SPEAKER-PHONETIC VECTOR ESTIMATION FOR SHORT DURATION SPEAKER VERIFICATION

Jianbo Ma[1,2], *Vidhyasaharan Sethu*[1], *Eliathamby Ambikairajah*[1,2], *Kong Aik Lee*[3]

[1]School of Electrical Engineering and Telecommunications, UNSW Sydney
[2]DATA61, CSIRO, Sydney, Australia
[3]Data Science Research Laboratories, NEC Corporation, Japan

jianbo.ma@unsw.edu.au

ABSTRACT

Phonetic variability is one of the primary challenges in short duration speaker verification. This paper proposes a novel method that modifies the standard normal distribution prior in the total variability model to use a mixture of Gaussians as the prior distribution. The proposed speaker-phonetic vectors are then estimated from the posterior probability of latent variables, and each vector has a phonetic meaning. Unlike the standard total variability model, the proposed method can incorporate a phoneme classifier to perform soft content matching, which has the potential to solve the phonetic variability problem. Parameter estimation and scoring formulae for speaker-phonetic vectors method are presented. Experimental results obtained using NIST 2010 data show that the proposed technique leads to relative improvements of more than 30% when fused with total variability model and tested on 3 second duration test files.

Index Terms— automatic speaker verification, short duration speaker verification, i-vector, phonetic variability, speaker-phonetic vector

1. INTRODUCTION

Most state-of-the-art text-independent speaker verification systems are comprised of i-vectors, which model speaker and channel variability with a low-dimensional representation of speech utterances [1]. These are combined with probabilistic linear discriminant analysis (PLDA), which serves as the back-end to the speaker verification system [2]. Text-independent speaker verification systems conventionally require long enrolment utterances and operate on long test utterances (2-3 minutes). In practical applications, short duration speaker verification is more desirable. It should be noted that it is reasonable to assume that long utterances can be used for enrolment purposes, since this is carried out only once and in an offline manner. Discussion in this paper is therefore confined to a scenario of long enrolment and short test utterances.

In recent years there has been increasing interest in short duration text-independent speaker verification systems. Duration compensation in an i-vector framework is major idea for short duration speaker verification. For example, twin model Gaussian-PLDA (GPLDA) has been proposed to compensate for the duration mismatch between i-vectors of long enrolment and short test utterances [3]. The covariance of the i-vector posterior probability was integrated into the PLDA model in [4-6]. Score domain compensation for duration mismatch using a quality measure function (QMF), which takes durations of enrolment and test utterances into account, was introduced in [7]. The mismatch between long training and short test duration was compensated for, in the training phase for the total variability matrix and hyper-parameters of PLDA, by adding short utterances [8]. A latent variable space that has less duration variability has been proposed to compensate duration variability and much better performance has been obtained [9]. These techniques are proposed under the fact that i-vectors from long and short utterances do not have the same distribution, and techniques are proposed to compensate this mismatch. Mismatch between long and short utterances arises primarily from the varying amounts of information in those utterances. However, the total variability model in i-vector framework is trained on long utterances, and therefore will contributes to the mismatch in i-vector space when enrolment and test utterance are in different lengths. To relieve this problem, a content aware local vector has been proposed [10]. Although different senones have been clustered agglomeratively in this method, it does not take into account the fact that different clusters may overlap. In [11], informative prior knowledge is used to compensate for channel variability, but the prior assumption is still a Gaussian distribution. Though a Gaussian mixture model (GMM) was incorporated into i-vector extraction for language identification in [12], it has no phonetic meaning and thus class assignments may not be accurate. Most importantly, this method was not fully mathematically developed.

In this paper, we aim to revise the i-vector extraction procedure to have speaker-phonetic vector representations for short duration utterances. The idea is to relax the standard normal distribution prior in the total variability model to a mixture of Gaussians. This idea has not been explored for short duration speaker verification. In this paper, we show that the total variability model with a standard normal distribution is not accurate for short duration utterances. More accurate models for short duration utterances and better performance can be obtained using a mixture of Gaussians as the prior distribution. In order to do this, a DNN trained for automatic speech recognition (ASR), e.g., a phoneme classifier, can be incorporated into the system. The mathematical development of this algorithm as well as the scoring function is presented herein.

2. PHONETIC VARIABILITY IN I-VECTORS

2.1. Total variability model

Let M be a supervector obtained by concatenating the mean vectors of all components of a Gaussian mixture model (GMM). The i-

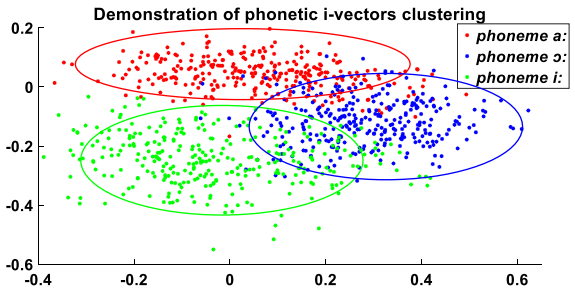

Figure 1: *Demonstration of phonetic i-vector clustering.*

vector corresponding to M is given by the well-established total variability model as follows:

$$M = M_0 + T\omega \tag{1}$$

where, M_0 is the supervector corresponding to the universal background model (UBM), T is the total variability matrix of a low rank R (e.g 400), and ω denotes latent variables that follow a normal distribution. The i-vector is the expected value of the latent variables ω and is given by:

$$E(\omega) = \left(I + \sum_{c=1}^{C} N_c T_c^* \Sigma_c^{-1} T_c\right)^{-1} \left(\sum_{c=1}^{C} T_c^* \Sigma_c^{-1} F_c\right) \tag{2}$$

where, T_c is the $D \times R$ dimensional sub-matrix of T corresponding to the c^{th} Gaussian mixture component of the UBM, C is the number of components in the UBM and D is the dimensionality of the feature space. Σ_c is the covariance of the c^{th} component of UBM and N_c and F_c are the zeroth and first order statistics of c^{th} component for a given utterance respectively.

2.2. Phonetic i-vector analysis

An underlying assumption of the total variability model is that i-vectors estimated from the same speaker should be clustered into the same group and that the contents will be essentially normalized by the long duration of the utterances. It is argued in [12] that the hypothesized standard normal prior on i-vectors only serves as check on the magnitude of the obtained i-vectors. It does not promote any form of clustering for i-vectors to be estimated from the data. We argue that i-vectors estimated from long utterances are more likely to be clustered together for a given speaker because the relative frequency of occurrences of different phonemes are more likely to be consistent, while short duration utterances are not. I-vectors of shorter utterances are more sensitive than those from long utterances. This adds weight to the suggestion that i-vectors from short duration have larger within-class variation.

To support this argument, i-vectors of different phonemes from different utterances are collected together using a phoneme decoder. Frames that are recognized as the same phonetic class for a given utterance are grouped together to estimate the corresponding i-vector. Those i-vectors are then projected into two-dimensional space by principle component analysis (PCA). Figure 1 shows the result of this analysis. 304 utterances randomly selected from background databases are used. It is clear that different groups of phonemes have different distributions and supports, that an i-vector is not phonetically invariant and that the output of the total variability model is not simply speaker discriminative information. In this case, a mixture of Gaussians should model the phonetic i-vectors more accurately compared to a single Gaussian. In total variability model, this would not be a problem for long duration utterances as the amount of information is sufficient and the statistical patterns for each group are relatively stable [13]. The extracted i-vector will not be biased toward a particular group and the within-class covariance does not increase. However, for short durations, the amount of information in each group is not statistically consistent. This will make the extracted i-vector biased toward some dominant groups and differ from one to another, resulting in larger within-class covariance. Consequently, in the long enrolment and short test situation, mismatch in terms of within-class covariance is introduced.

3. PROPOSED SPEAKER-PHONEITC VECTOR

From Section 2, i-vectors of different phonetic groups will be mapped into the same group with different distributions. But the total variability model fails to account for this. This is problematic for short duration speaker verification. In order to alleviate this problem, speaker-phonetic vector is proposed.

Kenny et al. mentioned that a single Gaussian assumption may not be optimal in the case of the eigenvoice model [14]. A prior specified with a mixture of Gaussians should be beneficial if a large number of speakers is available for training purposes. Based on the observations in Section 2, we make the assumption for our case that different phonemes will have different priors and that the latent variables are generated from different sources. Each source has its own prior and each prior bears the full burden of mapping phonetic information into different groups. The latent variable distribution is a combination of these different groups. The idea presented here is similar to independent factor analysis (IFA) [15]. However, comparing the two; in IFA, each source is univariate while we assume multivariate sources. We specify the posterior probability of each source (phoneme) by a phoneme decoder, rather than a GMM in IFA. The proposed method is developed in the following section.

3.1. Expectation of latent variables

The generative equation of the proposed method is the same as (1) and is illustrated in Figure 2. Suppose K states are specified in this model, similar to [15], and denote one particular state as q_k. The prior is specified by a combination of K Gaussians. This essentially means that there are several sources that generate the latent variables. In generating the latent variables, which source to choose depends on the probability that is generated by the same model or a separate model. For example, a separate phoneme decoder can be applied to provide this probability.

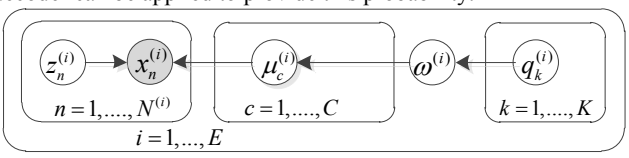

Figure 2: *Graphical model of proposed method. The variables are: z - labeling variables; x - feature frames; μ - means of the supervectors; ω - latent variable; q - state variables. The indexes are: superscript e – utterance index; subscripts c - mixture component in UBM, k - state index, and n - feature frame index.*

In this model, the prior $p(\omega)$ is the mixture of Gaussians

$$p(\omega) = \sum_k p(\omega|q_k)p(q_k) \tag{3}$$

where $p(q_k)$ is the mixture weight. The collective feature frames for a given utterance is denoted as vector X (note that the

superscript denoting utterance i is omitted in this subsection for simplicity). The probabilities of the latent variable ω given the state q_k given by:

$$p(\omega|q_k) = \mathcal{N}(m_k, B_k) \tag{4}$$

$$p(q_k) = 1/K \tag{5}$$

The conditional likelihood of observed feature frames of one utterance given latent variable and state is calculated as [14]

$$\log[p(X|\omega, q_k)] = \log[p(X|\omega)] = \sum_{c=1}^{C}\left(N_c log \frac{1}{(2\pi)^{\frac{F}{2}}|\Sigma_c|^{\frac{1}{2}}} - \frac{1}{2}tr(\Sigma_c^{-1}S_c) + \omega^* T_c^* \Sigma_c^{-1} F_c - \frac{1}{2} N_c \omega^* T_c^* \Sigma_c^{-1} T_c \omega \right) \tag{6}$$

where $S_c = \sum_{n=1}^{N}(x_n - \mu_c)(x_n - \mu_c)^*$. The likelihood term is

$$p(X) = \sum_k p(X|q_k)p(q_k) \tag{7}$$

where,

$$p(X|q_k) = \int p(X|\omega, q_k)p(\omega|q_k)d\omega = \int p(X|\omega)p(\omega|q_k)d\omega$$

As shown in the graphical model of Figure 2, $p(X|\omega, q_k) = p(X|\omega)$, which means that once ω is produced, the identity of the state q that generated it is no longer relevant [15].

Directly optimizing the likelihood term $p(X)$ is hard to do as latent variables will be inside the logarithm. Thus, the expectation-maximization (EM) algorithm is used.

To this end, the auxiliary function of the EM algorithm is

$$Q(\theta, \theta_{old}) = \int p(H|X, \theta_{old}) log[p(X|H)p(H)]dH \tag{8}$$

where H denotes the collected latent variables including ω, z and q, X denotes the observed features, and θ_{old} denotes the model hyper-parameters from the previous iteration of the EM algorithm. The value of Q is regarded as a lower bound of the log-likelihood of observable data. It will increase with every iteration, leading to a local optimum of parameters.

As the alignment of observed data to state q_k is not intended to be changed in this paper, the posterior probability to be estimated is

$$p(\omega|X) = \sum_k p(q_k)p(\omega|q_k, X) \tag{9}$$

where

$$p(\omega|q_k, X) \propto p(X|\omega, q_k)p(\omega|q_k)p(q_k) \tag{10}$$

The three terms on the right-hand side of (10) are given by equations (4), (5) and (6).

The posterior probability given the state and observed data is still a Gaussian. After some algebraic manipulations, the first and second moments of the latent variables given the state and observed data are calculated as,

$$cov(\omega|q_k, X) = (B_k^{-1} + T^*\Sigma^{-1}NT)^{-1} \tag{11}$$

$$E[\omega|q_k, X] = cov(\omega|q_k, X)(T^*\Sigma^{-1}F + B_k^{-1}m_k) \tag{12}$$

$$E[\omega\omega^*|q_k, X] = cov(\omega|q_k, X) + E[\omega|q_k, X]E[\omega^*|q_k, X] \tag{13}$$

where N is the stacked form of N_c, the same F and Σ are used as in [14], and $cov(\cdot)$ and $E[\cdot]$ are the covariance and expectation operator respectively.

According to [15], the following equation holds

$$E[f(\omega)|X] = \sum_k p(q_k|X)E[f(\omega)|q_k, X] \tag{14}$$

Taking $f(\omega) = \omega$ or $\omega\omega^T$, it is straightforward to calculate the expectations of latent variables given the observed data.

3.2. Parameter estimation

The T matrix training procedure is different to the one commonly used in [14] in how to estimate the posterior probabilities of latent variable.

The labelling variable z can be omitted as it will not change across the training and inference stages. Starting from (8), the auxiliary function is written as

$$\begin{aligned} &Q(\theta, \theta_{old}) \\ &= \sum_k \int p(\omega, q_k|X)\log[p(X|\omega, q_k)p(\omega, q_k)]d\omega \\ &= \sum_k p(q_k|X) \int p(\omega|q_k, X)\log[p(X|\omega)p(\omega|q_k)p(q_k)]d\omega \end{aligned} \tag{15}$$

where $p(q|X)$ is the posterior probabilities of state given observed data. As the T matrix is the only parameter that needs to be estimated and it is only included in $p(\omega|q, X)$ and $p(X|\omega)$, to optimize (15) is equivalent to optimize,

$$\begin{aligned} \hat{Q}(\theta, \theta_{old}) &= \sum_k p(q_k|X) \int p(\omega|q_k, X)\log[p(X|\omega)]d\omega \\ &= \sum_k p(q_k|X)E\big[\log[p(X|\omega)]\big]_{p(\omega|q_k, X, \theta_{old})} \end{aligned} \tag{16}$$

Add the utterance subscript i,

$$\begin{aligned} &\hat{Q}(\theta, \theta_{old}) \\ &= \sum_i \sum_k p(q_k|X_i)E\big[\log[p(X_i|\omega_k)]\big]_{p(\omega_i|q_k, X_i, \theta_{old})} \end{aligned} \tag{17}$$

After some algebraic manipulations, the auxiliary function can be expressed as

$$\tilde{Q}(\theta, \theta_{old}) = \sum_i \sum_k p(q_k|X_i) tr\left(\Sigma^{-1} \left(F_i E[\omega_i^*|q_k, X_i]T^* - \frac{1}{2} N_i T E[\omega_i \omega_i^*|q_k, X_i]T^* \right) \right) \tag{18}$$

Setting the gradient of the expression regarding parameter T to 0, the following updating scheme can be obtained:

$$T = AB^{-1} \tag{19}$$

where

$$A = \sum_i \sum_k p(q_k|X_i)E[\omega_i|q_k, X_i]F_i \tag{20}$$

$$B = \sum_i \sum_k p(q_k|X_i)N_i E[\omega_i \omega_i^*|q_k, X_i] \tag{21}$$

3.3. Scoring Method

In this method, the posterior probability of the latent variable will be a mixture of Gaussians, which means that one utterance can be

represented by a number of vectors based on the proposed model. A bank of GPLDAs is then estimated to obtain scores for each phonetic vector. The final score is then calculated by the formula,

$$Score(X_e, X_t) = \sum_k \gamma_k Score(X_{ik}, X_{tk}) \quad (22)$$

where $\gamma_k = N_{tk} / \sum_k N_{tk}$, $N_{tk}$is the zeroth-order statistics of state k and $Score(X_{ik}, X_{tk})$ is the GPLDA score for a single phonetic class. The weighted average considers the relative amount of information in each phonetic group, which is expected to be beneficial for short duration speaker verification.

3.4. Underlying meaning of Mixture of Gaussians prior

From the previous sections, the posterior probability of the latent variables is mixture of Gaussians. Thus, relaxing the prior of ω from (1) to a mixture of Gaussians means that the supervector M must also be a mixture of Gaussians.

The rationale behind the relaxation in this paper is that one utterance will be represented by a weighted combination of several supervectors with phonetic meaning. They are expected to have different distributions thus different priors are used to group phonetic and speaker discriminative information. Short duration utterances have unstable phonetic statistics, and weights used to combine phonetic supervectors provided by these phoneme statistics have the potential to account for uncertainty in each group. This mechanism should be beneficial for short duration utterances as it potentially performs content matching. What is more, each utterance is no longer represented by a single i-vector, rather by a number of phonetic grouped vectors that we name the speaker-phonetic vectors. The uncertainty within each group is considered in the scoring stage.

4. EXPERIMENTS AND RESULTS

A number of experiments were conducted to analyse the effectiveness of the proposed method. The 8CONV-10SEC condition of the NIST SRE'10 database [16] was chosen for these experiments along with the 8CONV-5SEC and 8CONV-3SEC conditions where test utterances are truncated to 5 and 3 seconds. The baseline system is an i-vector/G-PLDA system. Standard 13-dimensional MFCC features and their first and second derivatives were used in conjunction with a vector quantization model based voice activity detector [17] prior to feature warping [18]. Gender-dependent UBMs of 1024 Gaussian mixtures were created using utterances from the NIST SRE'04, 05, 06, 08, Switchboard II Part 1, 2, 3 and Switchboard Cellular Part 1 and 2 databases, which serve as background databases. One utterance was chosen from each speaker's available data to retain speaker diversity while reducing the overall data [19]. T matrices of rank 400 were estimated using the MSR Toolbox [20]. Linear discriminant analysis was then applied to further reduce the dimension to 200. The i-vectors were then radially Gaussianised followed by length normalization as described in [21]. The dimensionality of the speaker factors in the baseline is set to 200. For the proposed method, identical MFCC features and UBMs were used. Identical development, training and test sets were employed for the baseline and proposed systems.

The BUT group's phoneme decoder of Hungarian language [22] is used to obtain phonetic posterior probabilities in this paper. To further simplify the system, similar phonemes (e.g. long and short duration phonemes) are clustered together, resulting in 14 phonetic groups. The corresponding phonetic posterior probabilities are then added up. The prior mixture of Gaussians is estimated in this paper in the following way. First, a conventional total variability model is trained. Second, phonetic based zero- and first-order statistics are calculated as $N_{kc} = \sum_n p(k|x_n)p(c|x_n)$, $F_{kc} = \sum_n p(k|x_n)p(c|x_n)x_n$. Phonetic vectors were estimated based on these phonetic statistics. One Gaussian was then assigned to each phonetic group to fit the vectors, resulting with 14 Gaussians.

Table 1 summarises the performances of the i-vector/G-PLDA baseline system and the proposed system when the LDA dimension is 200. We can see that the proposed method has better performance when the test utterances are shorter, especially in the 3 seconds condition, where 18.2% relative improvement is observed in male condition. This supports the argument that the i-vector framework adds additional mismatch to the situation of long enrolment and short test utterance. The proposed speaker-phonetic vectors are able to relieve this mismatch.

Given that the proposed speaker phonetic vectors have the ability to group phonetic local information, which is then expected to complement the total variability framework of the baseline i-vector system, the baseline and the proposed system can be expected to be complementary and fuse well. In the experiments reported in this paper, we fused systems at the score level. Scores from the baseline system and the proposed system were fused using the BOSARIS Toolkit [23] and denoted as Fusion1. Based on the results it is clear that the two approaches are complementary, and the fusion leads to substantial improvements for all three different duration conditions, leads to 25.4%, 23.7% and 30.1% relative improvement for 10, 5 and 3s respectively in male condition. Finally, the baseline was also fused with pre-proposed acoustic local variability model [13] and denoted as Fusion 2. It shows that the proposed method in this paper outperformed the acoustic local variability model in both individual and fused system.

Table 1. *Performance* (*equal error rate %*) *of baseline and proposed systems on SRE'10 8CONV-10SEC and additional 5 and 3 second conditions.*

	Male			Female		
	10s	5s	3s	10s	5s	3s
Baseline	5.12	10.61	17.43	6.16	12.43	18.90
Proposed	5.34	10.26	14.26	6.68	11.54	16.52
Fusion1	3.82	8.10	12.19	4.94	8.90	14.15
SGPLDA	12.34	14.69	17.27	12.18	16.00	18.76
Fusion2	4.40	8.99	14.06	5.92	11.24	15.31

5. CONCLUSIONS

In this paper, we aim to revise the i-vector extraction procedure to have better representations for short utterances. We first find that different phonemes in the same utterance (with the same channel variability) tend to have different distributions, which makes short utterance i-vectors mismatched with those of long utterances. To mitigate this mismatch, we propose replacing the Gaussian prior with mixture of Gaussians in total variability model and the subsequent algorithms for training, inferring and scoring were developed. The efficacy of the proposed technique was validated on the NIST SRE'10 8CONV-10SEC condition and additional shorter duration conditions using truncated 5 and 3 second test data. The proposed method is found to be effective for shorter duration conditions and fusion with the total variability model leads to substantial improvement.

6. REFERENCES

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 788-798, 2011.

[2] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey*, 2010, p. 14.

[3] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Twin Model G-PLDA for Duration Mismatch Compensation in Text-Independent Speaker Verification," in *INTERSPEECH*, 2016.

[4] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7649-7653.

[5] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7644-7648.

[6] S. Cumani, O. Plchot, and P. Laface, "On the use of i–vector posterior distributions in Probabilistic Linear Discriminant Analysis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on,* vol. 22, pp. 846-857, 2014.

[7] T. Hasan, R. Saeidi, J. H. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7663-7667.

[8] A. K. Sarkar, D. Matrouf, P.-M. Bousquet, and J.-F. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification," in *INTERSPEECH*, 2012, pp. 2662-2665.

[9] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Duration compensation of i-vectors for short duration speaker verification," *Electronics Letters,* vol. 53, pp. 405-407, 2017.

[10] L. Chen, K. A. Lee, E.-S. Chng, B. Ma, H. Li, and L. R. Dai, "Content-aware local variability vector for speaker verification with short utterance," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5485-5489.

[11] S. E. Shepstone, K. A. Lee, H. Li, Z.-H. Tan, and S. H. Jensen, "Total variability modeling using source-specific priors," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 24, pp. 504-517, 2016.

[12] R. Travadi, M. V. Segbroeck, and S. S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *INTERSPEECH*, 2014.

[13] J. Ma, V. Sethu, E. Ambikairajah, and K. A. Lee, "Incorporating Local Acoustic Variability Information into Short Duration Speaker Verification," *Proc. Interspeech 2017,* pp. 1502-1506, 2017.

[14] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing,* vol. 13, pp. 345-354, 2005.

[15] H. Attias, "Independent factor analysis with temporally structured sources," in *Advances in neural information processing systems*, 2000, pp. 386-392.

[16] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[17] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *ICASSP*, 2013, pp. 7229-7233.

[18] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[19] T. Hasan and J. H. Hansen, "A study on universal background model training in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 19, pp. 1890-1899, 2011.

[20] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," Microsoft Research Technical Report , MSR-TR-2013-133, September 2013.

[21] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249-252.

[22] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, 2006, pp. I-I.

[23] N. Brümmer and E. de Villiers, "The bosaris toolkit: Theory, algorithms and code for surviving the new dcf," *arXiv preprint arXiv:1304.2865,* 2013.