker adaptation. Due to its complexity, a much larger amount of data than is available at present would be required to model the vocal ageing process. The framework presented here is a solid basis for compensating for ageing given the amount of development data available.

There are several ways the framework could be refined, given additional data. In dealing with ageing, we have not considered the absolute age of speakers, only the progression of age. The process of vocal ageing is not constant. In adult speakers, the rate of change typically increases over the age of 60. A more refined modelling of ageing progression would incorporate absolute age. Secondly, the decision boundary for each speaker was trained using a set of development speakers. An improvement would be to tune the boundary to each speaker. Lastly, it is unlikely the best decision boundary is a linear one - a non-linear SVM kernel would likely improve performance.

Aside from the stacked-classifier framework, there are options for extending the GMM-UBM system (incorporating Eigenchannel compensation for example). It will first be necessary to identify additional development data sources. This will be the focus of future work.

Acknowledgments

This research has been funded by the Irish Research Council.

References

Ben, M., Blouet, R., Bimbot, F., 2002. 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). pp. 689–692.

Bengio, S., Marithoz, J., 2006. A Statistical Significance Test for Person Authentication. In: Odyssey 2006. pp. 279–284.

Benjamin, B. J., 1997. Speech Production of Normally Aging Adults. Seminars in Speech and Language 18, 135–141.

Bilmes, J., 1988. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Tech. rep., International Computer Science Institute.

Brandschain, L., Graff, D., Cieri, C., Walker, K., Caruso, C., Neely, A., 2010. Greybeard - Voice and Aging. In: Seventh Conference on International Language Resources and Evaluation (LREC '10).

Campbell, W. M., Sturim, D. E., Reynolds, D. A., Solomonoff, A., 2006. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. In: ICASSP. pp. 97–100.

Chang Huai, Y., Kong Aik, L., Haizhou, L., 2010. GMM-SVM Kernel with a Bhattacharyya-based Distance for Speaker Recognition. IEEE Transactions on Audio, Speech, and Language Processing 18 (6), 1300–1312.

Cole, R., Noel, M., Noel, V., 1998. The CSLU Speaker Recognition Corpus. In: International Conference on Spoken Language Processing. pp. 3167–3170.

Cristianini, N., Shawe-Taylor, J., 2000. An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech & Language Processing 19 (4), 788–798.

Dobry, G., Hecht, R. M., Avigal, M., Zigel, Y., 2011. 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.

Doddington, G., 2012. The Effect of Target/Non-Target Age Difference on Speaker Recognition Performance. In: Odyssey 2012.

Drygajlo, A., Li, W., Qiu, H., 2011. Adult Face Recognition in Score-Age-Quality Classification Space. In: Vielhauer, C., Dittmann, J., Drygajlo, A., Juul, N., Fairhurst, M. (Eds.), Biometrics and ID Management. Vol. 6583 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 205–216.

Drygajlo, A., Li, W., Zhu, K., 2009. Q-stack Aging Model for Face Verification. In: EUSIPCO 2009. Glasgow, Scotland.

Farrell, K. R., 2002. Adaptation of Data Fusion-based Speaker Verification Models. In: IEEE International Symposium on Circuits and Systems, ISCAS, 2002. Vol. 2. pp. II–851–II–854 vol.2.

FG-NET Aging Database, 2010. `http://www.fgnet.rsunit.com`.

French, J. P. F., Harrison, P., Windsor-Lewis, J., 2006. R v John Samuel Humble: The Yorkshire Ripper Hoaxer trial. The International Journal of Speech, Language and the Law 13 (2), 256–273.

Frigge, M., Hoaglin, D. C., Iglewicz, B., 1989. Some Implementations of the Boxplot. The American Statistician 43 (1), 50–54.

Garcia-Romero, D., Fierrez-Aguilar, J., Gonzalez-Rodriguez, J., Ortega-Garcia, J., 2005. Using Quality Measures for Multilevel Speaker Recognition. Computer Speech & Language 20 (2-3), 192–209.

Grother, P., Tabassi, E., 2007. Performance of Biometric Quality Measures. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4), 531–543.

Hansen, J. H. L., Lei, Y., 2009. The Role of Age in Factor Analysis for Speaker Identification. In: Interspeeech 2009. Brighton.

Harnsberger, J. D., Shrivastav, R., Brown, W., 2010. Modeling Perceived Vocal Age in American English. In: Interspeech 2010.

Harnsberger, J. D., Shrivastav, R., Brown Jr, W. S., Rothman, H., Hollien, H., 2008. Speaking Rate and Fundamental Frequency as Speech Cues to Perceived Age. Journal of Voice 22 (1), 58–69.

Harriero, A., Ramos, D., Gonzalez-Rodriguez, J., Fierrez, J., 2009. Analysis of the Utility of Classical and Novel Speech Quality Measures for Speaker Verification. In: Third International Conference on Advances in Biometrics. Springer-Verlag, pp. 434–442.

Hasan, T., Hansen, J. H., 2011. A Study on Universal Background Model Training in Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 19 (7), 1890–1899.

Hermansky, H., Morgan, N., 1994. RASTA Processing of Speech. IEEE Transactions on Speech and Audio Processing 2 (4), 578–589.

Juefei-Xu, F., Luu, K., Savvides, M., Bui, T., Suen, C., oct. 2011. Investigating Age Invariant Face Recognition based on Periocular Biometrics. In: International Joint Conference on Biometrics, IJCB, 2011.

Junyan, W., Yan, S., Guangda, S., Xinggang, L., 2006. Age Simulation for Face Recognition. In: 18th International Conference on Pattern Recognition, ICPR, 2006. Vol. 3. pp. 913–916.

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

Kinnunen, T., Li, H., 2010. An Overview of Text-Independent Speaker Recognition: From Features to Supervectors. Speech Communication 52 (1), 12–40.

Kryszczuk, K., Drygajlo, A., 2007. Q-stack: Uni- and Multimodal Classifier Stacking with Quality Measures. In: 7th International Conference on Multiple Classifier Systems. Springer-Verlag, pp. 367–376.

Kryszczuk, K., Richiardi, J., Prodanov, P., Drygajlo, A., 2007. Reliability-based Decision Fusion in Multimodal Biometric Verification Systems. EURASIP Journal of Applied Signal Processing 2007 (1), 74–74.

Lanitis, A., 2010. A Survey of the Effects of Aging on Biometric Identity Verification. International Journal of Biometrics 2 (1), 34–52.

Lawson, A. D., Stauffer, A. R., Cupples, E. J., S.J., W., Bray, W., Grieco.J.J., 2009a. The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial findings. In: INTERSPEECH 2009. Brighton, U.K.

Lawson, A. D., Stauffer, A. R., Smolenski, B. Y., Pokines, B. B., Leonard, M., Cupples, E. J., 2009b. Long-term Examination of Intra-session and Inter-session Speaker Variability. In: INTERSPEECH 2009. Brighton, United Kingdom.

Li, W., Drygajlo, A., Qiu, H., 2010. Aging Face Verification in Score-Age Space using Single Reference Image Template. In: IEEE International Conference on Biometrics: Theory, Applications And Systems (BTAS).

Linville, S. E., 1995. Vocal Aging. Current Opinion in Otolaryngology & Head and Neck Surgery 3 (3), 183–187.

Linville, S. E., 2004. The Aging Voice. The American Speech-Language-Hearing Association (ASHA) Leader, 12–21.

Malfait, L., Berger, J., Kastner, M., 2006. P.563 - the ITU-T Standard for Single-Ended Speech Quality Assessment. IEEE Transactions on Audio, Speech, and Language Processing 14 (6), 1924–1934.

McLaren, M., Mandasari, M. I., Leeuwen, D. A. v., 2012. Source normalization for language-independent speaker recognition using i-vectors. In: Odyssey 2012.

Mueller, P. B., 1997. The Aging Voice. Seminars in Speech and Language 18 (02), 159,169.

Park, U., Tong, Y., Jain, A. K., 2010. Age-Invariant Face Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (5), 947–954.

Presidential Speech Archive, 2012. Miller Center, University of Virginia, `http://www.millercenter.org/scripps/archive/speeches`.

Reubold, U., Harrington, J., Kleber, F., 2010. Vocal Aging Effects on F0 and the first Formant: A Longitudinal Analysis in Adult Speakers. Speech Communication 52 (7-8), 638–651.

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

Rhodes, R., 2011. Changes in the Voice across the Early Adult Lifespan. In: The International Association of Forensic Phonetics and Acoustics, IAFPA, 2011.

Ricanek, K., Tesafaye, T., 2006. MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. In: 7th International Conference on Automatic Face and Gesture Recognition, 2006. pp. 341–345.

Richiardi, J., Drygajlo, A., 2008. Evaluation of Speech Quality Measures for the purpose of Speaker Verification. In: Odyssey 2008: The Speaker and Language Recognition Workshop.

Richiardi, J., Kryszczuk, K., Drygajlo, A., 2007. Quality Measures in Unimodal and Multimodal Biometric Verification. In: EUSIPCO 2007.

Richiardi, J., Prodanov, P., Drygajlo, A., 2006. Speaker Verification with Confidence and Reliability Measures. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2006.

Rose, P., 2002. Forensic Speaker Identification. Taylor & Francis.

Rosenberg, A. E., Parthasarathy, S., 1996. Speaker Background Models for Connected Digit Password Speaker Verification. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 1996. pp. 81–84.

Schötz, S., 2006. Perception, Analysis and Synthesis of Speaker Age. Ph.D. thesis, Lund University, Sweden.

Skoglund, J., Kozica, E., Linden, J., Hagen, R., Kleijn, W. B., 2008. Voice over IP: Speech Transmission over Packet Networks. Springer, pp. 307–330.

Stathopoulos, E. T., Huber, J. E., Sussman, J. E., 2011. 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.

Ting, K. M., Witten, I. H., 1999. Issues in Stacked Generalization. Journal of Artificial Intelligence Research 10, 271–289.

Vipperla, R., Renals, S., Frankel, J., 2010. Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance. EURASIP Journal on Audio, Speech, and Music Processing 2010.

Wolpert, D. H., 1992. Stacked Generalization. Neural Networks 5 (2), 241–259.

Yuan, D., Liang, L., Xian-Yu, Z., Jian, Z., 2009. Studies on Model Distance Normalization Approach in Text-independent Speaker Verification. ACTA AUTOMATICA SINICA 35 (5).