Effects of Long-Term Ageing on Speaker Verification

Finnian Kelly and Naomi Harte*

Department of Electronic and Electrical Engineering,
Trinity College Dublin, Ireland
{kellyfp,nharte}@tcd.ie

Abstract. The changes that occur in the human voice due to ageing have been well documented. The impact of these changes on speaker verification is less clear. In this work, we examine the effect of long-term vocal ageing on a speaker verification system. On a cohort of 13 adult speakers, using a conventional GMM-UBM system, we carry out longitudinal testing of each speaker across a time span of 30-40 years. We uncover a progressive degradation in verification score as the time span between the training and test material increases. The addition of temporal information to the features causes the rate of degradation to increase. No significant difference was found between MFCC and PLP features. Subsequent experiments show that the effect of short-term ageing (<5 years) is not significant compared with normal inter-session variability. Above this time span however, ageing has a detrimental effect on verification. Finally, we show that the age of the speaker at the time of training influences the rate at which the verification scores degrade. Our results suggest that the verification score drop-off accelerates for speakers over the age of 60. The results presented are the first of their kind to quantify the effect of long-term vocal ageing on speaker verification.

1 Introduction

With ageing, the subsystems which make up the human speech production system undergo progressive physiological change, bringing about significant changes in the voice. The respiratory system is affected by the decreasing rate and strength of muscle contraction. In the larynx, ossification of cartilages and atrophy of muscle tissue are the primary anatomic changes. Changes to the supralaryngeal system include loss of functionality of the tongue and facial muscles. These changes have been documented in numerous studies [1,2,3,4]. These anatomical changes affect the acoustic properties of the voice in a number of ways. Pitch, the rate and intensity of speech, and the 'quality' of the voice are the properties of the voice most affected [3,5]. In general, elderly speakers' voices experience more variability than younger speakers' [1,2].

Much research attention has been payed to the characteristics of the ageing voice. Very little attention however has been devoted to the effect of vocal ageing

* This research has been funded by the Irish Research Council for Science, Engineering and Technology.

C. Vielhauer et al. (Eds.): BioID 2011, LNCS 6583, pp. 113–124, 2011.

on the accuracy of speaker verification. With the increased use of biometric technology for security and forensic applications, understanding the impact ageing has on speaker verification is important. The primary difficulty in assessing this effect experimentally is a lack of longitudinal data. The effect of the ageing voice on the accuracy of speech recognition has been studied using two different sets of speakers from 'adult' and 'older' populations [6] . For speaker verification, a database of the same speakers over an extended time period is required. Some available databases [7,8] contain 'long-term' data covering a time span of 2-3 years. In [9], an attempt is made to observe vocal ageing effects on speaker verification over a 3 year period. However, in the context of vocal ageing, where the onset of change as well as the rate at which it progresses is speaker specific [1], a significantly longer time span would be required to uncover any definite trend. In this work, we examine the effect of a 30-40 year time span on the speaker verification accuracy of a number of subjects. We experimentally uncover a long term degradation in performance which is outside the bounds of expected session variability. We also present results which show that the rate of verification drop-off is not constant across all ages, with the rate of degradation appearing to increase above the age of 60. These results correlate well with expectations in terms of ageing [1,3], but to our knowledge this is the first work to quantify long-term ageing effects in terms of their impact on speaker verification.

These are early stage findings and are investigative in nature. The emphasis is not on system performance but rather uncovering previously unquantified effects of age on a speaker verification system. Finally, we recognise that although our database is limited in terms of the number of speakers, there is a sufficient quantity and variation of speech to reach some important conclusions.

2 Speech Data

To carry out the longitudinal analysis in this paper, an ageing database of 13 speakers was compiled. The database contains 15 hours of speech from 7 males and 6 females and was obtained from the archive material of the national broadcasters of the U.K. and Ireland: the BBC (British Broadcasting Corporation) and RTÉ (Raidió Teilifís Éireann). It contains audio recordings of interviews and speeches from a variety of radio broadcasts. The earliest recording is from 1953 and the most recent from 2010. The age profile of the speakers ranges from 19 at the time of the first recording to 96 at the time of the last recording.

The amount of material available for each speaker is varied. For two speakers (one male and one female) from the BBC archives, there are recordings for every 2-3 years over the entire time span. For the remainder of the speakers in the database we have compiled recordings approximately 10 years apart. To minimise any large noise and channel variations, the spectral content of the recordings was examined, and a number of early recordings, deemed to vary too greatly from the later recordings in terms of frequency content, were discarded. In addition to our ageing database, for background modelling two other data sources were used; the TIMIT corpus [10] and the 'University of Florida Vocal Aging Database 2 - Extemporaneous' (UFvadEX) [11].

3 The Speaker Verification System

A Gaussian Mixture Model and Universal Background Model (GMM-UBM) system, as introduced by Reynolds [12] was used in this work. A gender-independent UBM is first created. This is a GMM trained using the Expectation-Maximisation (EM) algorithm using data from a large population of speakers. The individual speaker models are then generated by Bayesian adaption of the UBM. In this work, a 1024 mixture UBM (as in [12]) was generated from 1 hour of speech taken in equal amounts from TIMIT and UFvadEX. The UBM data was carefully composed to avoid biasing it towards any of the speakers or recording channels. Rosenberg [13] notes that a UBM composed with gender-balanced speech with recording conditions matching the test conditions achieves good performance. We applied this finding to our database by ensuring our UBM contained age-balanced as well as gender-balanced data. Age balanced data was retrieved by taking equal amounts of speech from the following age profiles: under 35, 36-55, over 55 (the ages of speakers are given in the documentation of both databases). As our database covers a range of 40 years, it inherently contains a variety of recording conditions. To reflect this variation in our UBM, we used data from both TIMIT and UFvadEX, where TIMIT data consists of clean recordings of scripted speech and UFvadEX contains conversational speech over a wide variety of channels and speaking styles. Composing the UBM content in this way was an effort to ensure that it contained a balanced variety of recording conditions, phonetic content, accents, ages and genders.

Front end processing of the speech consisted of downsampling to 16kHz, energy-based silence removal, and pre-emphasis. 12-dimensional Mel-Frequency Cepstral Coefficients (MFCCs) were extracted over a 20ms windows with 50% overlap. Mean and variance normalisation was applied after RASTA filtering [15]. GMMs for each speaker were trained by adaption [12] of the UBM using 30 second segments of data. During testing, the likelihood of the test data given both a speaker's GMM and the UBM were calculated. Scoring was then done using the standard likelihood ratio framework [14], by subtracting the log likelihood score of the UBM from the log likelihood score of the speaker model.

4 Experimental Study

To uncover any effects of vocal ageing on the speaker verification system described in Section 3, several experiments were conducted on our ageing database. Our aim was to address several questions of interest:

1. How does a speaker's verification score change as the test data moves further away in time from the data on which the model was trained?
2. Is this trend consistent across different feature sets?
3. Accounting for inter and intra-session variability, is any trend in Question 1 significant?
4. Does the age of the speaker at time of model generation influence a long-term trend?

4.1 Long-Term Speaker Verification

The first experiment was designed to answer Question 1 above. Two models were trained for each speaker, one using 30 seconds of data from their first year of available speech and the other using 30 seconds of data from their last year of speech. 'Forward' testing was done by testing each speaker's first model with data from all subsequent years of their speech material. 'Reverse' testing was done by testing each speaker's last model with data from all previous years of their material. Each test score was generated by computing the log likelihood ratio for three separate 30 second segments and taking the average. An initial assumption is made that performance degrades linearly with time and thus a linear least squares fit was computed for each speaker's scores. The test scores along with their line fits for each of the 13 speakers (from 'QUEEN' to 'PLOMLEY', as indicated by the legend) are given for the forward direction in Fig 1 and the reverse direction in Fig 2. The average of the speaker line slopes in the forward and reverse directions are -0.011 and 0.015 respectively. It is evident that there is a significant degradation of verification score that is reasonably consistent across speakers in both forward and reverse testing.

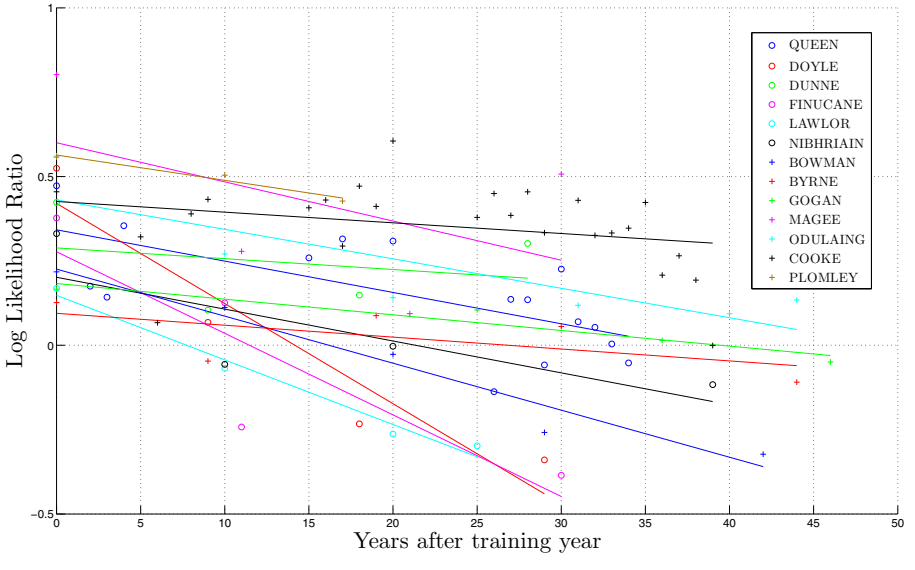

Fig. 1. Long-term verification, testing forward in time, with MFCC features

Speaker verification systems typically incorporate temporal information by taking first and second order time derivatives of the feature vector (referred to as delta and double-delta coefficients) and appending them to the original feature vector. Including dynamic information in this way has been shown to improve verification accuracy [16]. Our experiment above was repeated using MFCCs with both delta and double-delta coefficients appended. Deltas and double-deltas were extracted as time differences over a window of length ±2 samples. Results

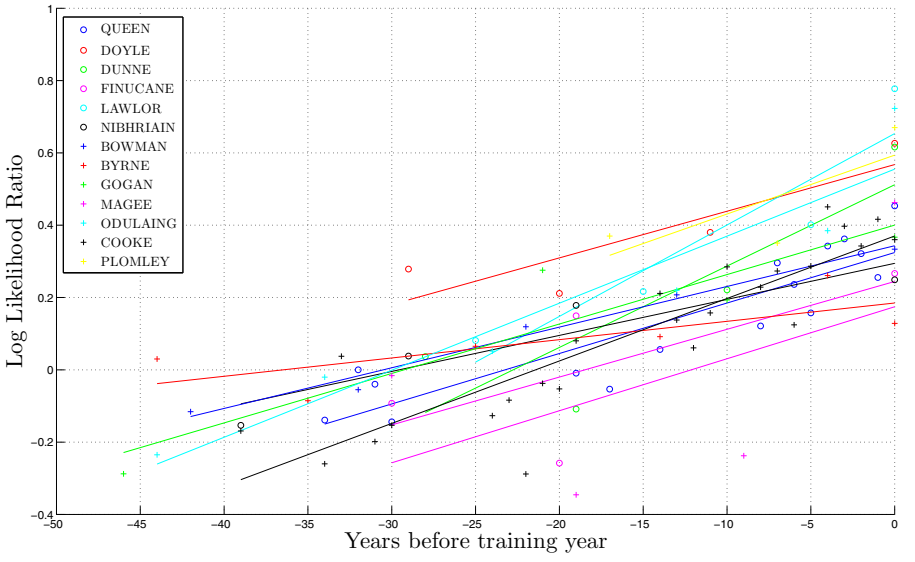

Fig. 2. Long-term verification, testing backwards in time, with MFCC features

of the log likelihood score versus age are plotted for the forward direction for MFCCs with delta coefficients in Fig 3 and delta & delta-delta coefficients in Fig 4. A trend consistent with 1 is seen in these results. The average of the speaker slopes in Fig 3 and 4 are -0.026 and -0.030 respectively. Testing in the reverse direction yields average slopes of 0.027 and 0.036. Thus the rate of decrease of verification score increases progressively with the addition of temporal information.

For comparison to MFCCs, an alternative feature set, Perceptual Linear Predictive (PLP) [17] coefficients were extracted. In [18], it is suggested that there is no clear advantage to using PLPs over MFCCs. However, it has been observed that MFCCs can outperform PLPs in clean conditions, while PLPs offer better performance in noise [21]. The long-term verification experiment was rerun using 12-dimensional PLPs extracted over 20ms windows with 50% overlap. The resulting scores are very similar to those using MFCC results, with forward and reverse slopes of -0.016 and 0.021 respectively. Based on these initial results, MFCCs (without dynamic coefficients) were used exclusively for subsequent experiments.

4.2 Comparison with Inter and Intra-session Variability

The results presented in Section 4.1 demonstrate a consistent decrease in verification score as the time span between training and testing grows. Caution must be observed before attributing this effect solely to ageing however. In [9], Lawson concluded that the influence of 'long-term' ageing of 3 years on speaker verification scores was consistent with simple inter-session variability. Degradation due

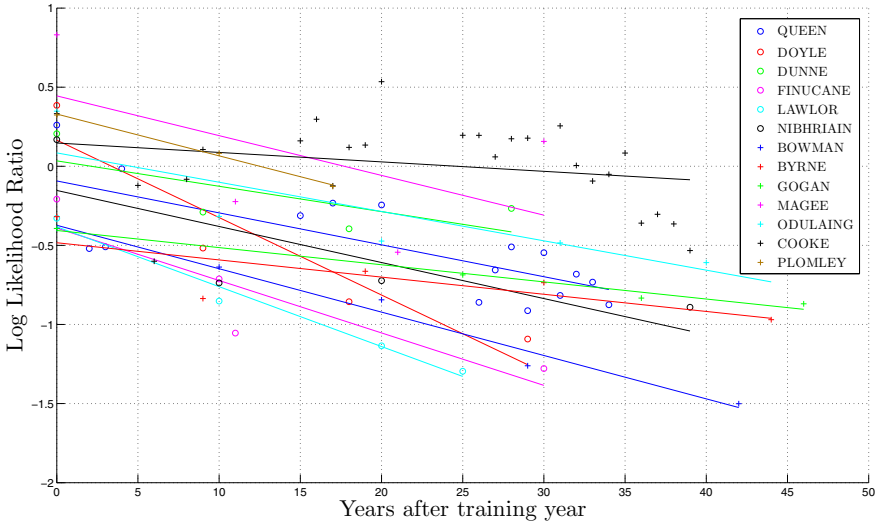

Fig. 3. Long-term verification, testing forward in time, with MFCC features + first order dynamic coefficients

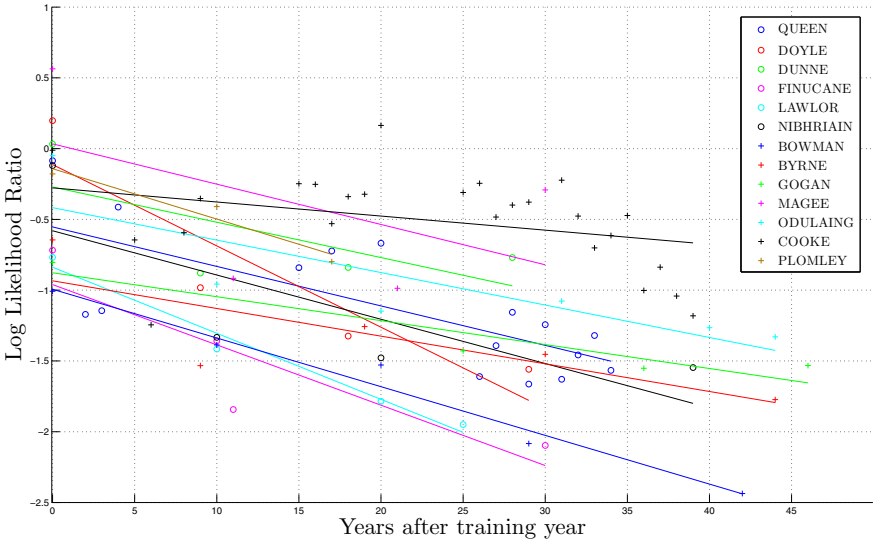

Fig. 4. Long-term verification, testing forward in time, with MFCC features + first and second order dynamic coefficients

to inter-session variability was demonstrated on the MARP corpus by [19]. Similarly, in [20], it is mentioned that results presented on NIST-SRE '05 showing a fall in verification accuracy over a period of one month is more attributable to variabilities other than ageing.

Our second experiment was designed to compare intra and inter-session variability with the potential ageing effect uncovered in Section 4.1 and answer Question 3 above. As short-term inter-session data (recordings from different sessions within a given year) was available for one speaker only, Alistair Cooke, we based our analysis on his speech only.

Short-term inter-session scores were obtained by training a model for each session with 30 seconds of data and testing it against 30 second segments from all other sessions in that year. Intra-session scores were found by training a model with the first 30 seconds of a session and testing it with all subsequent 30 second segments from that session. This was done for all sessions. Long-term inter-session scores were generated by training a model for each session with 30 seconds of data and testing it with 30 second segments from all other sessions across all years. The score distributions of these three sets of results are given in Fig 5. As expected, intra-session scores and long-term inter-session scores at a time span of 0 years are closely aligned. Short-term inter-session scores lie below this range. Interestingly, long-term inter-session scores at a time span of 5 years occupy a similar range to short-term inter-session results. This agrees with previous findings that ageing effects of ≈ 3 years are insignificant compared to normal inter-session variability.

At time spans of 10, 20 and 30 years however, the verification score distribution shifts progressively downwards, beyond the range of the short-term inter-session score distribution. This supports the existence of a negative long-term (> 5 years) effect of vocal ageing on speaker verification.

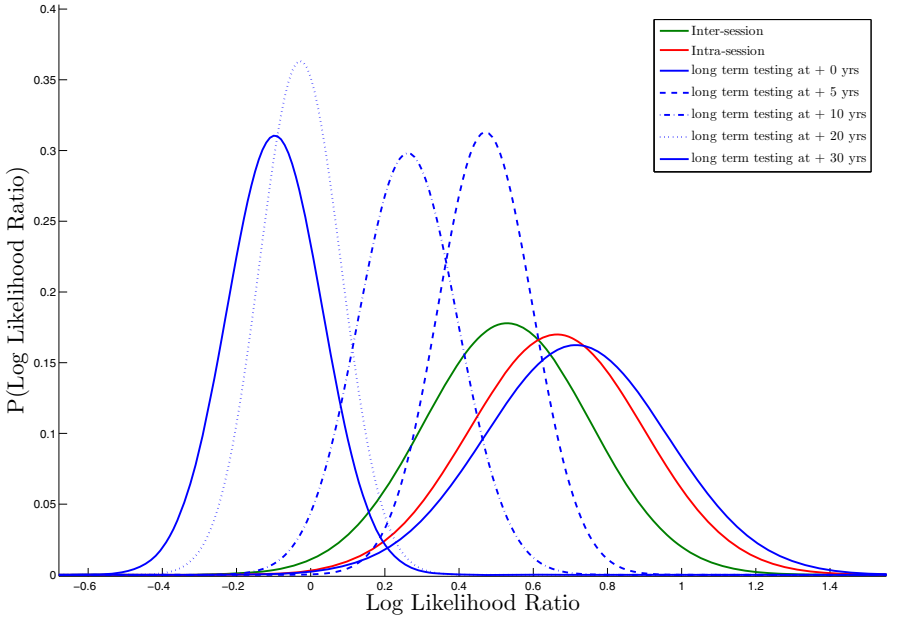

Fig. 5. Distributions of inter/intra-session and long-term verification scores for Alistair Cooke

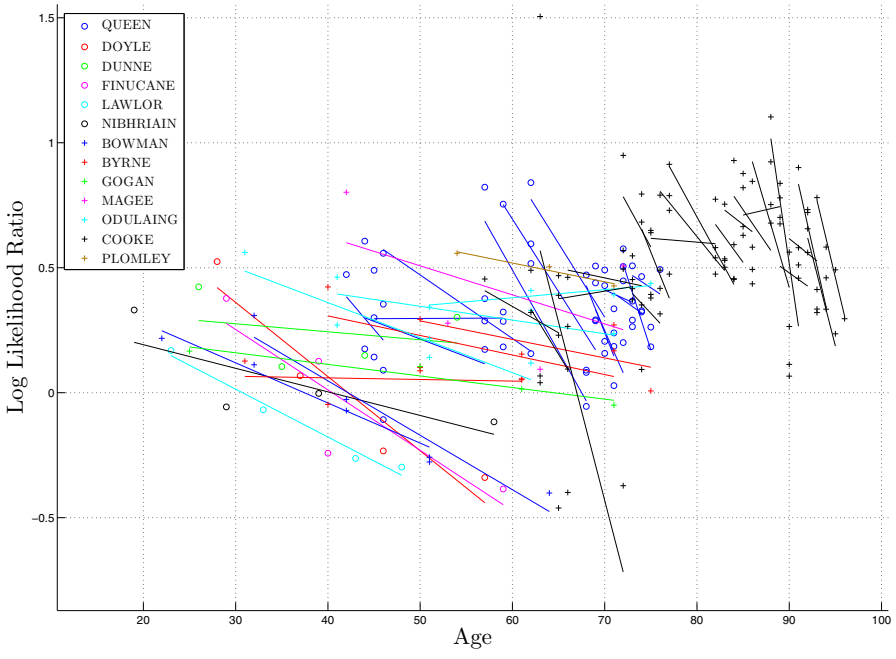

Fig. 6. Long-term verification results, testing forward in time, for individual speakers over multiple age ranges. Taking the speaker 'BYRNE' (symbol +) as an example: three lines are plotted for this speaker, each representing the score trend over different intervals. The first line is fitted to the scores of this speaker's model trained at age 31 and tested with data from age 31, 40, 50 and 61. The second is fitted to the scores of the model trained at age 40 and tested with data from age 40, 50, 61 and 71. Finally, the third is fitted to the model trained at age 50 and tested with data from age 50, 61, 71 and 75.

4.3 Age Dependent Long-Term Speaker Verification

In Section 4.1 we had modelled the drop in verification score between a speaker's first and last recordings as a linear relationship. In reality however, vocal ageing is not constant over time. One of the indicators of an 'elderly' voice is its variability (in pitch, intensity etc) relative to a young speaker [2]. It would be expected then, that the drop in verification scores would be somewhat dependent on the age of the speaker. Furthermore, the onset of vocal changes and the degree of change varies between individuals [1]. We would expect to see evidence of this in verification scores.

To investigate these issues, and address Question 4, the experiment in Section 4.1 was repeated over multiple time spans. A model was trained using data from year 1 and tested with data from year 1 to year $1+N$. A new model was created with data from year 2 and tested with data from year 2 to $2+N$ and so on. N was taken as 3. Note that N was not the span in years, but rather the span

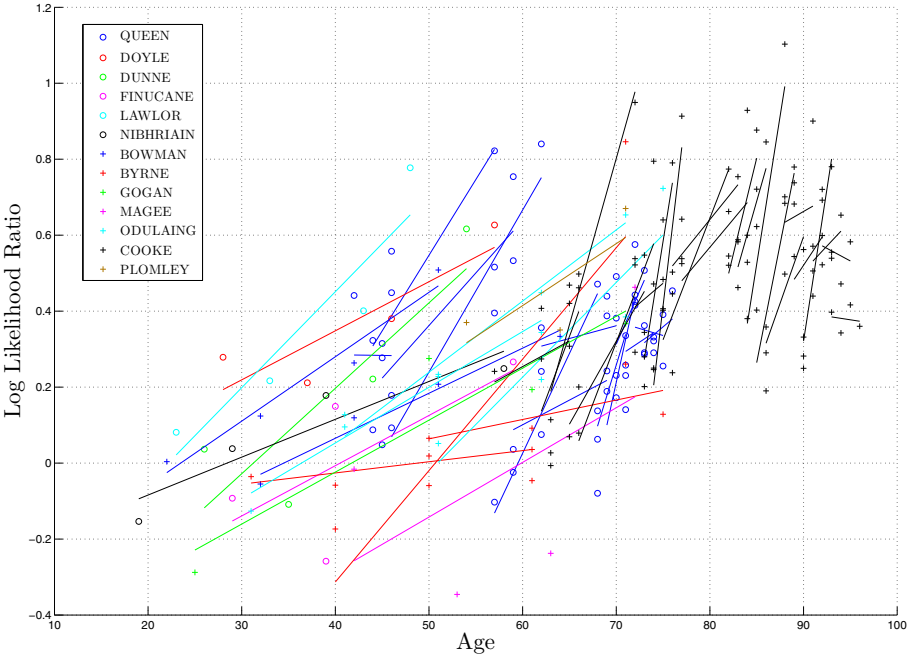

Fig. 7. Long-term verification results, testing backwards in time, for individual speakers over multiple age ranges

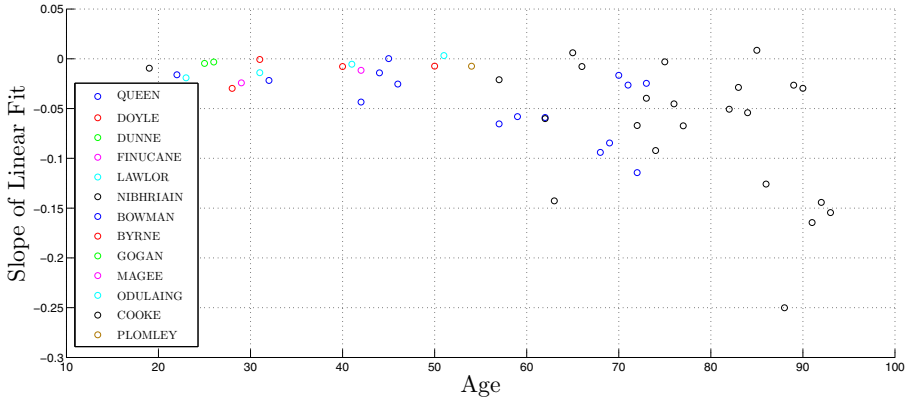

Fig. 8. Slopes of line fits over age ranges in Fig 6

in *available* years of data for a speaker. This was also done in reverse, testing a model from the most recent year Y with data from year Y to $Y - N$, and so on. This was done for all 13 speakers. Results for each of the speakers are presented in Fig 6 and 7. Again, the assumption is made that the score degradation across $N + 1$ points can be approximated linearly.

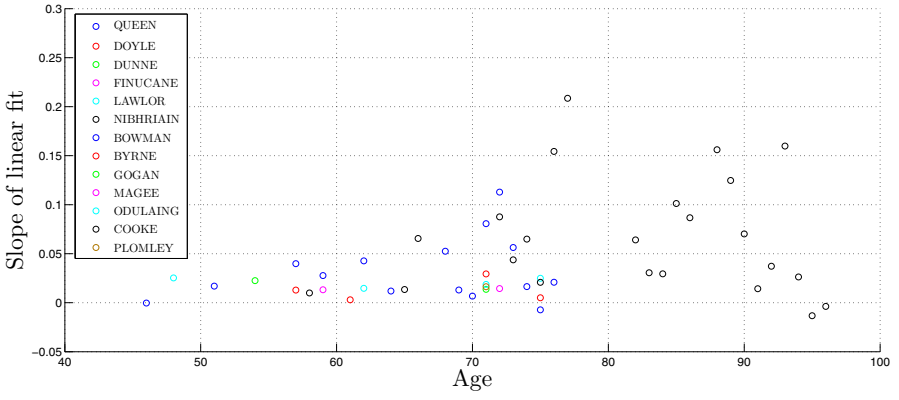

Fig. 9. Slopes of line fits over age ranges in Fig 7

While there are some outliers, a trend emerges in which speaker's models experience a sharper drop off in verification score as their age increases. This change is non-linear, with age of 60 appearing to be a turning point after which the rate of decrease of verification score increases. For clarity, the slopes of each line plot in Fig 6 and 7 are plotted against age in Fig 8 and 9.

5 Conclusions

In this work, we have presented some early-stage results on the effect of ageing on speaker verification.

We have shown that there is a degradation in verification score as the time span between model training and testing increases. This trend is consistent in forward and reverse directions. This behaviour agrees with expectations based on physiological research around vocal ageing. We found little difference in using either MFCC or PLP features.

Including temporal information in the extracted features increases the rate of verification score degradation. As noted in the introduction, a major change in the voice with age is a change in the rate of speech production. Therefore incorporating temporal coefficients, which capture rate information, leads to a fall in accuracy. This introduces an interesting dilemma for building a speaker verification system. In the short-term, temporal information has been shown to increase accuracy, as it captures person-specific information. However, this trait is far less robust to ageing. It is conceivable that other features, such as those derived from pitch and energy, which are advantageous in the short-term, will be similarly detrimental in the long-term.

A major issue in speaker verification is session variability. As discussed, previous studies have considered speaker ageing as insignificant compared with normal inter-session variabilities. We have attempted to separate the effects of session variability from a longer term ageing effect. Our experiment shows how

score variation over a time span of up to 5 years lies within the range of short-term inter-session variability. At greater time spans, of 10, 20 and 30 years, this score distribution shifts outside the expected inter-session variation. This demonstrates a clear effect of vocal ageing outside the realm of normal inter-session variability. This has obvious implications for the life cycle management of biometric templates.

Our final experiment showed the effect of ageing is not constant across all ages. A greater rate of score degradation is seen in older speakers. Based on our limited database, an acceleration in score drop-off is seen above the age of 60. While the degree of vocal change and the time of onset differs between individuals, changes in the voice become generally more marked in older speakers. This is reflected in the increased score variability of older speakers in our examination.

Future work will incorporate a larger cohort of speakers and consider feature sets which are more robust to the changing voice. Different modelling approaches, particularly concerning the UBM composition and training strategy should also be considered.

Acknowledgements

The authors would like to thank James Harnsberger and Rahul Shrivastav, University of Florida, for providing the UFvadEX database.

References

1. Mueller, P.B.: The Aging Voice. Seminars in speech and language 18(2), 159–168 (1997)
2. Linville, S.E.: Vocal aging. Current Opinion in Otolaryngology & Head and Neck Surgery 3, 183–187 (1995)
3. Linville, S.E.: The Sound of Senescence. Journal of Voice 10(2), 190–200 (1996)
4. Sataloff, R.T.: Vocal aging. Current Opinion in Otolaryngology & Head and Neck Surgery 6, 421–428 (1998)
5. Reubold, U., et al.: Vocal aging effects on F0 and the first formant: A longitudinal analysis in adult speakers. Speech Communication 52, 638–651 (2010)
6. Vipperla, R., et al.: Ageing Voices: The Effect of Changes in Voice Parameters on ASR Performance. EURASIP Journal on Audio, Speech, and Music Processing (2010)
7. Cole, R., et al.: The CSLU speaker recognition corpus. In: Proceedings of the International Conference on Spoken Language Processing, pp. 3167–3170 (1998)
8. Lawson, A.D., et al.: The Multi-Session Audio Research Project (MARP) Corpus: Goals, Design and Initial Findings. In: INTERSPEECH 2009, Brighton (2009)
9. Lawson, A.D., et al.: Long term examination of intra-session and inter-session speaker variability. In: INTERSPEECH 2009, Brighton, United Kingdom (2009)
10. Garofolo, J.S.: TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium, Philadelphia (1993)
11. Harnsberger, J.D., et al.: Modeling perceived vocal age in American English. To be presented at Interspeech 2010 (2010)

12. Reynolds, D.A., et al.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing 10, 19–41 (2000)
13. Rosenberg, A.E., et al.: Speaker background models for connected digit password speaker verification. In: ICASSP 1996 (1996)
14. Bimbot, F., et al.: A Tutorial on Text-Independent Speaker Verification. EURASIP Journal on Applied Signal Processing 4, 430–451 (2004)
15. Hermansky, H., et al.: RASTA processing of speech. IEEE Transactions on Speech and Audio Processing 2, 578–589 (1994)
16. Furui, S.: Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Transactions on Acoustics, Speech and Signal Processing 29(3), 342–350 (1981)
17. Hermansky, H., et al.: Perceptual Linear Predictive (PLP) Analysis-Resynthesis Technique. In: IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics, Final Program and Paper Summaries, pp. 37–38 (1991)
18. Kinnunen, T., et al.: An overview of text-independent speaker recognition: From features to supervectors. Speech Communication 52, 12–40 (2010)
19. Lawson, A.D., et al.: External factors influencing the performance of speaker identification of the multisession audio research project (MARP) corpus, 153rd Meeting of the Acoustical Society of America (June 2007)
20. Campbell, J.P., et al.: Forensic speaker recognition. IEEE Signal Processing Magazine 26(2), 95–103 (2009)
21. Kinnunen, T.: Optimizing Spectral Feature Based Text-Independent Speaker Recognition, PhD thesis, Department of Computer Science, University of Joensuu (2005)