Eigenageing Compensation for Speaker Verification

Finnian Kelly[1], *Niko Brümmer*[2], *Naomi Harte*[1]

[1]Department of Electronic & Electrical Engineering, Trinity College Dublin, Ireland
[2]AGNITIO Research, Somerset West, South Africa
kellyfp@tcd.ie, niko.brummer@gmail.com, nharte@tcd.ie

Abstract

Dealing with the effect of vocal ageing on speaker verification is an important challenge. In this paper, a new approach to improving speaker verification performance in the presence of long-term ageing is presented. Analogous to eigenchannel compensation, the proposed eigenageing compensation method operates by adapting a speaker model to a test sample based on a predetermined ageing subspace. An experimental evaluation of the new method, using the Trinity College Dublin Speaker Ageing database, demonstrates it to be very effective at reducing long-term speaker verification error rates, and shows it to compare favourably with our previous stacked classifier technique.

Index Terms: speaker verification, ageing, eigenanalysis

1. Introduction

The effect of vocal ageing is an important consideration for automatic speaker verification systems. As biometric systems increase in scale and operating duration, the issue of ageing becomes increasingly relevant [1, 2]. The physiological changes of the vocal mechanism, and the associated change in the properties of the voice have been well documented [3, 4, 5, 6, 7, 8, 9]. The effect of vocal ageing on speaker verification has been highlighted as an important issue [10]. Vocal ageing is relevant also to forensic speaker recognition, where there is frequently a need to compare non-contemporary speech samples [11, 12]. Despite this, there has been little research attention devoted to the topic.

In our recent work [13, 14], a longitudinal speaker verification evaluation of 18 speakers over a 30-60 year period was presented. As the time span between enrolment and verification increased, the verification scores of genuine speakers decreased, leading to large error rates. We proposed a solution to the problem by adopting a 'stacked classifier' decision boundary, dependent on ageing progression and quality measures, and demonstrated its effectiveness.

A new approach to speaker verification across ageing, eigenageing compensation, is presented in this paper. The proposed technique compensates for the effect of ageing by adapting a speaker model to a test sample at the verification stage, based on a vocal ageing subspace. The term eigenageing has been adopted as a reference to the closely related eigenchannel compensation [15, 16], which compensates for channel variability by adapting a speaker model to a test sample, based on a channel subspace.

The longitudinal ageing data used for the experiments in this paper is the Trinity College Dublin Speaker Ageing (TCDSA) Database [13], which has recently been expanded to 26 speakers. To our knowledge, this is the largest publicly available longitudinal speaker database, and is freely available for academic research.

An evaluation of the proposed eigenageing approach demonstrates its effectiveness at improving ageing speaker verification performance. A relative reduction in the half total error rate (HTER) of 36% for male speakers and 39% for female speakers is observed after applying eigenageing compensation to a baseline Gaussian mixture model - universal background model (GMM-UBM) system [17]. The new method also outperforms our previous stacked classifier approach, without the need for ageing progression information at the verification stage.

2. Speaker Ageing Data

The speaker ageing data used for the experiments in this paper is an updated version of the Trinity College Dublin Speaker Ageing (TCDSA) Database [13, 14]. The TCDSA database is a longitudinal speaker database with recordings spanning 30-60 years per speaker. For this work, the total number of speakers has been increased from 18 to 26 (15 males and 11 females). The 8 new speakers were sourced from 'The Up series' television documentary series [18]. For each of the new speakers, there are 6 recordings spanning 35 years, at intervals of 7 years. An analysis of the vocal ageing changes of some of these speakers was presented in [19].

Accompanying the main database is a development database for background modelling, 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. As an additional data source, conversational excerpts from the CSLU [20] database were used.

3. Eigenageing Compensation

Eigenchannel compensation was introduced by Kenny et al. [21, 22], and subsequently used by several groups for speaker verification evaluations [15, 16, 23, 24]. The aim of the technique is to model the variability in speaker models that occurs due to changing channel conditions, and to use this to improve speaker verification performance by adapting a speaker model at the verification stage to the conditions of the test sample.

We propose to use a similar idea to compensate for the variability in speaker models that occurs due to ageing. The aim of eigenageing compensation is to model the ageing change in speakers, and then use this to adapt a speaker model at verification time to a sample of unknown age. The adaptation is constrained to an ageing subspace so that the speaker is adapted to the age of the test sample speaker rather than the identity of the test sample speaker. An advantage of this approach over our previous stacked classifier method is that no ageing information (e.g. length of time elapsed since enrolment) is required at verification.

The eigenageing method was integrated into a standard Gaussian mixture model - universal background model (GMM-UBM) system [17], whereby GMMs were trained for each speaker by maximum a posteriori (MAP) adaptation of a UBM. Only the GMM means were adapted. Since all speaker GMMs differ only in their means, each GMM can be fully described by the concatenation of its mean vectors. This representation of a GMM is commonly referred to as a supervector [10, 16]. Before concatenating the GMM means, each was normalized by its corresponding standard deviation [24].

3.1. Ageing Subspace Estimation

The aim of ageing subspace estimation is to find a low-rank matrix that defines the directions of greatest change in the models of ageing speakers.

Given the ageing data in the TCDSA database suitable for this purpose, models were trained for a subset according to several constraints. Several studies have highlighted differences in the way that male and female voices change with ageing [3, 4]. It was therefore considered appropriate to train separate ageing subspaces for males and females. The greatest vocal ageing changes occur in childhood and above the age of 60 [3, 5]. To constrain our ageing subspace to ‘typically ageing adults’, the data used for training was restricted to a subset of speaker recordings within the age range of 20 to 60. Finally, to balance the contribution of different speakers, the number of recordings per speaker was limited to 6, with a 5 to 10 year time-lapse between recordings.

For each recording in this reduced dataset, a GMM was trained and converted to a supervector representation. Unlike implementations of eigenchannel compensation, e.g. [16], the mean supervector of each speaker was not subtracted from the set of that speaker’s supervectors. Instead, a set of ageing difference supervectors were found by subtracting each speaker’s year 1 supervector from their subsequent (i.e. year 2, year 3, $\cdots$, year N, where $N \leq 6$) supervectors. A second and third set of ageing-difference supervectors were found by subtracting each speaker’s year 2 and year 3 supervectors from their subsequent supervectors. The aim of this three stage difference operation was to emphasise the temporal shift in a speaker model that we assume happens with ageing progression.

The set of ageing difference supervectors for each speaker were assembled to form columns of the $MD \times J$ matrix S, where MD is the supervector dimension, a product of the number of GMM components M and the feature dimension D. In this case $MD = 512 * 24 = 12288$. J is the number of ageing difference supervectors, which averaged 100 for males and 80 for females. The ageing subspace matrix V is then given by the R principal eigenvectors of the within speaker covariance matrix $\frac{1}{J}SS^{\top}$ (those corresponding to its R largest eigenvalues). Thus V is of dimension $MD \times R$. In our system, R was taken as 20.

3.2. Eigenageing adaptation

A speaker model s can be ageing-adapted to a set of test features $O = [o_1, o_2, \cdots, o_T]$ by shifting its supervector elements in the directions specified by the ageing subspace matrix V (Section 3.1). In this work, a constrained maximum likelihood (ML) adaptation, as opposed to a maximum a posteriori (MAP) adaptation, is used. This is achieved by finding a low-dimensional vector x such that the following criterion is maximised:

$$p\left(O|s + Vx\right) \tag{1}$$

As shown in [23], x is maximised by:

$$x = A^{-1} \sum_{m=1}^{M} V_m^{\top} \sum_{t=1}^{T} \gamma_m\left(t\right) \frac{o_t - \mu_m}{\sigma_m} \tag{2}$$

where o_t is the tth feature frame, $V_m^{\top}$ is the transpose of the $D \times R$ block of matrix V corresponding to the mth mixture component, $\gamma_m\left(t\right)$ is the probability of occupation of mixture component m at time t. μ_m and σ_m are the component mean and standard deviation vectors. A is given by:

$$A = \sum_{m=1}^{M} V_m^{\top} V_m \sum_{t=1}^{T} \gamma_i\left(t\right) \tag{3}$$

The occupation probabilities $\gamma_i\left(t\right)$ are computed using the test speaker GMM, s. In our experiments, x was initialised as 0 and three iterations of the maximisation in Equation 1 were computed, each time using the updated GMM to calculate $\gamma_i\left(t\right)$. The log likelihood ratio (LLR) score for each trial was then calculated in the usual manner [10], using the UBM and the adapted speaker model s_a translated back into GMM representation:

$$\log p\left(O|s_a\right) - \log p\left(O|UBM\right) \tag{4}$$

4. Baseline GMM system

A Gaussian mixture model - universal background model (GMM-UBM) system [17] was used for the baseline speaker verification experiments. As discussed in [13], the GMM-UBM system has been used for clarity and simplicity in investigating the ageing issue. The unsupervised ageing compensation method proposed in this paper moves in the direction of a factor analysis [25, 15] approach, and we plan to integrate ageing compensation into such a framework in subsequent work. Additional suitable development data will be required to pursue this.

All speech material was pre-processed by downsampling to 8kHz, removing silences with an energy-based voice activity detector and applying pre-emphasis. MFCC features of length 12 were extracted over 20ms windows with 50% overlap and a Mel filterbank of 26 bands. Mean and variance normalization were applied to the features followed by RASTA filtering. Finally, delta coefficients were appended. Further details of all preprocessing and feature extraction steps can be found in the review by Kinnunen and Li [10].

The GMM-UBM system was created by first training a 512 component gender-dependent UBM using 1 hour of data comprised of TCDSA development data (30 minutes) and conversational excerpts from CSLU (3 minutes from each of 10 speakers). 10 iterations of the expectation maximisation (EM) algorithm [26] were applied to train the UBM. Speaker-specific models for each speaker in the TCDSA database were trained by adapting the UBM means with 1 minute of training speech [17]. A relevance factor of 16 was used, and all UBM components were considered in adaptation and scoring.

5. Speaker Verification Experiment

A speaker verification experiment was designed to evaluate the performance of eigenageing compensation. For comparison, the performance of our stacked classifier approach [13, 14] was also evaluated.

5.1. Eigenageing compensation evaluation

To train an ageing subspace matrix, a leave-one-out scheme was used. For every male speaker under test, the remaining TCDSA 14 males were used to train a subspace matrix, as detailed in Section 3.1. Similarly, for every female speaker under test, the remaining 10 were used to train a subspace matrix.

A GMM was adapted from the UBM using 1 minute of data from the test speaker's earliest recording. The remainder of the training recording ('year 1') and all subsequent years of data were used as genuine speaker trials. No age restrictions were placed on test recordings. Thus, the full set of recordings, within the age range 19 to 96 were used (as opposed to the reduced set used for subspace estimation, Section 3.1). For imposter trials, recordings from each of 20 gender-dependent speakers from the CSLU database were used.

For all trials, between one and three 30 second segments (dependent on data availability) were scored against the test speaker GMM and the UBM, and an average log likelihood ratio (LLR) [10] was calculated. Eigenageing adaptation was applied to the test GMM for each trial, and an ageing adapted average LLR was acquired (Equation 4). This test design resulted in a set of genuine speaker LLRs (average of 8 per speaker) and imposter LLRs (20 per speaker) with and without eigenageing compensation.

A Z-normalisation [10] was applied to all trials. Statistics for Z-normalisation were calculated from a set of 10 CSLU speakers on a gender-dependent basis. The subsets of CSLU data used for UBM training, imposter trials and Z-normalisation were all independent, i.e. no overlap of speakers.

To determine a decision threshold, a second leave-one-out scheme was used: for each test speaker, the LLRs from the genuine and imposter trials of all other speakers of the same gender were used to train the threshold. From the set of genuine LLRs, only those from the same year as the training year, 'year 1', were used. This ensured that no ageing information was present in the decision boundary training set.

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 must reach a tradeoff between the FAR and FRR. Here, a threshold was found such that the half total error rate (HTER) on the training set was minimised, where the HTER is the average of the FAR and FRR. Thresholds determined in this way are superimposed on the LLR scores in each plot in Figure 1.

5.2. Stacked classifier evaluation

For comparison with eigenageing compensation, the performance of a stacked classifier approach [13, 14] was evaluated. This approach uses a multi-dimensional decision boundary incorporating ageing information and/or quality measures. The evaluation of the stacked classifier follows the same procedure as the eigenageing experiment, with the difference in the decision boundary.

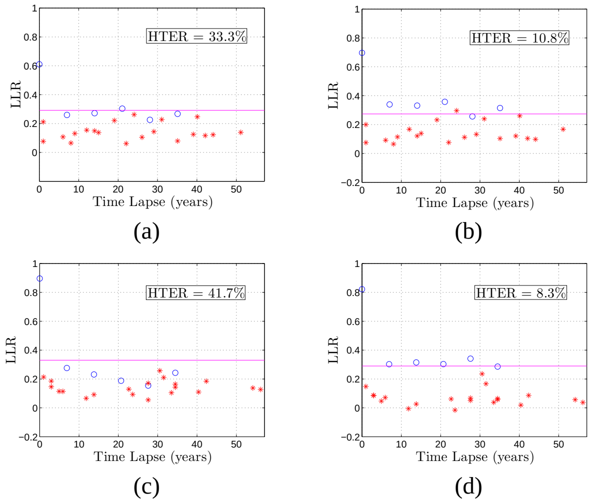
A leave-one-out decision threshold training scheme, similar to that Section 5.1, was used. However, for each test speaker, *all* LLRs from the genuine and imposter trials of all other speakers of the same gender we used. Thus, ageing information was present in the decision threshold training set. Genuine speaker LLR scores were associated with their corresponding ageing progressions (i.e. the time-lapse in years between enrolment and verification). As imposter trials had no ageing information associated with them, they were each assigned a random time-lapse within the range of the genuine speaker trial ageing progressions. A linear support vector machine (SVM) boundary was trained on these scores, and applied as a two-dimensional decision boundary to the test speaker.

For completeness, a three-dimensional decision boundary was also evaluated. In our previous work, the best performing stacked classifier incorporated both ageing progression and quality (measured by our model-based 'Wnorm' measure [13]). This configuration was evaluated in the same manner as the two-dimensional stacked-classifier. To compare to eigenageing compensation, an additional two-dimensional stacked classifier, incorporating eigenageing adapted scores and their associated Wnorm measures, was also evaluated.

To reduce any effects of the random assignment of time-lapses to imposters, each of the stacked classifier experiments involving ageing progression was repeated with 10 different random time-lapse assignments. Results presented here are the average of those 10 iterations.

6. Speaker Verification Results

The LLR scores of an example male and female speaker, before and after eigenageing compensation, are given in Figure 1. Also indicated on each plot is the decision threshold, as determined in Section 5.1, and the HTER. There is a noticeable shift in the scores of genuine trials compared with imposter trials after eigenageing compensation, and the improvement in verification performance can be seen in the associated HTERs.



(a)

(b)

(c)

(d)

Figure 1: Examples of individual speaker LLRs before and after eigenageing compensation. Genuine speaker LLRs (blue circles), Imposter LLRs (red asterisks) and the decision threshold are shown on each plot: (a) Neill (male), GMM-UBM (b) Neill, GMM-UBM + eigenageing (c) Suzy (female), GMM-UBM (d) Suzy, GMM-UBM + eigenageing

The average HTER for all 15 males and 11 females is given in Table 1 for all the systems evaluated. Also included are gender-independent HTERs, determined by pooling male and female LLRs at the decision stage. From the first two rows of Table 1, there is a significant relative improvement in HTER after eigenageing compensation: 36% for males, 39% for females, and 27% for a combination. The eigenageing performance compares favourably to the stacked classifier with ageing progression; the eigenageing HTERs are lower in all cases, and unlike the stacked classifier, no ageing information is needed at the verification stage.

The HTERs for the stacked classifier incorporating quality measures, the bottom two rows of Table 1, are the lowest by some margin. We will refer to these results further in the discussion.

A detection error tradeoff (DET) plot, from the combined LLRs of males and females is provided in Figure 2. At low false acceptance rates, the eigenageing approach clearly outperforms the baseline GMM-UBM.

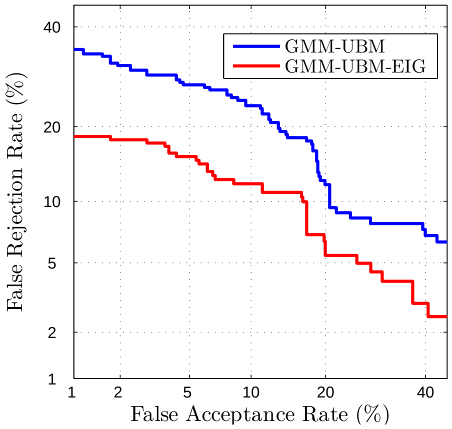


Figure 2: DET plot of scores from all 15 males and 11 females for the GMM-UBM and GMM-UBM-EIG (eigenageing compensated) systems

HTER(%)

	male	female	all
GMM-UBM	19.8	21.4	18.0
GMM-UBM-EA	12.7	13.0	13.2
SC: GMM-UBM + ageing	14.2	17.2	14.8
SC: GMM-UBM + ageing + Wnorm	5.9	7.7	6.6
SC: GMM-UBM-EA + Wnorm	3.9	2.5	3.5

Table 1: Average HTER (%) for the baseline GMM system (GMM-UBM), the GMM system with eigenageing compensation (GMM-UBM-EA), and three stacked classifier (SC) configurations combining GMM-UBM/GMM-UBM-EA scores with ageing progression (ageing) and quality (Wnorm)

7. Discussion

Eigenageing compensation provides significant improvement over a baseline GMM-UBM system at the task of verifying ageing speakers, as evident from the HTER improvements and the DET plot.

It achieves lower HTERs than our stacked classifier approach, with the advantage that no ageing meta-data is required at verification time. This protects against the possibility of an imposter attack using false ageing information. Furthermore, the eigenageing subspace training used less data (a maximum of 6 recordings in the age range 20-60 per speaker) than used for the stacked classifier boundary training.

The performance of eigenageing compensation is likely to improve with additional data. Considering that implementations of the related eigenchannel compensation technique have used thousands of recordings of speakers in different conditions to model a channel subspace [24], and here we have use at most 100 ageing difference supervectors to model the ageing subspace, there is certainly scope for improvement.

The performance of the three-dimensional stacked-classifier boundary, 'SC: GMM-UBM + ageing + Wnorm' in Table 1, has been included for completeness, as this was the best performing system in our previous work [13]. The two-dimensional boundary, 'SC: GMM-UBM-EA + Wnorm' in Table 1 essentially represents both the score and ageing dimensions with the eigenageing adapted scores, and thus the two can be compared. The two-dimensional boundary outperforms the three-dimensional one, demonstrating that the effectiveness of eigenageing compensation holds after accounting for quality information.

It is important to note that the low HTERs achieved with the addition of Wnorm are very optimistic, as the use of different datasets for imposter (CSLU) and genuine speaker trials (TCDSA) maximises the discrimination ability of the quality measure Wnorm (TCDSA is exclusively microphone data and CSLU is telephone). With a more closely matched test, the improvement with Wnorm would not be as dramatic. The reason that the TCDSA speakers were not used as imposters in a leave-one-out approach was that they were already in use in the leave-one-out scheme for ageing subspace training.

As mentioned previously, males and females experience vocal ageing in physiologically different ways. This results in different acoustic correlates with ageing between genders. For example, the speaking fundamental frequency of females decreases with age in a more consistent manner than in males [3, 4]. Given the same feature set (e.g. MFCCs), gender differences are likely to affect the quality of the ageing subspace estimations. Here, there was a slightly greater relative HTER improvement with eigenageing compensation in females than males (39% compared to 36%). However, there were less females than males used to train the gender-dependent ageing subspaces (11 compared to 14), so this may be an indication of a better female ageing subspace estimation. Ageing aside, there are performance differences between genders commonly reported in speaker recognition evaluations. It is difficult to make assertions on gender with so few subjects, but it plays a defining role in the vocal ageing process, and is an interesting topic for future research.

There are differences between individual speakers in the extent of vocal change with ageing. (Mueller [6] refers to this as the 'heterogeneity of ageing'). It would be interesting to investigate if there is any correlation between particular vocal ageing characteristics and the performance improvement observed with eigenageing compensation.

8. Conclusions

An eigenageing compensation approach to speaker verification has been presented in this paper. The method demonstrates promising performance on a dataset of long-term ageing speakers. Additional ageing data for subspace modelling will likely bring about further performance improvements, and enable the integration of ageing compensation into a factor analysis framework.

9. 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. 10, no. 6, pp. 52–62, 2012.

[3] 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.

[4] P. Torre III and J. A. Barlow, “Age-related changes in acoustic characteristics of adult speech,” *Journal of Communication Disorders*, vol. 42, no. 5, pp. 324–333, 2009.

[5] W. Decoster and F. Debruyne, “Longitudinal voice changes: Facts and interpretation,” *Journal of Voice*, vol. 14, no. 2, pp. 184–193, 2000.

[6] P. B. Mueller, “The Aging Voice,” *Seminars in Speech and Language*, vol. 18, no. 2, pp. 159–169, 1997.

[7] S. Schötz, “Perception, Analysis and Synthesis of Speaker Age,” Ph.D. dissertation, 2006, Lund University, Sweden.

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

[9] J. Harrington, S. Palethorpe, and C. Watson, “Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers,” in *InterSpeech 2007*.

[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] 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.

[12] 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*, vol. 13, no. 2, pp. 256–273, 2006.

[13] 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.

[14] ——, “Compensating for ageing and quality variation in speaker verification,” in *Interspeech 2012*.

[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*, vol. 15, no. 4, pp. 1435–1447, 2007.

[16] N. Brümmer, L. Burget, J. H. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, “Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 2072–2084, 2007.

[17] 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.

[18] “The Up series,” 1977-2012, directed by Michael Apted, produced by Granada Television.

[19] R. Rhodes, “Changes in the Voice across the Early Adult Lifespan,” in *The International Association of Forensic Phonetics and Acoustics, IAFPA, 2011*, 2011.

[20] R. Cole, M. Noel, and V. Noel, “The CSLU Speaker Recognition Corpus,” in *International Conference on Spoken Language Processing*, 1998, pp. 3167–3170.

[21] P. Kenny, M. Mihoubi, and P. Dumouchel, “New MAP estimators for speaker recognition,” in *Eurospeech*, 2003.

[22] P. Kenny and P. Dumouchel, “Disentangling speaker and channel effects in speaker verification,” in *ICASSP*, 2004, pp. 37–40.

[23] N. Brummer, “Spescom DataVoice NIST 2004 system description,” in *NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, 2004.

[24] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, “Analysis of feature extraction and channel compensation in a GMM speaker recognition system,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 7, pp. 1979–1986, 2007.

[25] P. Kenny, “Joint factor analysis of speaker and session variability: theory and algorithms - technical report,” CRIM-06/08-13, Tech. Rep., 2006.

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