Compensating for Ageing and Quality variation in Speaker Verification

Finnian Kelly[1], *Andrzej Drygajlo*[2] *and Naomi Harte*[1]

[1]Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland
[2]Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland
kellyfp@tcd.ie, andrzej.drygajlo@epfl.ch, nharte@tcd.ie

Abstract

Performing speaker verification in the simultaneous presence of ageing progression and changing speech sample quality is an important, open problem. The issues of ageing and quality variation go hand in hand; the effect of ageing increases with time, while variations in quality are also more likely to be encountered as time passes. In this work we demonstrate the effect of ageing on speaker verification performance, and show the relationship between quality variation and verification score via a range of established quality measures. We employ a stacked classifier framework to combine the output of the baseline verification system with ageing information and quality measures. This new approach to long-term speaker verification allows for a multi-dimensional decision boundary that significantly improves upon the baseline performance. The proposed framework is evaluated on the Trinity College Dublin Speaker Ageing Database.

Index Terms: speaker verification, ageing, stacked classifier

1. Introduction

In any large-scale biometric system, the accuracy degradation due to ageing is an important consideration [1, 2]. The effect of ageing-related variability in speaker verification has received marginal research attention however.

The physiological changes to the vocal mechanism and the associated acoustic changes to the voice have been extensively studied [3, 4, 5]. While the most noticeable changes in the voice occur in children and the elderly, progressive change occurs throughout adulthood. The most significant changes observed are a downward shift in the fundamental frequency and a change in timbre. Instability of pitch and intensity of speech, and a slowed rate of delivery are typical in older ($\geq$ 60 years) speakers.

An existing solution to the problem of ageing in speaker verification is to periodically update speaker models with new data [6]. This is not a feasible solution for large-scale systems, and introduces a weak point in terms of security. A more appropriate approach would be to automatically adjust to ageing-related change. This is not a straightforward task however. The first difficulty faced when researching the problem is a lack of data. There are currently no publicly available longitudinal speaker databases covering a time range of greater than 3 years. In our previous work [7] we presented the Trinity College Dublin Speaker Ageing (TCDSA) database, a new collection of longitudinal data from 18 speakers spanning a range of 30-60 years per speaker. In the same work, we demonstrated the progressive degradation in verification performance that occurs due to the ageing effect, and proposed a stacked classifier framework for improving long-term verification via an ageing-dependent decision boundary. Although the TCDSA database was compiled such that ageing was the dominant variability, variations in quality were unavoidable over the extremely long time period concerned.

A quality measure [8] provides an indication of the utility of a speech sample in a speaker verification system. In this work, we extend our previous investigation by considering a range of quality measures from the literature. After investigating the effectiveness of the measures on the TCDSA database, they are incorporated into the stacked classifier framework. This allows for a three dimensional decision boundary in score-ageing-quality space. This new approach, jointly accounting for both ageing and quality, proves more effective on challenging long-term data than accounting for either factor alone.

2. Effect of Ageing on Speaker Verification

A speaker verification evaluation on the TCDSA database was designed to observe the ageing effect.

2.1. TCDSA database

The TCDSA database [7] is a longitudinal speaker database spanning 30-60 years for a set of 18 speakers. The majority of the recordings were obtained from the national broadcasters of the U.K. and Ireland. There is a variable quantity of data per speaker, ranging between 2-35 recordings of length 1-30 minutes each. Accompanying the main database is a development database for background modeling, containing 30 seconds of speech from each of 120 speakers, balanced across gender, age and accent, and containing quality variation similar to that of the main database.

2.2. GMM-UBM speaker verification

A Gaussian Mixture Model - Universal Background Model (GMM-UBM) system [9] was used for the speaker verification experiments. Recent developments, e.g. Joint Factor Analysis (JFA) [10], extend the GMM-UBM approach to compensate for intersession variability. As discussed in [7], GMM-UBM provides a clearer insight at this investigative stage than techniques like JFA. Furthermore, the application of JFA to databases with widely varying conditions and limited content is a challenging problem in its own right.

2.2.1. Feature Extraction and the GMM-UBM system

All samples were pre-processed by downsampling to 16kHz, removing silences with an energy-based voice activity detector and applying pre-emphasis. MFCC features of length 12 were extracted and appended with delta and acceleration coefficients. Mean and variance normalization were applied to the features followed by RASTA filtering. These steps are typical in current verification systems [10]. A GMM-UBM verification system was developed by first training a UBM of 1024 mixtures using the contents of the TCDSA development set (Section 2.1). To generate speaker-specific models for each speaker in the TCDSA database, the UBM means were adapted with one minute of training speech [9]. The standard log-likelihood ratio (LLR) [10], was used for scoring test speech.

2.3. Ageing Speaker Verification evaluation

A model was trained for each speaker using one minute of their first year of available data. All recordings in the database from after the date of the training sample were considered test material. Thus for each speaker, a set of genuine speaker and imposter scores was generated from the set of recordings occurring after the date of their model training recording. Each test recording was split into 10 segments of 30 seconds duration. The resulting LLR scores were averaged to give one LLR score per speaker per test year. Given that there was a variable number of data years available per speaker, to ensure that the imposter tests were balanced across all imposters it was necessary to reduce the number of years per imposter to a random subset of 5. The genuine and impostor LLR scores are shown in Figure 1.

Errors in a speaker verification system are described by the false acceptance rate (FAR), the percentage of imposters incorrectly accepted, and the false rejection rate (FRR), the percentage of genuine speakers incorrectly rejected. A decision threshold is typically determined by reaching some tradeoff between FAR and FRR using verification scores from the time of enrolment. In our evaluation, in the majority of cases there is only one recording per speaker per year. To set a robust threshold, a global, rather than speaker-specific approach was adopted, in which the threshold was found by minimising the half total error rate (HTER) of the pooled enrolment scores from all speakers. The HTER is given by the average of the FAR and FRR. The enrolment score in this case is the LLR of a test segment usually taken from the same session as, but always distinct from, the training session. A threshold trained in this way is superimposed on the LLR scores in Figure 1.

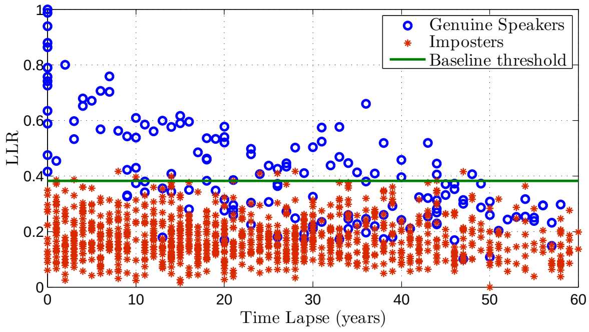


Figure 1: LLR scores of genuine speakers and imposters on a 60 year verification evaluation, and a conventional decision threshold, determined from scores at 'Time Lapse' = 0.

In Figure 1, genuine speaker scores decrease progressively until they converge with the set of imposter scores. It is clear that applying the baseline threshold as a verification boundary will result in an increasing error rate as time-lapse increases - the effect of ageing on the baseline system is therefore significant.

3. Effect of Quality Variation on Speaker Verification

The effect of sample quality variation has a significant effect on biometric verification [8, 2]. A quality measure that can predict the utility of a sample in a verification system can be jointly modeled with the verification score to reduce error rates [11]. A range of measures were extracted from the TCDSA data, based on those measures from previous studies that exhibited promising utility.

3.0.1. SNR

Signal to noise ratio (SNR) [12, 13, 14] was calculated using an energy-based voice activity detector. A sample is divided into 20ms frames designated as either speech or non-speech by an energy threshold:

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

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

3.0.2. *Skewness and Kurtosis*

Since clean speech has a distinctive distribution, higher order statistics like skewness and kurtosis [12, 13] can be used to measure quality. They are both described by the general equation:

$$S = \frac{1}{T} \sum_{t=1}^{T} \left(\frac{s_t - \mu_t}{\sigma_t} \right)^n \qquad (2)$$

Where s_t is the tth frame of the speech sample and μ_t and σ_t are its mean and variance respectively. S gives skewness when $n = 3$ and kurtosis when $n = 4$. The frame length was 20ms.

3.0.3. *UBM likelihood*

In [12], the likelihood of the test sample given the UBM was proposed as a quality measure for GMM-based speaker verification. The motivation for this measure is that assuming a UBM represents the acoustic space of the expected operating conditions, then degraded signals are more likely to have a lower likelihood given the UBM than non-degraded signals. The authors of [12] acknowledge that the likelihood with respect to the UBM reflects speaker-specific traits as well as quality related information. Their evaluation of this measure on NIST telephone data suggests promising utility however. In the GMM-UBM verification framework, given a speech sample O and a GMM λ_s for a particular speaker, the LLR is computed as:

$$LLR = \log p \left(O | \lambda_s \right) - \log p \left(O | \lambda_{UBM} \right) \qquad (3)$$

Where $p \left(. | \lambda \right)$ is the probability density function for any model λ. The second term, $\log p \left(O | \lambda_{UBM} \right)$ is the UBM likelihood.

3.1. Quality evaluation

Each of the four quality measures were extracted from the TCDSA database in 30 second segments and a $[0, 1]$ mapping [8] was applied. In [8, 12], utility of quality measures was demonstrated by plotting the LLR scores of genuine speakers and imposters against their corresponding quality measures. A quality measure was deemed useful if, as its value was increased, the scores of genuine speakers increased relative to the imposters. In the case of the TCDSA database, the ageing effect influences the LLR score alongside quality variation, and cannot be controlled for. This considered, the correlation between the LLR of genuine speakers and both ageing and quality measures was examined. The figures are given in Table 1. Measures that display stronger correlation with LLR score than with ageing can be regarded as better predictors of non-ageing related variability, and thus are expected to be most useful for reducing HTER when incorporated in the verification decision in some way.

	SNR	Skewness	Kurtosis	UBML
LLR	0.26	0.11	0.19	0.01
Ageing	-0.19	-0.13	-0.10	-0.15

Table 1: Correlation between quality measures, LLR score and ageing progression (UBML denotes UBM likelihood).

4. Ageing-Quality Stacked Classifer

Having observed that the effects of both ageing and quality influence the LLR score, an effective strategy for long-term verification needs to compensate for both.

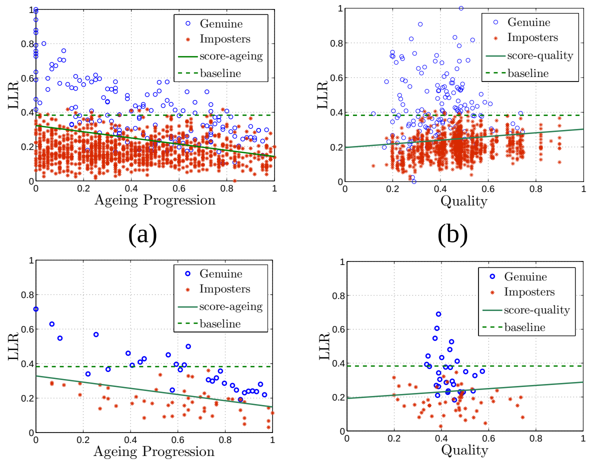
4.1. Stacked Classifier framework

In this approach, the output of several lower-level classifiers, together with the original class labels, form the input to a higher-level classifier. This is an effective way to combine different sources of 'evidence' in verification. In our previous work [7], a GMM-UBM system and ageing information constituted the lower-level information. A higher-level classification was carried out in score-ageing space with a Support Vector Machine (SVM) decision boundary. Here, we extend this framework by including each of the proposed quality measures as lower-level inputs. The final higher-level classification will thus involve a three-dimensional decision boundary in score-ageing-quality space. This approach can easily be extended further to include multiple quality measures. Here, we have considered only one quality measure at a time, as we are interested in the effectiveness of individual measures, rather than overall system performance.

In the proposed framework, each of LLR scores from the baseline evaluation, Section 2.3, together with their associated ageing (time-lapse) and quality information are concatenated into a vector. A Z-normalization [10] is applied to the LLR score, based on the statistics of the training data. The normalized LLR score and ageing information are then scaled to the range $[0, 1]$ based on the extreme values in the training set. The scaled vector forms the input to a linear SVM classifier. The SVM parameters are set such that the HTER on the training data is minimised.

4.2. Stacked Classifier experimental evaluation

An evaluation of the proposed framework was designed such that the performance could be directly compared to that of the baseline classifier described in Section 2.3. For each speaker, the set of all other (17) speakers over their complete 30-60 year time span constituted the training set. Pooling the training data, a score-ageing-quality boundary was trained. Applying this boundary to the test data of the speaker, the HTER was determined. An example of the trained boundary in score-ageing and score-quality planes is shown in Figure 2 (a) and (b). An application of the boundary to test data, in both cases, is given in Figure 2 (c) and (d). Note that the baseline and score-ageing thresholds are not identical at ‘Ageing Progression’ = 0, as the score-ageing threshold is optimised over all values of ‘Ageing Progression’. It is evident that the ageing and quality dependent boundaries track the progression of genuine speaker scores more closely than the baseline linear classifier. The average HTER for all speakers is shown in Table 2. Incorporating ageing and quality into the framework independently reduces HTER. The combination of both brings best performance - a score-ageing-SNR combination reduces HTER by 34% relative to the baseline, and by 4% compared to score and ageing alone. The relative performance of the quality measures aligns with the utility predicted in Table 1 - SNR and Kurtosis, the best performing measures, have a greater correlation with LLR score than with ageing progression.



(a) (b)

(c) (d)

Figure 2: Stacked Classifier example: (a) Training, score-ageing, (b) Training, score-quality, (c) Testing, score-ageing, (d) Testing, score-quality

	HTER(%)			
GMM-UBM	23.85			
+Ageing	16.31			
+Quality	SNR	Skew.	Kurt.	UBML
	18.39	19.08	18.93	19.99
+Ageing+Quality	SNR	Skew.	Kurt.	UBML
	15.66	16.27	16.08	16.29

Table 2: Average HTER (%) for all 18 speakers for the Baseline GMM-UBM system and the stacked classifier. ‘+Ageing’ denotes a score-ageing boundary, ‘+Quality’ denotes a score-quality boundary and ‘+Ageing+Quality’ denotes a score-ageing-quality boundary

5. Conclusions

To exploit the dependencies between ageing progression, quality variation and LLR score on the TCDSA database, a stacked classifier framework has been used to develop a verification boundary in score-ageing-quality space. The novel use of this framework in a long-term speaker verification evaluation was successful in significantly reducing HTER. This framework provides an effective and flexible approach to the challenging problem of speaker verification in the simultaneous presence of ageing progression and quality variation.

6. Acknowledgements

This research was funded by the Irish Research Council for Science, Engineering and Technology (IRCSET).

7. References

[1] A. Lanitis, “A Survey of the Effects of Aging on Biometric Identity Verification,” *International Journal of Biometrics*, vol. 2, no. 1, pp. 34–52, 2010.

[2] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia, “Quality measures in biometric systems,” *IEEE Security & Privacy*, vol. Preprint, 2011.

[3] S. E. Linville, “The Aging Voice,” *The American Speech-Language-Hearing Association (ASHA) Leader*, pp. 12–21, 2004.

[4] 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*, vol. 54, pp. 1011–1021, 2011.

[5] U. Reubold, J. Harrington, and F. Kleber, “Vocal Aging Effects on F0 and the first Formant: A Longitudinal Analysis in Adult Speakers,” *Speech Communication*, vol. 52, no. 7-8, pp. 638–651, 2010.

[6] K. R. Farrell, “Adaptation of Data Fusion-based Speaker Verification Models,” in *IEEE International Symposium on Circuits and Systems, ISCAS, 2002*, vol. 2, 2002, pp. II–851–II–854 vol.2.

[7] F. Kelly, A. Drygajlo, and N. Harte, “Speaker Verification with Long-Term Ageing Data,” in *International Conference on Biometrics, ICB, 2012*, 2012.

[8] P. Grother and E. Tabassi, “Performance of Biometric Quality Measures,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.

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

[10] T. Kinnunen and H. Li, “An Overview of Text-Independent Speaker Recognition: From Features to Supervectors,” *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[11] A. Drygajlo, W. Li, and K. Zhu, “Q-stack Aging Model for Face Verification,” in *EUSIPCO 2009*, Glasgow, Scotland, 2009.

[12] 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*. Springer-Verlag, 2009, pp. 434–442.

[13] 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.

[14] D. Garcia-Romero, J. Fierrez-Aguilar, J. Gonzalez-Rodriguez, and J. Ortega-Garcia, “Using Quality Measures for Multilevel Speaker Recognition,” *Computer Speech & Language*, vol. 20, no. 2-3, pp. 192–209, 2005.