Evaluation and calibration of short-term aging effects in speaker verification

Finnian Kelly, John H.L. Hansen

Center for Robust Speech Systems (CRSS),
University of Texas at Dallas, Richardson, Texas, USA
{finnian.kelly, john.hansen}@utdallas.edu

Abstract

A speaker verification evaluation is presented on the Multi-session Audio Research Project (MARP) corpus, for which speakers were recorded at regular intervals, in consistent conditions, over a period of three years. It is observed that the performance of an i-vector system with probabilistic linear discriminant analysis (PLDA) modelling decreases progressively, in terms of both discrimination and calibration, as the time intervals between train and test sessions increase. For male speakers, the equal error rate (EER) increases from 2.4% to 4.4% when the interval between sessions grows from several months to three years. An extension to conventional linear score calibration is proposed, whereby short-term aging information is incorporated as an additional factor in the score transformation. This new approach improves discrimination and calibration performance in the presence of increasing time intervals between train and test sessions, compared with score-only calibration.
Index Terms: speaker verification, speaker recognition, calibration, longitudinal variability, aging.

1. Introduction

The performance of automatic speaker verification on challenging datasets has improved consistently in recent years, largely driven by the NIST speaker recognition evaluation (SRE) series [1]. Current speaker verification systems, many of which are based on an i-vector probabilistic linear discriminant analysis (PLDA) framework [2], have demonstrated robustness across varying microphone, channel, noise, and room-acoustic types. However, there remain sources of variability that have not been assessed in large-scale evaluations, including longitudinal speaker variability, i.e. the change in the voice due to aging.

Variability in the voice exists at multiple levels: within-conversation, due to communication dynamics between speakers, for example [3]; from day-to-day, due to changes in emotion, physical and mental health [4]; from months-to-years, due to the progressive physiological effects of aging [5, 6], and other factors, such as the effects of long-term health and lifestyle [5, 7], and geographical mobility [8].

Previous studies on longitudinal variability in speaker recognition [7, 9, 10, 11] considered the effect of long-term aging, across time intervals ranging from 7 to 60 years. Several proposals to compensate for aging-related performance degradation, at the score [11] and model [10] levels, were proposed.

In this study, the focus is on short-term aging, across time intervals of several months to three years. Intervals within this range are of direct relevance to real-world operating scenarios, particularly in the application of speaker recognition to forensics, where samples under comparison are typically separated by time intervals in range of several months [7].

To enable the study of short-term aging, the Multi-session Audio Research Project (MARP) corpus [12] is utilized. MARP contains speech recorded at regular intervals over a three year time span. Recordings were collected in controlled conditions to minimize external sources of variability.

Previous studies to analyse aging variability in MARP found that there was a weak correlation between speaker recognition scores and the time-elapsed between training and test [13], and that within-session variability of speaker models exceeded that of longitudinal variability [14].

This study expands significantly on [13, 14], by assessing the impact of three-year aging on speaker verification performance with a carefully designed evaluation protocol, rather than score or model similarity, by considering a much greater number of trials, and by utilizing a current i-vector PLDA speaker verification system.

Alongside discrimination performance, calibration performance also considered in this study. The calibration of a system reflects its ability to make good decisions for a range of applications, and is therefore an important measure in practice [15, 16, 17]. An evaluation of the calibration performance across short-term aging leads to the proposal of several Quality Measure Functions (QMFs) [18], which improve upon conventional score calibration by utilizing aging information.

2. MARP corpus

The Multi-session Audio Research Project (MARP) Corpus [12] was collected as a resource for the study of speaker variability across multiple sessions. Over the course of three years, 21 sessions were recorded at regular intervals of 1–2 months. Data was released for all but the 1st and 6th sessions. A total of 73 speakers (46 male, 27 female) contributed, although not all speakers were present at all sessions. The recordings were collected in a soundproof booth using headset microphones. The recording environment and equipment remained consistent throughout. At each session, a range of read, whispered and conversational speech was elicited. Only the conversational portion of the corpus is considered in this study. For the conversational task, pairs of speakers were instructed to converse freely on any subject for 10 minutes. Recordings for both speakers in a session were made simultaneously. However, each speaker was recorded on a separate channel, with no cross-talk between channels. Some speakers had the same partner for all sessions, while others had several partners. The data was released as 8 kHz, 16 bit, raw mono audio.

This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen

3. Speaker verification system

A standard i-vector [19] speaker verification system with probabilistic linear discriminant analysis (PLDA) modeling [20] was used for the experiments in this study. At the front-end, 13-dimensional MFCCs were extracted over 20 ms windows at 10 ms intervals, and appended with first and second derivatives. Mean and variance normalisation and RASTA filtering were subsequently applied. Speech activity detection (SAD) was applied via Combo-SAD [22], which leverages a combination of speech voicing and spectral flux features.

Gender-dependent UBMs of 1024 components were trained with a subset of NIST SREs 2008 and 2010 consisting of microphone speech from speakers of US English. A maximum of 30 seconds of speech post-SAD was included from each session. Approximately 30 hours of data were allocated for training male and female UBMs respectively. An i-vector extractor matrix T of rank 400 was estimated from the same recordings used to train the UBM. Linear discriminant analysis (LDA) was applied, reducing the i-vector dimensionality to 200. Finally, the i-vectors were mean and length normalized, and whitened. The speaker- and session-dependent i-vector distribution was modelled with PLDA, again using the UBM development data.

4. Speaker verification evaluation

From the full set of MARP speakers, a reduced set of 35 males and 25 females was formed by discarding those speakers with less than five sessions of data. In the resulting set, the average number of sessions per-speaker was 14. Since each session was a conversation side of 10 minutes duration, the quantity of active speech for a given speaker was typically less than five minutes. Each session was therefore divided into 'chunks' of exactly 60 seconds duration post-SAD. Fixing the length of these chunks removes the effect of duration [18, 23] from the subsequent experiments. Between one and four 60-second chunks were extracted from each session (dependent on the quantity of active speech). Chunks were extracted such that they were maximally dispersed throughout the session.

An 'all-vs-all' evaluation protocol was adopted, whereby all chunks were considered, in turn, as training data, with the remainder of same-gender (different-session) chunks serving as testing data. Performance was then evaluated according to the absolute difference (in time) between the sessions from which the training and testing chunks were drawn. We define the absolute session difference (ASD), for a given trial, as the absolute difference between the indices (2–21) of the corresponding train and test sessions. Each ASD unit therefore represents a time interval of 1–2 months.

Figure 1 displays the number of target trials for all female speakers in the all-vs-all protocol, across all ASDs. The distribution of target trials across ASD is similar for the 35 male speakers. It is clear that the number of target trials varies between speakers, and decreases progressively with ASD. The imbalance in the number of trials per-speaker and per-ASD is accounted for at the performance analysis stage via a trial weighting scheme [24, 9], whereby the trials corresponding to a particular speaker-ASD combination are weighted by the inverse of the total number of trials of that combination.

4.1. Experimental results

In Figure 2, Detection Error Trade-off (DET) curves are plotted for the full set of male and female trials. The associated Equal Error Rates (EERs) are provided. The effect of trial weighting

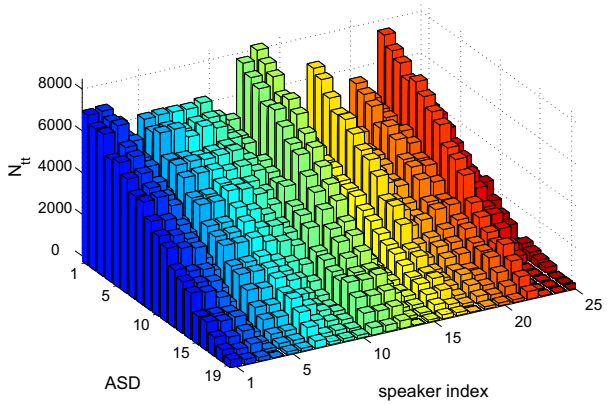

Figure 1: The number of target trials N_{tt} for each of the 25 female speakers across absolute session difference (ASD) with an all-vs-all evaluation protocol.

is demonstrated by the shift in the DET curve after the application of speaker, ASD, and speaker-ASD (both speaker and ASD) trial weighting. All subsequent performance metrics in this paper are based on speaker-ASD weighted trials.

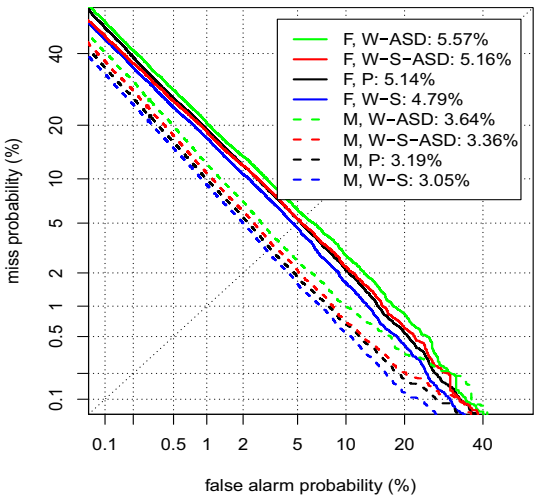

Figure 2: DET curves for 'F' (female) and 'M' (male) speakers for the full set of trials. 'P': pooled (unweighted), 'W-S': speaker-weighted, 'W-ASD': ASD-weighted, 'W-S-ASD': speaker- and ASD-weighted. EER% values are indicated in each case.

A breakdown of performance over different ASD ranges is shown in Table 1. For males, the EER increases progressively with ASD, from 2.44% to 4.35%. Aside from the EERs at ASD ranges 8–11 and 12–15 being close in value, this increase is approximately linear. The rate of EER increase is similar in the female case, and aside from ASD range 8–11, is also approximately linear. The difference in EER between the first and last ASD ranges is significant: a 78% and 56% relative increase for males and females respectively.

As the time elapsed between training and testing increases, the overall discrimination ability of the speaker verification system decreases. It is for this reason that the EER increases after ASD weighting, Figure 2, due to the bias toward trials with smaller ASDs being removed.

	ASD range	1-19	1-3	4-7	8-11	12-15	16-19
male	EER%	3.36	2.44	3.02	3.47	3.50	4.35
	$N_{tt} \times 10^4$	7.6	2.3	2.4	1.7	0.8	0.4
	$N_{nt} \times 10^6$	2.5	0.7	0.8	0.6	0.3	0.1
female	EER%	5.16	4.36	4.78	4.55	6.21	6.87
	$N_{tt} \times 10^4$	5.9	1.9	1.9	1.3	0.6	0.2
	$N_{nt} \times 10^6$	1.3	0.4	0.4	0.3	0.1	0.1

Table 1: EER% for subsets of trials grouped according to absolute session difference (ASD). N_{tt} and N_{nt} denote the number of target and non-target trials respectively.

5. Calibration evaluation

While the EER is an effective summary of the discrimination of the system, it is also important to evaluate the calibration of the system [16]. In this section, the effect of longitudinal speaker variability on calibration is assessed, focusing on the case of male speakers. A suitable performance metric is the log-likelihood ratio cost, C_{llr} [25], which provides a measure of calibration over all effective priors. An associated measure is C_{llr}^{min}, the minimum value of C_{llr} obtained via an optimal score transformation [16, 25]. The C_{llr} is a measure of both discrimination and calibration, while C_{llr}^{min} is a measure of discrimination only. Thus, a measure of pure calibration can be defined as: $R_{mc} = \left(C_{llr} - C_{llr}^{min}\right) / \left(C_{llr}^{min}\right)$, where R_{mc} is the relative miscalibration [18].

5.1. Score calibration

To convert raw (uncalibrated) scores s output from the speaker verification system into linearly calibrated likelihood ratios x, the following function was applied:

$$x = w_0 + w_1 s \tag{1}$$

where w_0 and w_1 are offset and scaling parameters respectively. Given a set of development data, w_0 and w_1 were optimized via logistic regression. This optimization relies on the assumption that the score distributions of the development and test datasets are similar. Due to the unique (longitudinal) nature of the MARP corpus, there are no other readily available databases suitable for this optimization. Thus, for the first set of calibration experiments in this section, offset and scaling parameters are optimized on the test scores directly (self-calibration). In a follow-up experiment, independence between development and test datasets is introduced by adopting both cross-validation and cross-gender approaches.

To assess the impact of longitudinal speaker variability on conventional score calibration, Equation 1, three scenarios are considered:

- **Full (F):** one set of calibration parameters is optimized on the full set of scores, i.e. from all ASDs.
- **Mismatched (MM):** one set of calibration parameters is optimized on scores in the ASD range 1–3.
- **Matched (M):** five sets of calibration parameters are optimized: one set from the scores of each ASD range considered in Table 1.

5.2. Aging calibration

Since a relationship was observed between ASD and speaker verification performance, in terms of EER, Table 1, here we propose to extend the score calibration given in Equation 1 by incorporating aging information (i.e. ASD) as an additional term. We adopt the approach introduced in [18], which incorporated recording duration information in calibration via 'Quality Measure Functions' (QMFs). The extended score calibration is given by:

$$x = w_0 + w_1 s + Q\left(w_2, ASD, ...\right) \tag{2}$$

Where Q denotes a QMF defining the way in which ASD (and any additional parameters) are incorporated into the calibration. w_2 is a new calibration parameter to be optimized on the development set.

To inspire suitable functions Q for aging calibration, the evaluation scores were first calibrated according to the 'Matched' scenario (Section 5.1). However, in optimizing the parameters at each ASD range, the scaling parameter, w_1, was fixed, and the offset parameter w_0 was allowed to vary. This process was similar to the 'shared scaling' experiments presented in [18]. The resulting offset parameters for each ASD range, and each individual ASD, are plotted in Figure 3. Three candidate QMFs are proposed in Table 2 to approximate the relationship between offset parameter and ASD in Figure 3.

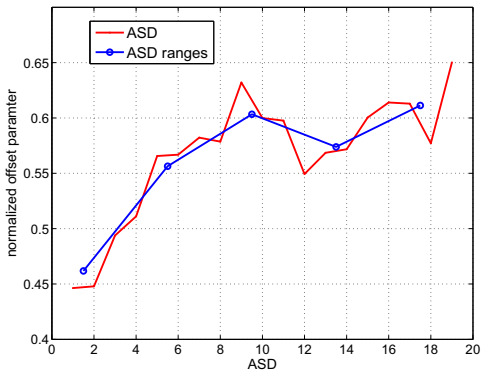

Figure 3: Normlized offset parameters w_0 for individual ASDs and ASD ranges.

QMF: $Q_n\left(w_2, ASD, ...\right)$	additional parameters
$Q_1 = w_2 ASD$	w_2
$Q_2 = w_2 \log\left(ASD\right)$	w_2
$Q_3 = w_2\left(1 - e^{-\lambda . ASD}\right)$	w_2, λ

Table 2: Proposed quality measure functions (QMFs) for incorporating aging information, i.e. ASD, as an additional parameter in conventional score calibration. λ was set at 0.1 in our experiments.

5.3. Experimental results

The evaluation scores of male speakers were self-calibrated according to the three score calibration scenarios in Section 5.1: Full, Matched and Mismatched, via Equation 1, and according to the three aging calibration proposals, Table 2. Performance in terms of C_{llr} and $R_{mc}\%$ is given in Table 3 for each of the five ASD ranges.

Expectations for the score calibration proposals are for the 'Mismatched' (MM) to perform worst, as it uses a limited range of ASDs for parameter estimation, and for 'Matched' (M) to perform best, as it represents an ideal (but unrealistic) case where data is available to optimize separate parameters at each ASD range. The performance of 'Full' (F) is expected to fall somewhere in between, as it uses the full range of ASDs to optimize one set of parameters. It can be seen from the first three rows of Table 3, that this expected behaviour is observed. In all cases, discrimination error and miscalibration increase with ASD range.

The bottom three rows of Table 3 indicate the performance of the three QMF proposals. At all but the first ASD range, all decrease error relative to the 'Mismatched' and 'Full' cases, and come close to the 'Matched' performance. The relative improvement with each QMF increases with ASD: at the 16-19 range, a 37% reduction in R_{mc} is achieved with Q_1 relative to the 'Full' case. Although Q_1 performs slightly better than the other QMFs, there is little difference between their performance.

	ASD range	1-3	4-7	8-11	12-15	16-19
MM	C_{llr}	.094	.121	.150	.153	.226
	$R_{mc}\%$	2.17	7.08	13.64	15.04	34.54
F	C_{llr}	.099	.116	.136	.139	.195
	$R_{mc}\%$	7.61	2.65	3.03	4.50	16.07
M	C_{llr}	.094	.115	.134	.137	.184
	$R_{mc}\%$	2.17	1.77	1.52	3.01	9.53
Q_1	C_{llr}	.095	.115	.135	.137	.185
	$R_{mc}\%$	3.26	1.77	2.27	2.24	10.12
Q_2	C_{llr}	.095	.115	.135	.137	.187
	$R_{mc}\%$	3.26	1.77	2.27	2.24	11.31
Q_3	C_{llr}	.095	.115	.135	.137	.186
	$R_{mc}\%$	3.26	1.77	2.27	2.24	10.71

Table 3: Performance of score-only calibration approaches: Full (F), Mismatched (MM), and Matched (M) and proposed score-aging calibration approaches Q_1–Q_3.

To evaluate score-aging calibration with parameters optimized on independent data, two approaches were taken. The first was a 10-fold cross-validation (CV) scheme, where 21 male speakers were used for parameter optimization and the remaining 14 males were used for testing. The second was a cross-gender (CG) scheme, where the full set of scores from female speakers were used for parameter optimization, and the full set of male scores for testing. The results of both approaches, and best performing QMFs, are presented in Tables 4 and 5.

The absolute values of C_{llr} and R_{mc} are much larger than in Table 3. In the CV case, this is due to the differences between score and ASD distributions across speakers. In the CG case, the R_{mc} values are an order of magnitude greater than those in Table 3. Along with the difference in score and ASD distributions between genders, this is due to the separate development data used for male and female systems. Although neither scenario provides ideal calibration data, relative reductions in R_{mc} of 12% and 8% at the 16-19 ASD range are obtained in CV and CG cases respectively, by including aging information. Q_1 performs slightly better than Q_2 and Q_3 in the CG case, suggesting it may generalize best across datasets.

	ASD range	1-3	4-7	8-11	12-15	16-19
F	C_{llr}	.105	.123	.155	.151	.229
	$R_{mc}\%$	22.09	13.89	18.32	23.77	51.66
Q_3	C_{llr}	.101	.123	.154	.149	.220
	$R_{mc}\%$	17.44	13.89	17.56	22.13	45.70

Table 4: 10-fold cross validation performance of score-only calibration, Full (F), and a score-aging calibration proposal, Q_3. The mean values across the 10 folds are presented.

	ASD range	1-3	4-7	8-11	12-15	16-19
F	C_{llr}	.282	.434	.579	.611	.843
	$R_{mc}\%\times10^1$	20.65	28.41	33.86	35.94	40.18
Q_1	C_{llr}	.292	.433	.568	.584	.786
	$R_{mc}\% \times10^1$	21.74	28.32	33.30	33.91	36.79

Table 5: Cross-gender calibration, Full (F), and score-aging, Q_1

6. Discussion

This study presented a speaker verification evaluation of controlled three-year longitudinal data. Discrimination and calibration rates were observed to increase progressively over this time interval. An EER increase over an interval as short as 6 months (between ASD ranges 1–3 to 4–7, Table 1) was observed, indicating that this phenomenon is an important consideration for speaker recognition systems in practice. An extension to conventional score calibration, incorporating aging as side-information, was shown to reduce discrimination and calibration error, particularly at larger ASDs.

The drop in discrimination ability of the system is in line with previous studies (e.g. [9]), where proportionally similar decreases in EER were observed over longer time intervals. In [11], a score-aging decision threshold was applied to verification scores using an SVM classifier. The calibration proposal in this study uses aging information in similar way. Thus, a score-level aging compensation approach can be applied in both long- and short-term applications.

Longitudinal studies are inherently limited by data availability. This factor makes it difficult to train calibration parameters independently; the large miscalibration values in the cross-speaker and cross-gender experiments emphasize this point. There are other, limited, sources of longitudinal data: TCDSA [9] and Greybeard [26]. There is scope for combining these datasets for calibration optimization or other compensation schemes.

Inter-speaker differences in longitudinal score trajectories, and resulting error rates, can be observed from the output of the speaker verification system in this study. Longitudinal speaker-dependent behaviour has also been observed in long-term aging studies [7, 27]. The 'Doddington Zoo' effect [28], where some speakers are inherently more difficult to recognize than others, may be compounded when speakers change in different ways over time. The characteristics which set these 'problem' speakers apart are worth investigating.

7. References

[1] C. Greenberg, V. Stanford, A. Martin, M. Yadagiri, G. Doddington, J. Godfrey, and J. Hernandez-Cordero, "The 2012 NIST speaker recognition evaluation," in *InterSpeech 2013*, 2013.

[2] R. Saeidi, K. A. Lee, T. Kinnunen, T. Hasan, B. Fauve, P.-M. Bousquet, E. Khoury, P. L. S. Martinez, J. M. K. Kua, C. You, H. Sun, A. Larcher, P. Rajan, V. Hautamki, C. Hanilci, B. Braithwaite, G.-H. Rosa, S. O. Sadjadi, G. Liu, H. Boril, N. Shokouhi, D. Matrouf, L. El Shafey, P. Mowlaee, J. Epps, T. Thiruvaran, D. Van Leeuwen, B. Ma, H. Li, J.-F. Bonastre, S. Marcel, J. Mason, and E. Ambikairajah, "I4U submission to NIST SRE 2012: a large-scale collaborative effort for noise-robust speaker verification," in *InterSpeech*, 2013.

[3] C. De Looze, S. Scherer, B. Vaughan, and N. Campbell, "Investigating automatic measurements of prosodic accommodation and its dynamics in social interaction," *Speech Communication*, vol. 58, no. 0, pp. 11–34, 2014.

[4] 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.

[5] J. M. Beck, "Organic Variation of the Vocal Apparatus," in *The Handbook of Phonetic Sciences*. Blackwell Publishing Ltd., 2010, pp. 153–201.

[6] S. E. Linville, *Vocal Aging*. Canada: Singular, 2001.

[7] R. Rhodes, "Assessing non-contemporaneous forensic speech evidence: acoustic features, formant frequency-based likelihood ratios and asr performance," *The International Journal of Speech, Language and the Law*, vol. 20, no. 1, pp. 147–150, 2013.

[8] D. Bowie, "The effect of geographical mobility on the retention of a local dialect." Ph.D. dissertation, 2000.

[9] F. Kelly, R. Saeidi, N. Harte, and D. v. Leeuwen, "Effect of long-term ageing on i-vector speaker verification," in *InterSpeech 2014*, Singapore, 2014.

[10] F. Kelly, N. Brümmer, and N. Harte, "Eigenageing compensation for speaker verification," in *InterSpeech 2013*, Lyon, France, 2013.

[11] F. Kelly, A. Drygajlo, and N. Harte, "Speaker verification in score-ageing-quality classification space," *Computer Speech & Language*, vol. 27, no. 5, pp. 1068–1084, 2013.

[12] A. D. Lawson, A. R. Stauffer, E. J. Cupples, W. S.J., W. Bray, and Grieco.J.J., "The multi-session audio research project (MARP) corpus: Goals, design and initial findings," in *InterSpeech 2009*, Brighton, U.K., 2009.

[13] A. D. Lawson, A. R. Stauffer, B. Y. Smolenski, B. B. Pokines, M. Leonard, and E. J. Cupples, "Long term examination of intra-session and inter-session speaker variability," in *InterSpeech 2009*, Brighton, U.K., 2009.

[14] K. W. Godin and J. H. Hansen, "Session variability contrasts in the MARP corpus," in *InterSpeech 2010*, Makuhari, Japan, 2010.

[15] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, "The NIST speaker recognition evaluation overview, methodology, systems, results, perspective," *Speech Communication*, vol. 31, no. 23, pp. 225–254, 2000.

[16] D. A. v. Leeuwen and N. Brümmer, "An introduction to application-independent evaluation of speaker recognition systems," *Speaker Classification*, vol. I, pp. 330–353, 2007.

[17] J. Gonzalez-Rodriguez, P. Rose, D. Ramos, D. T. Toledano, and J. Ortega-Garcia, "Emulating DNA: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2104–2115, 2007.

[18] M. Mandasari, R. Saeidi, M. McLaren, and D. van Leeuwen, "Quality measure functions for calibration of speaker recognition system in various duration conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 11, pp. 2425–2438, 2013.

[19] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verication," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[20] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matejka, and N. Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011*, pp. 4832–4835.

[21] S. O. Sadjadi, M. Slaney, and L. Heck, "MSR identity toolbox v1.0: A MATLAB toolbox for speaker recognition research," Tech. Rep., November 2013.

[22] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[23] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. v. Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *ICASSP 2013*, Vancouver, Canada, 2013.

[24] D. A. v. Leeuwen, "A note on performance metrics for speaker recognition using multiple conditions in an evaluation," Tech. Rep., 9 June 2008.

[25] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 23, pp. 230–275, 2006.

[26] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Greybeard voice and aging," in *Seventh conference on International Language Resources and Evaluation (LREC '10)*, 2010.

[27] H. J. Künzel, "Non-contemporary speech samples: Auditory detectability of an 11 year delay and its effect on automatic speaker identification," *The International Journal of Speech, Language and the Law*, vol. 14, no. 1, pp. 109–136, 2007.

[28] G. Doddington, W. Liggett, A. F. Martin, M. A. Przybocki, and D. A. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES: a statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation," in *International Conference on Spoken Language Processing (ICSLP)*, 1998.