

Speaker Verification with Long-Term Ageing Data

Finnian Kelly¹, Andrzej Drygajlo² and Naomi Harte¹

¹Department of Electronic and Electrical Engineering, Trinity College Dublin, Ireland

²Swiss Federal Institute of Technology Lausanne (EPFL), Lausanne, Switzerland

kellyfp@tcd.ie, andrzej.drygajlo@epfl.ch, nharte@tcd.ie

Abstract

The change experienced by the voice due to ageing must be considered in the development of a long-term speaker verification system. This difficult, largely open, research problem has received little attention to date. For this study, a new Speaker Ageing Database has been collected, containing speech from 18 speakers over a 30-60 year time span. A speaker verification evaluation of this data with a Gaussian Mixture Model - Universal Background Model system reveals that the verification scores of genuine speakers decrease progressively as the time span between training and testing increases, while the imposter scores are less affected. As a consequence, applying a decision threshold fixed at time of enrolment results in a high classification error rate after only a few years. A stacked classifier method of introducing an ageing-dependent decision boundary is applied, significantly improving long-term verification accuracy. Due to score variability at extended time spans however, accurate classification remains a challenging research problem. The ageing-dependent classification approach introduced here represents a first step towards dealing with long-term ageing in speaker verification systems.

1. Introduction

The accuracy of speaker verification systems is essentially limited by variability between enrolment and testing sessions. Variability can originate from environmental, channel and within-speaker sources. In recent years, there has been much research focused on compensating for such variability in speaker verification [16]. Long-term within-speaker variability, due to ageing, has received marginal attention however.

Detailed coverage of the changes in the adult vocal sys-

tem occurring with age, and their effect on the acoustic properties of the voice, can be found in [21, 22, 23]. In summary, ageing-related change of the vocal system affects the pitch, intensity level and rate of speech. The level of hoarseness and tremor in the voice is also increased. Therefore it is a reasonable assumption that a speaker model created at enrolment becomes gradually less relevant to that speaker as time progresses. This has often been mentioned as a potential problem facing long-term speaker verification [16, 3]. The effect of ageing is not exclusive to the voice; other biometric modalities including face, iris, signature and gait all change with age [17]. Ageing is a challenge facing biometrics in general.

There are important motivations to investigate the issue of ageing and speaker verification. In a typical speaker verification application, a speaker is first enrolled and some time later attempts to access the system via verification. We will refer to this as *forwards* speaker verification. Considering the change in the voice with ageing, a speaker model may need updating after a number of months [17]. In a large-scale application, the logistics and cost involved make this a difficult task. Furthermore, there is the question of when and how to update the speakers' models. To develop a forwards-ageing-aware system, with the ability to adapt to ageing-related change, would be a more favourable option.

An application of *backwards* speaker verification is in the area of forensic investigation. This scenario involves hypothesis testing between a suspected-speaker model and a trace recording. Due to the nature of forensic investigation, it is likely that a significant time period will have passed between the date of the trace recording and the time at which a suspect's speech can be recorded to create a speaker model. An example of a forensic case with a large time interval was the 'Yorkshire Ripper Hoaxer Trial' [8] where there was a gap of 27 years between the 'scene of the crime' and the custody recordings. It is of clear interest to develop a speaker verification approach that is backwards-ageing-aware for such scenarios.

To date, research into the voice and age for speech tech-

This research has been funded by the Irish Research Council for Science, Engineering and Technology (IRCSET) and Science Foundation Ireland (SFI) (grant number 09/RFP/ECE2196).

nology applications has included age perception [11], age estimation and synthesis [25], the effect of absolute speaker age on speaker verification [10] and speech recognition [26]. An attempt to uncover an ageing effect on speaker verification over a three year period was presented in [19]. It was concluded that over a time interval of this length, variability in genuine speaker scores is more attributable to inter-session variability than to ageing. Our previous work [14] analysed longitudinal speaker data over a 30-40 year period. Using a Gaussian Mixture Model - Universal Background Model (GMM-UBM) system, the log likelihood ratio (LLR) scores of the speakers' recordings against their models were presented as a function of time span between training and testing. It was observed that at a time span of greater than 5 years, the LLR scores had fallen outside the range of expected inter-session variability for the speaker. The impact of this score degradation trend on the accuracy of a verification task has not yet been established.

To our knowledge, there are currently a limited number of longitudinal speaker databases available. The CSLU and MARP databases [4, 18] contain longitudinal data over a 2-3 year period. The Greybeard Corpus [2], compiled by NIST for their Speaker Recognition Evaluation (SRE) 2010, contains longitudinal data over a 2-12 year period. The Greybeard Corpus is currently available only to SRE 2010 participants however, and there are no publications that the authors are aware of that report results on the database. The novel contributions of this work are:

- A new Speaker Ageing Database, containing 18 speakers with a 30 to 60 year data range per speaker, and an age-balanced UBM development database.
- The presentation of genuine speaker and imposter speaker verification scores as a function of time. Speaker verification in both forwards and backwards directions is considered, replicating both conventional and forensic long-term scenarios.
- A stacked classifier method of incorporating ageing information into the classification decision, improving long-term speaker verification accuracy compared with a threshold fixed at time of enrolment.

Section 2 details the Speaker Ageing Database. Section 3 describes the baseline GMM-UBM speaker verification system. In Section 4, the effect of ageing on genuine speaker and imposter verification scores is presented. In Section 5, a stacked classifier approach is introduced. An evaluation of the stacked classifier along with the baseline classifier is outlined in Section 6. Finally, conclusions are presented in Section 7.

2. Speaker Ageing Data

Obtaining suitable data is a major challenge faced when approaching the study of speaker ageing and verification.

As evident from Lawson's work [19], the effect of ageing across a time span of 3 years is not significant relative to inter-session variability. To understand the ageing effect, a much longer time span than this is clearly necessary. An ideal database would contain data for each speaker at regular intervals in controlled conditions. In reality, acquiring samples over several decades in controlled conditions is not achievable. The only way to proceed is to use data with acoustic properties that are as consistent as possible.

2.1. Speaker Ageing Database

A new Speaker Ageing Database has been compiled for this work, consisting of speech from 18 speakers (9 male, 9 female). The recordings span a range of between 30 to 60 years per subject, at irregular intervals of between 1 to 10 years. The majority of the material originates from the national broadcasters of the U.K. and Ireland, the BBC (British Broadcasting Corporation) and RTÉ (Raidió Teilifís Éireann). Further publicly available samples were obtained from YouTube and The Miller Center [1].

The recordings contain radio and television interviews and speeches. As would be expected with a database spanning such a large time period, there are variations in the quality of the recordings. Any audibly noisy recordings, or segments of recordings, were discarded. The spectral content of the recordings was examined, and any with significant frequency artefacts were removed. As an objective measure of quality, the likelihood of the recordings against an age-balanced Universal Background Model (UBM) was calculated (the construction of this UBM is detailed in Section 3.1). This approach has previously been applied to measure quality for speaker verification [12]. Assuming a UBM represents the common distribution of speaker features of the test population, degraded recordings will score lower against the UBM than those that are 'clean'. Any UBM scores that fell outside 1.5 times the interquartile range of the set of all scores was deemed an outlier, and the associated recording was discarded. Variability in channel and environmental conditions cannot be avoided in the collection of a long-term database such as this. However, every effort has been made at the stage of data collection to minimise the extent of this variability.

2.2. UBM development database

In addition to the Speaker Ageing Database, a separate database was compiled for UBM development. The University of Florida Vocal Aging Database (UF-VAD) (extemporaneous set) [11] was combined with data sourced from Youtube and the Miller Center [1]. The resulting dataset contains 1 hour of data from 120 speakers (30 seconds each). The speakers are balanced across gender, age and accent, and a variety of recording conditions are present. The age content is split evenly into three profiles: 35, 36-55 and

over 55. English, Irish and American accented speakers are included. Composing the dataset in this way ensured that it was as close as possible a representation of the population within the Speaker Ageing Database.

3. Speaker Verification System

While there are other systems for speaker verification that have come to prominence recently, we chose to present this work on a GMM-UBM system. This is the first work to explore speaker verification over such an extended time span. As the extent of the ageing effect in this domain is not fully understood, presenting work on a baseline system is therefore in the interest of clarity. Recent speaker verification developments, such as Joint Factor Analysis (JFA) [15, 16], are extensions of the GMM-UBM approach. This work is therefore a basis from which extension to these methods is possible. It should be mentioned however, that systems like JFA, while successful in recent NIST SRE events, do not provide an ‘out of the box’ solution to all applications. In this work, the limited and varied nature of the data introduces extra challenges in training an effective JFA system. This is also true of the forensic domain, where there are difficulties in using JFA due to limited development data [9]. In addition, the legal requirement for the explanation of the strength of evidence to be readily understandable to the lay persons of the court [5], make GMM-UBM a more favourable option in the forensic domain at present.

3.1. GMM-UBM system

The verification system used in this work was a standard GMM-UBM system [24]. Using the complete UBM dataset (as described in Section 2.2), separate male and female UBMs of 512 mixtures were trained via the EM algorithm and joined to create a UBM of 1024 mixtures. Individual speaker models were created by Bayesian adaptation [24] of the gender-independent UBM given the speaker data from the Speaker Ageing Database. One minute of speech was used to train each speaker model and the UBM means only were adapted.

3.2. Pre-processing and Feature Extraction

Standard pre-processing and feature extraction steps were applied (see the recent review by Kinnunen [16]). All recordings were first downsampled to 16 kHz. Energy-based silence removal and pre-emphasis were applied. 12-dimensional MFCC vectors were extracted over 20 ms windows with 50% overlap. Delta and acceleration coefficients were appended, resulting in a length 36 feature vector. Mean and variance normalisation were applied. Due to the varied sources of the data, RASTA filtering [13] was also applied to limit the influence of different channels.

4. Effect of Ageing on the Speaker Verification System

The first objective was to establish the effect of ageing on the baseline verification system using our Speaker Ageing Database. A test was designed to observe the behaviour of the genuine speaker scores (verification scores of speakers’ test data on their own models) and the imposter scores (verification scores of speakers’ test data on models that are not their own) over a long-term period.

In all cases, testing involved scoring (using the standard log likelihood ratio [16]) 10 segments of 30 seconds duration, all from one session, against the test model. These scores were averaged to give one log likelihood ratio (LLR) score per test year.

4.1. Forwards Speaker Verification

As discussed in Section 1, forwards speaker verification refers to the scenario where some time has passed between enrolment and testing. Forwards verification was conducted by training a speaker model with data from their first year of available speech. Genuine speaker scores were obtained by testing the model with recordings from each subsequent available year. The set of imposters was chosen as all other speakers in the database. Imposter scores were obtained by testing the model with each imposter recording occurring after the date of the training recording.

4.2. Backwards Speaker Verification

Backwards speaker verification refers to the scenario in which some time has passed between the recording of the test speech and the enrolment. Backwards verification was conducted by training a speaker model with data from their last year of available speech. Genuine speaker scores were obtained by testing the model with recordings from each previous available year. The set of imposters was chosen as all other speakers in the database. Imposter scores were obtained by testing the model with each imposter recording occurring before the date of the training recording.

4.3. Results

Figure 1 shows the genuine speaker and imposter LLR scores for all 18 speakers in forwards and backwards directions. Figure 2 shows LLR scores for an example speaker.

From the global score trends in Figure 1 it can be seen that as age progresses, the LLR scores of the genuine speakers decrease in general. This has been observed in [14], and a similar trend has been observed in the domain of face ageing [7]. The LLR scores of the imposters experience less change over time. It is significant that the score trends of the genuine speakers and imposters behave differently over time. This confirms that the speaker’s age is the source of the drop in LLR score, rather than a change in the acous-

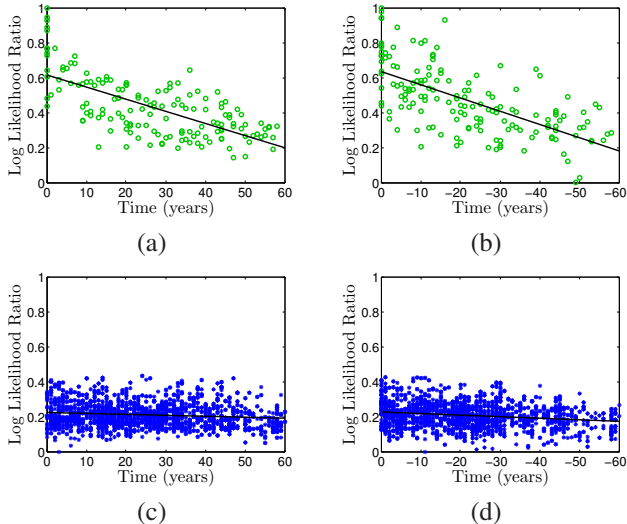


Figure 1. LLR scores for all 18 speakers: (a) genuine speakers, forwards, (b) genuine speakers, backwards, (c) imposters, forwards, (d) imposters, backwards

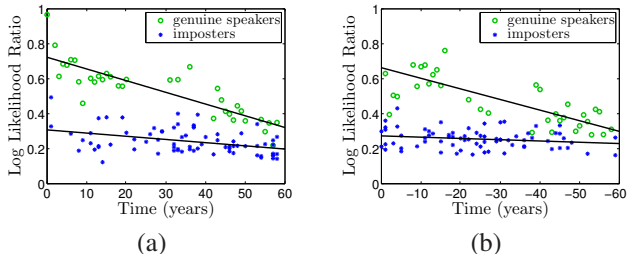


Figure 2. Genuine speaker and imposter score distributions for an example speaker: (a) forwards testing, (b) backwards testing

tic properties of the recordings, as this would affect both distributions in a similar way.

The example of the individual speaker and imposter score distributions in Figure 2 demonstrates the same trend as the global case. The variability of the scores can be attributed to numerous factors, including: the age of the speaker at time of training, the rate of ageing-related change for a given speaker, the characteristics of the training and testing recordings (content, channel, recording conditions). By inspection, it is clear that a decision threshold fixed during enrolment will fail after a number of years.

5. Ageing-aware Speaker Verification

On an individual speaker basis, the effect of ageing on the voice varies [21]. Approaching the problem of creating an ageing-aware system at a feature or modelling level is therefore a difficult task. At the level of scores however, it has been seen that there is a common tendency for speakers' scores to degrade over time. A logical extension to the GMM-UBM system is to use this general degradation to select a time-dependent decision boundary

5.1. Stacked Classifier Approach

A suitable method of combining the output of a GMM-UBM classifier with ageing information is a 'stacked classifier' system. This approach has been successfully applied to the similar scenario of ageing and face verification [7, 20]. In the stacked classifier scheme, the baseline classifier (or classifiers) is applied to the data. The output of this 'level 0' classification is combined with additional information to carry out a new 'level 1' classification. The additional information could be a quality measure extracted from the signal (*e.g.* SNR) or speaker information (*e.g.* age, gender). If there is a dependency between this additional information and the output of the level 0 classifier, it will potentially offer an improvement in classification of the testing data.

In the case of long-term speaker verification, we propose the use of age progression as an additional measure in the stacked classifier model. Age progression is defined as the time interval in years between the date of the recording used to train the model and the date of the test recording. Although the age of a speaker provides no inter-speaker discriminatory information, over a long time period, the age progression of a speaker has a clear influence on the verification scores (Figure 1).

The approach proceeded as follows: Given a test recording and a speaker model, the LLR score was obtained from the GMM-UBM system and the age progression was calculated. A Z-score normalization [16] was applied to the LLR score and it was passed, along with age progression information, to the level 1 classifier. This was repeated for all tests, resulting in a set of client and imposter score and age progression pairs. A Support Vector Machine (SVM) classifier with a linear kernel was employed for the final (level 1) decision boundary estimation and classification.

6. Experimental Evaluation

To establish if age information can be used to improve the accuracy of long-term speaker verification, two verification tasks were conducted. In the first, a decision threshold fixed at enrolment was used. In the second, an ageing-dependent decision boundary was applied via the stacked classifier method discussed in Section 5.1.

6.1. Baseline Classifier Evaluation

To set a decision threshold fixed at the time of enrolment, a model was trained for each speaker using one minute of speech. A set of enrolment scores were obtained by scoring the remainder of the enrolment session against the model in 30 second segments. For each speaker, the decision threshold between the speaker enrolment scores and a set of imposter scores was selected such that half total error rate (HTER) was minimised. The HTER is defined as the average of the false acceptance rate (FAR) and false rejection

rate (FRR).

The fixed decision threshold was then applied to classify the scores of the speaker in both forwards and backwards directions, as in Sections 4.1 and 4.2.

6.2. Stacked Classifier Evaluation

As detailed in Section 2.1, the minimum interval between data samples for each speaker is one year. To train an ageing-dependent decision boundary, scores and age progression information at multiple time intervals are required. In a realistic biometric application, recordings for each subject across a time span in the range of years would not be available.

A realistic approach was applied here, whereby a global ageing-dependent decision boundary was trained for each speaker. The training set consisted of all speakers other than the test speaker. Thus for each of the 18 subjects, a rotating training set of 17 speakers was used. Pooling the genuine speaker and imposter LLR scores from the training set, along with age progression information, a linear SVM decision boundary was trained such that the HTER on the training data was minimised. An example of a trained SVM for one speaker is shown in Figure 3.

The decision boundary was then applied to classify the scores of test speaker in both forwards and backwards directions, as in Sections 4.1 and 4.2.

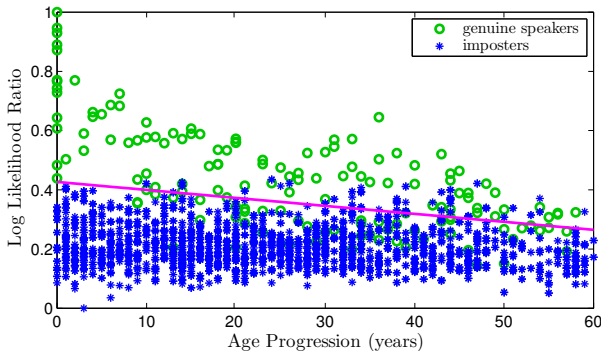


Figure 3. A global ageing-dependent decision boundary trained from 17 speakers’ LLR scores and age progression information

6.3. Evaluation Results

The baseline classifier and stacked classifier were evaluated on each of the 18 subjects of the Speaker Ageing Database. In Figure 4, evaluation results are presented for two example speakers in forwards and backwards directions. The LLR test scores of the genuine speaker and imposters along with the fixed decision threshold and ageing-dependent decision boundary are shown. It is clear from HTER rates of the two classification approaches that the ageing-dependent decision boundary is more accurate over the long-term.

The average HTER for all 18 speakers using both classification approaches, tested over different ranges of age progression, is given in Table 1. It can be observed that at an age progression range of five years, the fixed threshold and ageing-dependent boundary have similar accuracy. As the range of age progression increases, genuine speaker scores fall below the fixed threshold and are misclassified. The ageing-dependent boundary performs to a higher accuracy over the long-term, but at extended time spans however, the overlap of genuine and imposter speaker scores make accurate classification difficult.

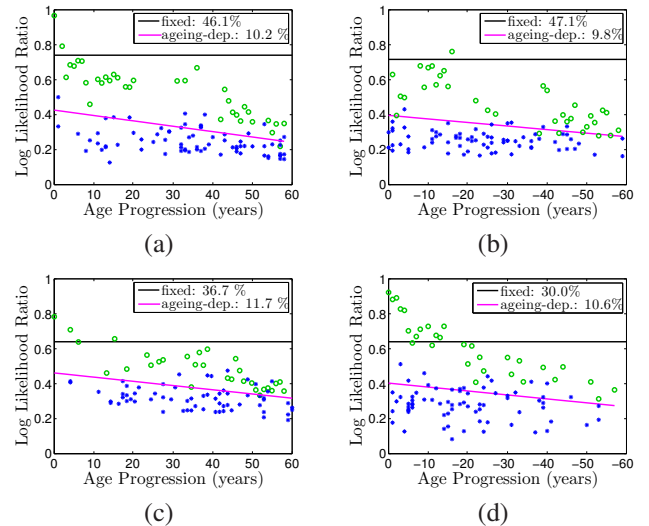


Figure 4. Fixed thresholds and ageing-dependent decision boundaries and their corresponding HTER, for two example speakers: (a) Female, forwards, (b) Female, backwards, (c) Male, forwards, (d) Male, backwards.

Age Progression (years)	5	10	20	40	60
Forwards					
Fixed Threshold	10.8	15.9	26.5	32.2	36.1
Ageing-dependent	7.3	9.2	10.4	17.2	17.5
Backwards					
Fixed Threshold	13.7	18.1	18.9	24.8	29.1
Ageing-dependent	10.2	11.4	12.3	17	21.9

Table 1. Average HTER for all 18 speakers in the Speaker Ageing Database across different ranges of age progression.

7. Conclusions

Variability between training and testing sessions is the major challenge facing speaker verification. One source of variability in the voice occurs due to ageing. Despite the fact it has received little research attention, it is of relevance to both standard biometric applications and the field of forensic investigation.

Uncovering the ageing effect on a speaker verification system is a challenging task, particularly in the acquisition of suitable data. A Speaker Ageing Database has been compiled for this work, containing 18 speakers over a 30–60 year period, which to our knowledge represents the largest collection of long-term speaker data suitable for use in a speaker verification evaluation.

In our speaker verification evaluation of this database, correlation between the scores of genuine speakers and age progression is observed. There is significantly less correlation between imposter scores and age progression. As a consequence, a classification threshold fixed at enrolment results in a decrease in accuracy over time. Using a stacked classifier approach to implement an ageing-dependent decision boundary, long-term verification results are significantly improved.

The data used in this evaluation presents challenges in terms of variability. Alongside the long-term trends presented in the verification evaluation, the short-term score variability is also observable. Including measures of recording quality in the stacked classifier model could potentially reduce score variability and improve accuracy [6]. It is also of interest to investigate whether systems such as JFA, which build on the GMM-UBM approach by incorporating session compensation, can be applied to this difficult scenario of uncontrolled, long-term data of limited quantity.

References

- [1] Presidential Speech Archive, Miller Center, University of Virginia. <http://millercenter.org/scripps/archive/speeches>, 2011.
- [2] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely. Greybeard longitudinal speech study. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, May 2010.
- [3] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J. F. Bonastre, and D. Matrouf. Forensic speaker recognition. *Signal Processing Magazine, IEEE*, 26(2):95–103, 2009.
- [4] R. Cole, M. Noel, and V. Noel. The CSLU speaker recognition corpus. In *International Conference on Spoken Language Processing*, 1998.
- [5] A. Drygajlo. Forensic automatic speaker recognition. *Signal Processing Magazine, IEEE*, 24(2):132–135, 2007.
- [6] A. Drygajlo, W. Li, and H. Qiu. Adult face recognition in score-age-quality classification space. In *Biometrics and ID Management*, volume 6583 of *Lecture Notes in Computer Science*, pages 205–216. Springer Berlin / Heidelberg, 2011.
- [7] A. Drygajlo, W. Li, and K. Zhu. Q-stack aging model for face verification. In *EUSIPCO 2009*, Glasgow, Scotland, 2009.
- [8] 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.
- [9] J. Gonzalez-Dominguez, B. Baker, R. Vogt, J. Gonzalez-Rodriguez, and S. Sridharan. On the use of factor analysis with restricted target data in speaker verification. In *Odyssey 2010*, Brno, Czech Republic, 2010.
- [10] J. H. L. Hansen and Y. Lei. The role of age in factor analysis for speaker identification. In *Interspeech 2009*, Brighton, 2009.
- [11] 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.
- [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 *Proceedings of the Third International Conference on Advances in Biometrics*, pages 434–442, 2009.
- [13] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–589, 1994.
- [14] F. Kelly and N. Harte. Effects of Long-Term Ageing on Speaker Verification. In *Biometrics and ID Management*, volume 6583 of *Lecture Notes in Computer Science*, pages 113–124. Springer Berlin / Heidelberg, 2011.
- [15] 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.
- [16] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, 52(1):12–40, 2010.
- [17] A. Lanitis. A survey of the effects of aging on biometric identity verification. *Int. J. Biometrics*, 2(1):34–52, 2010.
- [18] A. D. Lawson, A. R. Stauffer, E. J. Cupples, S. Wemndt, W. Bray, and J. Grieco. The multi-session audio research project (MARF) corpus: Goals, design and initial findings. In *INTERSPEECH*, Brighton, United Kingdom, 2009.
- [19] 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*, Brighton, United Kingdom, 2009.
- [20] W. Li, A. Drygajlo, and H. Qiu. Face verification in score-age space using single reference image template. In *IEEE International Conference on Biometrics: Theory, Applications And Systems (BTAS)*, 2010.
- [21] S. E. Linville. Vocal Aging. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 3(3):183–187, 1995.
- [22] P. B. Mueller. The Aging Voice. *Seminars in Speech and Language*, 18(02):159,169, 1997.
- [23] 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.
- [24] 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.
- [25] S. Schötz. *Perception, Analysis and Synthesis of Speaker Age*. PhD thesis, Lund University, Sweden, 2006.
- [26] 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.