Channel Variability Synthesis in i-vector Speaker Recognition

***Ahmed Isam Ahmed*[1], *John Chiverton*[1], *David Ndzi*[2], *Victor Becerra*[1]**

1. School of Engineering, University of Portsmouth, Portsmouth, UK, PO1 3DJ,
Email: ahmed.ahmed@port.ac.uk, john.chiverton@port.ac.uk, victor.becerra@port.ac.uk
2. School of Engineering and Computing, University of the West of Scotland,
Paisley, UK, PA1 2BE, Email: david.ndzi@uws.ac.uk

Keywords: multi-condition training, i-vector, session variability

Abstract

In this paper, we are tackling a practical problem which can be faced when establishing an i-vector speaker recognition system with limited resources. This addresses the problem of lack of development data of multiple recordings for each speaker. When we only have one recording for each speaker in the development set, phonetic variability can be simply synthesised by dividing the recordings if they are of sufficient length. For channel variability, we pass the recordings through a Gaussian channel to produce another set of recordings, referred to here as Gaussian version recordings. The proposed method for channel variability synthesis produces total relative improvements in EER of 5%.

1 Introduction

The introduction of factor analysis in speaker recognition has rapidly changed the narrative towards addressing the problem of session variability in addition to speaker variability [1]. That lead to a low dimensional representation of speech utterances, namely, the identity vectors (i-vectors) [2]. This representation enabled the application of further analyses such as Linear Discriminant Analysis (LDA) and Probabilistic Linear Discriminant Analysis (PLDA) and it introduced very fast scoring and high accuracy performance [3]. The development of an i-vector based speaker recognition system requires data that contains multiple recordings for each speaker. These multiple recording needs to have been recorded over different channels in order to model session variability. Note that session variability also includes other factors such as phonetic variation. In this paper, we propose a methodology to allow the establishment of an i-vector system when such development data is not accessible.

While inter-speakers variability modelling can be achieved using speakers' recordings from various datasets that is accessible and free, the multiple channel recordings requirement is not as easy. If the available one channel recording of the speaker is long enough, we can synthesise phonetic variability i-vectors by splitting the recording [4]. Then, we propose to pass these split recordings through a Gaussian channel to produce another set of recordings, referred to here as Gaussian version recordings.

Practically speaking, there is high expectation of channel and conditions mismatch between the test and enrolment utterances in real-life applications of speaker recognition. Factor analysis models this mismatch by learning session variability from development utterances embedded with various channel effects and recording conditions [2]. When the variability within these development utterances is limited, we propose to incur Gaussian channel effect on the available utterances. Then all of the original and new (Gaussian Channel) utterances are used together in the analysis to enrich session variability modelling.

In information theory, Gaussian noise is a basic statistical model used to mimic the effect of random processes that occur in nature [5]. It is used to model many practical channels like wired and wireless telephone channels. The additive noise in such channels may be due to various causes. By the central limit theorem, the cumulative effect of a number of random effects will be approximately normal thus the Gaussian assumption becomes valid [6]. By adding Gaussian noise as we are proposing, we aim to fit a broad channel effect in the modelling of session variability that accounts for general mismatch between test and enrolment utterances. It could also account for effects that may cause unknown forms of mismatch.

In **Section 2**, the idea of channel variability synthesis is described in light of similar work. In **Section 3**, we discuss the power of the Gaussian noise to be added. The theory and justification of channel effect synthesis are presented in **Section 4**, where we also explain how channel variability is important for i-vector development stages. In **Section 5** the experimental setup and corpora are described and in **Section 6** the system performance is illustrated and discussed.

2 Channel Synthesis and Related Work

A number of similar work can be found in the literature where the significance of using muti-channels or multi-conditions utterances has been demonstrated. In some occasions, specific types of noises are added to clean recordings in order to

account for various conditions regarding the test/detection samples. This strategy is genrally called multi-condition training and it has been mostly considered in the PLDA model training for speaker recognition using i-vectors.

The effect of multi-condition training of PLDA has been studied in [7]. It has been demonstrated that multi-condition training of PLDA is very important for the performance in noisy conditions. In [8], a mixture of channel-dependent PLDA models are trained to account for the channel conditions of each test utterance presented at the detection phase. Another mixture of PLDA Models is presented in [9], where each one is trained with different levels of noise and used according to the signal-to-noise ratio of the test utterance. The work in [10] assessed one channel features-domain noise compensation combined with multi-condition training.

The PLDA model is trained in [11] using clean recordings with the added effect of reverberation plus babble, car or helicopter noises. A full multi-condition training was presented in [12] where all the development stages of the i-vector system included clean speech samples with various types of noise.

The work presented in this paper is different in the sense that it is addressing the problem of lack of development data, and it is technically distinct in three main aspects:

- The development data used here is telephone and microphone speech, meaning that it is not strictly clean speech as in [11]. Hence, we anticipate that Gaussian version recordings can be generated from any available data in order to enrich session variability modelling.
- The work is not meant to account for certain conditions as is generally the case in multi-condition training. In other words, the test and enrolment are not subject to the Gaussian channel.
- The Gaussian channel through which the recordings are passed incur white Gaussian noise on all the frames of the speech signal, while the noise in similar multi-condition training is comparatively discrete and does not affect all the spectrum of the speech. Thus, we call it 'channel synthesis'.

In addition to the aforementioned points, similar work is not very clear about the system performance when test/enrolment is not as noisy, so it seems that the system might only be trained for noisy conditions. In order to show the validity of our proposed channel synthesis, the performance is evaluated on the same enrolment and test data before and after channel synthesis is deployed.

3 Signal to Noise Ratio

Noise is not usually a desirable effect in signals and signal processing. We need to specify a particular power of the Gaussian noise we are adding to the recordings to a level that can provide the anticipated benefit. The speech recordings we are using have different signal powers. Accordingly, we need to specify a noise power that maintains a desired signal-to-noise ratio (SNR) η_d for the new recording obtained by adding Gaussian noise to a particular original recording. The power in the context of this section is linear.

A speech signal is a random continuous-time signal that becomes discrete-time signal after sampling. Suppose we have a speech signal s with finite length of n samples expressed as

$$s = [s_1, s_2, ..., s_n]. \tag{1}$$

The power of such signal is defined by

$$P_s = \frac{1}{n}\sum_{i=1}^{n}(s(i))^2. \tag{2}$$

The power of the additive Gaussian noise that maintains a desired η_d is

$$P_{\mathrm{n}} = \frac{P_s}{\eta_d}. \tag{3}$$

Define the added Gaussian noise by a random normally distributed variable $x = [x_1, x_2, ..., x_n]$, where n is the number of values of x which is the same as the length of the speech signal s. As a normally distributed variable, x is defined by its mean μ and variance σ^2 and can be expressed by probability density function

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right). \tag{4}$$

The noise normally has zero mean which makes its power equal to its variance. Hence, our desirable Gaussian noise is a normally distributed random variable, denote by x_G, with zero mean and P_{n} is the variance of the normally distributed random variable with probability density function

$$f(x;0,P_{\mathrm{n}}) = \frac{1}{\sqrt{2\pi P_{\mathrm{n}}}}\exp\left(-\frac{1}{2}\left(\frac{x_G}{\sqrt{P_{\mathrm{n}}}}\right)^2\right). \tag{5}$$

where P_{n} is determined in equation (3). The new speech signal with added Gaussian noise will be

$$s_G = s + x_G. \tag{6}$$

where s and x_G have the same number of samples (values) which is defined earlier by n and it is the same number of samples of the resulting s_G. The MFCC features are extracted from s_G in the same way they are extracted from s. We have explained how to achieve the Gaussian noise with power that maintains a desirable signal-to-noise ratio. However, the choice of the desirable η_d is an empirical choice decided by the overall performance of the system. This will be illustrated in the Results and Discussion section.

4 Channel Variability Synthesis (Theory)

In this section, we present a theory for channel synthesis from a factor analysis perspective. In the interest of speaker verification and identification, session variability (mainly phonetic and channel variations) presents a problem when the recognition system results in, for example, the decision that two utterances are coming from different sources [13]. The problem is to determine if they are actually coming from different sources (inter-speakers variability) or if they are coming from different sessions of the same source (intra-speaker variability). In [14], a model of session variability

was the first attempt at this problem and it is referred to as eigenchannel maximum a posteriori probability estimation. It was presented to separately model session variability that can negatively affect the recognition decision. That model was then integrated with models of inter-speakers variability, in [1], to produce a model of speaker and session variability which is referred to as Joint Factor Analysis (JFA). This was defined by [15] as

$$M = m + Vy + Ux + Dz. \tag{7}$$

where *M* is a supervector of a speech utterance. *M* depends on the speaker and the session of that particular utterance and it involves combined components from speaker and the channel/session subspace. These components are: *m* is a global supervector independent of speakers and sessions (the Universal Background Model (UBM) supervector); *V* is the eigenvoice matrix that defines the speaker subspace (inter-speakers variability) and *D* is a residual term that represents variability not captured in *V*. *U* defines a session subspace (eigenchannel matrix). The vectors *x*, *y* and *z* are random variables assumed to be normally distributed. They are, respectively, the factors in the subspaces of *U*, *V* and *D* [2].

When we have two (enrolment and test) utterances of the same speaker with 'speaker and channel'-dependent supervectors M and M' respectively, a method of speaker model synthesis in [16] basically assumes that M' can be synthesised from M by adding a supervector *c* that depends only on the channel conditions of the two utterances [17],

$$\begin{cases} M' = M + c, \\ c = Ux. \end{cases} \tag{8}$$

since *c* is assumed to be a channel compensation supervector with normal distribution for the purpose of eigenchannel modelling, we assume that it is possible to flip the assumption and synthesise channel-variable utterances by adding Gaussian distributed noise in the role of channel effect to produce new recordings. See figure 1.

However, it has been found through the experiments in [18], that channel factors of JFA which are only expected to model channel effects also contain information about the speaker. This motivated the definition of the total variability space which simultaneously contains speaker and channel variabilities. Hence, the joint factor analysis in (7) became a simple factor analysis expressed as

$$M = m + Tw. \tag{9}$$

where *T* is a rectangular low-rank total variability matrix of the eigenvectors with the highest eigenvalues of total variability covariance matrix [2]. In the following sub-sections, we explain the development stages of the i-vector speaker recognition system. Channel-synthesised recordings are included in the the development stages illustrated in figure 1.

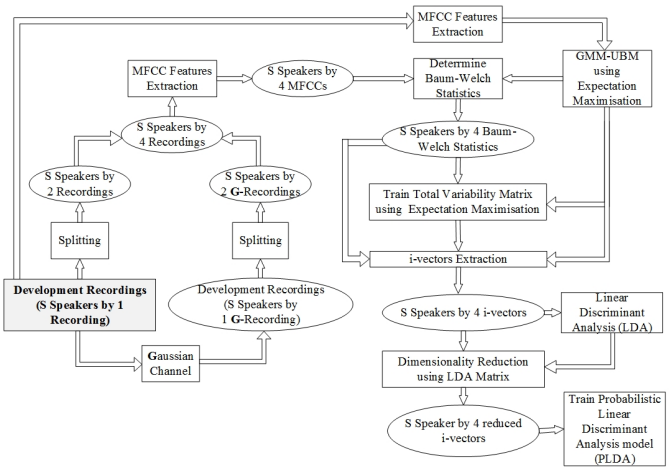

Figure 1: Development Stages of The Proposed i-vector System.

its components are the total factors. According to [2], where this simple factor analysis is proposed, *w* is a hidden variable that can be defined by its posterior distribution conditioned on Baum-Welch statistics which are calculated from a given speech utterance and the UBM. The i-vector is the mean of this distribution.

The total variability matrix is commonly obtained using the Expectation Maximisation (EM) algorithm using the Baum-Welch statistics of the speech utterances and unlike the case of classical JFA, the utterances of the same speaker are considered to belong to different speakers. These statistics are obtained by

$$\begin{cases} N_c = \sum_{l=1}^{L} P(c|y_l, \Omega). \\ F_c = \sum_{l=1}^{L} P(c|y_l, \Omega) y_l. \end{cases} \tag{10}$$

where *l* is the index of frame *y* of the speech utterance, *c = 1, 2, ..., C* are the mixture components of the UBM which is indicated by Ω and, $P(c|y_l, \Omega)$ is the posterior probability of the UBM mixture component *c* generating the feature vector y_l. For the extraction of the i-vector, centralised first order Baum-Welch statistics need to be computed based on the UBM mean mixture components m_c

$$\tilde{F}_c = \sum_{l=1}^{L} P(c|y_l, \Omega)(y_l - m_c). \tag{11}$$

Afterwards, the i-vector of any speech utterance *u* is estimated according to

$$w = (I + T^T \Sigma^{-1} N(u) T)^{-1} . T^T \Sigma^{-1} \tilde{F}(u). \tag{12}$$

where *I* is an identity matrix with the same size of the total variability dimension, 400 in this work. The structure of equation (12) is explained in details in [2].

4.1 Total Variability Matrix Training

In (9), *M* is assumed to be normally distributed of mean vector *m* with covariance matrix TT^T and, *w* is the i-vector and

4.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a statistical pattern classification technique that assumes, 1) feature vectors of

the same class are considered to be identically distributed around their class mean according to a within-class covariance matrix $\sum_W$ and, 2) the means of the classes are considered 'sample' vectors and are similarly distributed around a central mean with between-class covariance matrix $\sum_B$ [19]. The low dimension and fixed length of the i-vectors enabled effective application of LDA in order to minimise within class variance caused by channel effects. Channel synthesis is compatible with this purpose and assumption (1), because the new generated recordings are only different in channel effect, hence it enhances the within-class covariance matrix $\sum_W$ interpretation of the within-class variance.

4.3 Gaussian Probabilistic LDA Model

This model was first presented in [20] in order to address the face recognition problem of different pose and lighting of test and enrolment data. Thus it assumes that the data is resulting from a generative model which incorporates within and between class variance. Similarly used in speaker recognition, i-vectors are regarded as observations which can be decomposed by the following Gaussian-PLDA generative model [3]

$$w = m + \Phi_\beta + \Gamma_{\alpha_r} + \epsilon_r. \tag{13}$$

where $m + \Phi_\beta$ is a speaker dependent term and, $\Gamma_{\alpha_r} + \epsilon_r$ is a channel dependent term. They respectively describe between-speaker variability, eigenvoices Φ, and within-speaker variability, eigenchannels Γ. β and α are statistically independent latent vectors with standard normal distribution. m is a global offset (the mean of the development i-vectors) and ϵ_r is a residual term assumed to be Gaussian with zero mean and diagonal covariance matrix, however, a full covariance matrix $\sum$ of ϵ_r can compensate for $\Gamma_{\alpha_r} + \epsilon_r$ as proposed in [21].

The G-PLDA model training is simple and computationally efficient, however, it assumes Gaussian distribution of the input observations (i-vectors). It was reported in [3] that G-PLDA gives inferior performance compared to Heavy-Tailed PLDA [21] unless a transformation is applied to the i-vectors, where radial gaussianisation was used for this purpose.

As the concept behind this model is analogous to that of LDA, channel synthesised recordings will impose similar effect to that experienced in linear discriminant analysis, however, the purpose of using the G-PLDA model is to carry the scoring between test and enrolment i-vectors. This is a fast scoring procedure based on the log-likelihood ratio of the same(H_s)/different(H_d) speaker hypotheses which aims at determining if two utterances (test and enrolment) belong to the same speaker or to different speakers

$$score = \log_e \frac{p(w_1, w_2|H_s)}{p(w_1|H_d)p(w_2|H_d)}. \tag{14}$$

$$\begin{aligned} score = \log_e N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} \sum_{tot} & \sum_{ac} \\ \sum_{ac} & \sum_{tot} \end{bmatrix} \right) \\ - \log_e N\left(\begin{bmatrix} w_1 \\ w_2 \end{bmatrix}; \begin{bmatrix} m \\ m \end{bmatrix}, \begin{bmatrix} \sum_{tot} & 0 \\ 0 & \sum_{tot} \end{bmatrix} \right). \end{aligned} \tag{15}$$

where $\sum_{tot}$ and $\sum_{ac}$ are respectively $\Phi\Phi^T$+$\sum$ and $\Phi\Phi^T$ of equation (15).

The scores using this log-likelihood are achieved in closed-form solution. Which means that each test i-vector is scored against its target (same speaker) enrolment i-vectors and all other non-target i-vectors. N denote the number of i-vectors to be scored. Similar to LDA, G-PLDA is trained using i-vectors of the development set labelled as per each speaker's recordings.

5 Corpora and Experimental Setup

The development data included one recording for each speaker with average length of 2 minutes. Voice Activity Detection (VAD) is used to remove silences from these recordings then they are divided in half to make up two recordings. Next the two recordings are passed to the Gaussian channel to produce another two recordings, the result is four recordings for each speaker in total. This development data consisted of the NIST 2002 SRE telephone training data (English) [22], the NCHLT Speech Recognition microphone corpus (English and Afrikaans) [23] and the LWAZI Speech Recognition telephone corpus (English, Afrikaans, Sesotho and Zulu) [24]. The system is gender-independent and in order to balance the analyses, the number of development speakers is 639 male and 639 females speakers (1278 speakers).

The evaluation data used is the telephone speech test samples of the NIST 2002 Speaker Recognition Evaluation, which contains 191 female and 139 male speakers (330 in total). We used one target test utterance for each speaker which resulted in 108900 gender-independent detection trials.

The speech features used are 13 Mel-Frequency Cepstral Coefficients (MFCC) calculated from the Hamming windowed speech frames of 25 ms size and 40% overlap (10 ms shift), appended with their first and second derivatives and normalised using feature warping with window size of 3 s. The dimension of the total variability matrix is 400 factors resulting in 400 dimension i-vectors reduced to 150 using LDA. The dimension of the G-PLDA model is also 150. The evaluation procedure of the system is shown in the diagram of figure 2.

6 Results and Discussion

The system is firstly established using only two recordings per development speaker obtained by splitting the one recording we have for them. The reason for this step is to illustrate the system performance before and after including Gaussian version recordings in the development. Hence, we show the effect on each development stage and then for the overall system. The system performance is evaluated in terms of Equal Error Rate (EER) and Minimum Detection Cost Functions (minDCF) of the 2008 and 2010 NIST speaker recognition evaluation. As explained earlier, we need to specify the power of the Gaussian noise that we are adding to the original recording in order to produce new recordings. This is decided

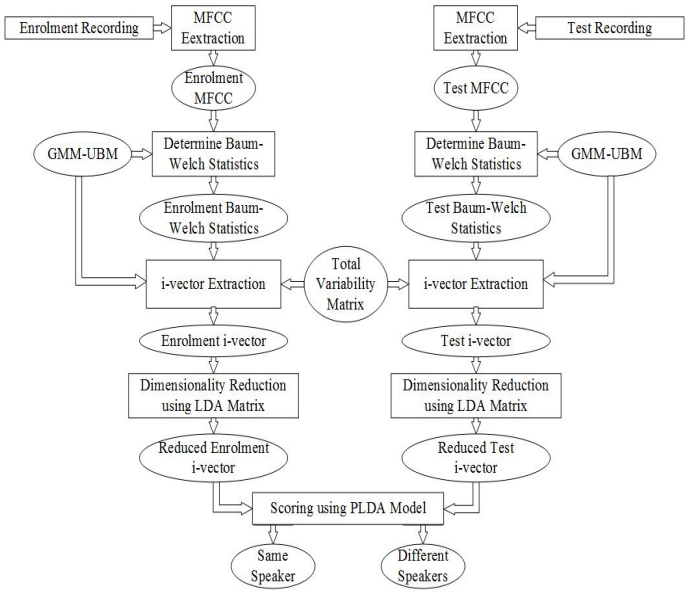

Figure 2: Evaluation Procedure in i-vector Speaker Recognition System.

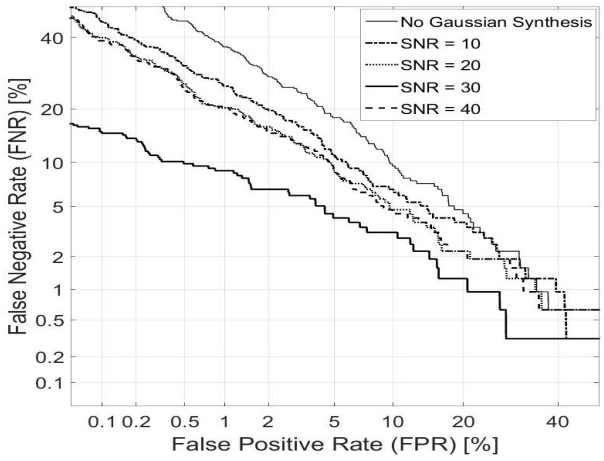

Figure 3: Detection Error Tradeoff curves of system performance at different SNRs of the resulting Gaussian recordings used for channel synthesis.

by the SNR of the resulting new recordings. However, the desired SNR is empirically decided based on the performance of the system. We evaluated four values of SNR as shown in figure 3. We can see that the best system performance was achieved at SNR of 30 dB. For 10 dB SNR, the new recordings became very noisy, however, channel synthesis was also achieved compared to the case where no channel synthesis is deployed. The performance at 20 dB SNR is better compared to that of 10 dB SNR, because the noise power is decreased. Following the best performance accomplished at 30 dB SNR, we can see that the performance at 40 dB SNR is comparable to that at 20 dB although the noise power is less. That is because at low noise power, the effect of channel synthesis decreased and the performance started to move closer to the case where channel synthesised recordings are not used.

Now that the best system performance is achieved at 30 dB SNR; per-stage performance improvements are illustrated in Figure 4 and Table 1. In LDA, the usage of Gaussian version recordings produced improvement of 1.35% EER. LDA is used for channel effect compensation. Since the Gaussian channel effect is not included on the test and enrolment data, we can see that channel synthesis is successful. It appears as if it is actually the recordings over a different channel since it has produced the anticipated positive effect on the LDA for defining more precisely the directions that minimise between speaker variability and potentially those directions that maximise between speaker variability. Similar behaviour in G-PLDA gave an improvement of 3.4% in EER. This is where we again point out the difference from multi-condition training, as significant enhancement occurs in the system although the test and enrolment data are not passed through a Gaussian channel unlike the development data. When Gaus-

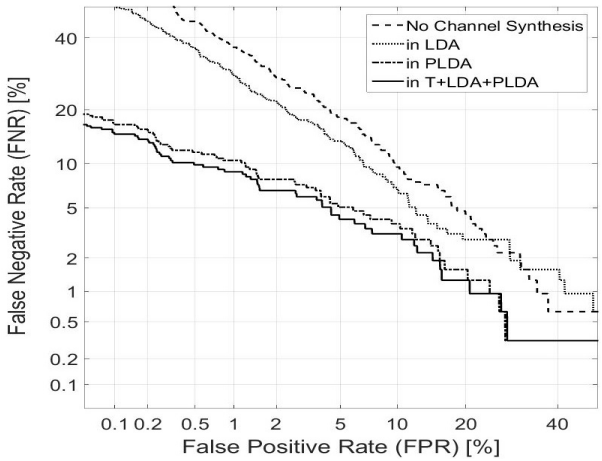

Figure 4: Detection Error Tradeoff curves of system performance. Illustrates the in the development stages where Gaussian version recordings are included [in LDA, in PLDA and in (T+LDA+PLDA)].

Development Stages	EER%	DCF 2008	DCF 2010
No Channel Synthesis	9.81	4.65	0.088
in LDA	8.46	3.76	0.085
in PLDA	5.06	1.52	0.029
in (T+LDA+PLDA)	**4.43**	**1.33**	**0.025**

Table 1: System Performance in terms of EER and DCF. It shows the effect of including Gaussian version recordings in the development stages of the system.

sian version recordings were involved in the total variability matrix training alone a slightly inferior accuracy was noticed. After providing improvements by involving Gaussian version recordings in LDA and G-PLDA, including these recordings in the total variability matrix training as well also presented further improvement. Hence, we can see that stages following the total variability training may place a burden on the channel subspace learned if poorer channel variability is involved in the subsequent development steps.

7 Conclusion

Multiple recordings of the speakers of the development data is an essential requirement for the establishment of factor analysis-based speaker recognition systems. It has been shown in this paper that in case such data is not accessible, channel variable recordings can be synthesised from the recording that we have for each speaker even if they are not clean recordings. The outcome of this paper also indicates that if clean recordings are available for the development and in order to carry broader channel variability modelling, various or desired channel effects can be incurred on the clean 'recordings as the telephone channel simulation in [25] as well as the possibility of adding different environmental noise for multi-condition training to account for test/detection conditions.

References

[1] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," *CRIM, Montreal,(Report) CRIM-06/08-13*, vol. 215, 2005.

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

[3] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, vol. 2011, 2011, pp. 249–252.

[4] M.-W. Mak and W. Rao, "Utterance partitioning with acoustic vector resampling for gmm–svm speaker verification," *Speech Communication*, vol. 53, no. 1, pp. 119–130, 2011.

[5] C. Houdré, D. Mason, P. Reynaud-Bouret, and J. Rosiński, *High Dimensional Probability VII: The Cargèse Volume*, ser. Progress in Probability. Springer International Publishing, 2016.

[6] T. Cover and J. Thomas, *Elements of Information Theory*. Wiley, 2012.

[7] P. Rajan, T. Kinnunen, and V. Hautamäki, "Effect of multi-condition training on i-vector plda configurations for speaker recognition." in *Interspeech*, 2013, pp. 3694–3697.

[8] J. Villalba and E. Lleida, "Handling i-vectors from different recording conditions using multi-channel simplified plda in speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.(ICASSP)*, 2013, pp. 6763–6767.

[9] M.-W. Mak, X. Pang, and J.-T. Chien, "Mixture of plda for noise robust i-vector speaker verification," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 1, pp. 130–142, 2016.

[10] D. Martinez, L. Burget, T. Stafylakis, Y. Lei, P. Kenny, and E. Lleida, "Unscented transform for ivector-based noisy speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4042–4046.

[11] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multi-condition training of gaussian plda models in i-vector space for noise and reverberation robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4257–4260.

[12] D. Ribas, E. Vincent, and J. R. Calvo, "Full multicondition training for robust i-vector based speaker recognition," in *Interspeech 2015*, 2015.

[13] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in gmm-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1448–1460, 2007.

[14] P. Kenny, M. Mihoubi, and P. Dumouchel, "New map estimators for speaker recognition." in *INTERSPEECH*, 2003.

[15] D. B. Rubin and D. T. Thayer, "Em algorithms for ml factor analysis," *Psychometrika*, vol. 47, no. 1, pp. 69–76, 1982.

[16] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Sixth International Conference on Spoken Language Processing*, 2000.

[17] 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.

[18] N. Dehak, "Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification," Ph.D. dissertation, École de technologie supérieure, 2009.

[19] H. Beigi, *Fundamentals of speaker recognition*. Springer Science & Business Media, 2011.

[20] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[21] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.

[22] A. Martin and P. Mark, "2002 nist speaker recognition evaluation ldc2004s04," *Web Download*, 2004. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2004S04

[23] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based asr data collection tool for under-resourced languages," *Speech communication*, vol. 56, pp. 119–131, 2014. [Online]. Available: https://rma.nwu.ac.za/

[24] E. Barnard, M. Davel, and C. Van Heerden, "Asr corpus design for resource-scarce languages." ISCA, 2009. [Online]. Available: https://rma.nwu.ac.za/

[25] G. Zuo, W. Liu, and X. Ruan, "Speech conversion from clean conditions to telephone ones," in *Intelligent Control and Automation, 2004. WCICA 2004. Fifth World Congress on*, vol. 5. IEEE, 2004, pp. 4211–4214.