Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

Tipo de artículo: Artículo de revisión
Temática: Reconocimiento de patrones
Recibido: 22/09/2015 | Aceptado: 16/12/2015

Variability compensation for speaker verification with short utterances

Compensación de la variabilidad para la verificación de locutores con señales cortas

Flavio J. Reyes-Díaz[1*], Gabriel Hernández-Sierra[1], José R. Calvo de Lara[1]

[1]Advanced Technologies Application Center (CENATAV). 7a.A # 21406 e/ 214 y 216, Playa, Havana, C.P. 12200, Cuba. Email: freyes,gsierra,jcalvo@cenatav.co.cu

[*]Autor para correspondencia: freyes@cenatav.co.cu

Abstract
Nowadays, represents an attractive challenge the application of Automatic speaker recognition in real scenarios, where the use of short duration signals for forensic or biometric speaker verification is very common. In this paper we perform an analysis of the behavior of within-class and between-classes scatter matrices, showing the importance to reduce within-class scatter to face the speaker recognition with short duration utterances. In addition, two duration compensation methods for short duration utterances on i-vector framework were proposed. Both of them were evaluated through speaker verification experiments on NIST-SRE 2008 dataset. The proposed methods shown an improvements under enrollment-test matched conditions regard to the duration.

Keywords: short utterance, variability compensation, speaker verification, i-vector.

Resumen
En la actualidad representa un desafío atractivo la aplicación del reconocimiento automático de locutores en escenarios reales, debido a que es muy común el uso de señales de corta duración para la verificación biométrica y forense de locutores. En esta investigación realizamos un análisis del comportamiento de las matrices de dispesión dentro de las clases y entre clases, mostrando la importancia de reducir la dispersión dentro de las clases para hacer frente al reconocimiento de locutores a partir de expresiones de corta duración. Además, se propusieron dos métodos de compensación de la duración sobre el enfoque i-vector. Ambos métodos fueron evaluados a través de experimentos de verificación del locutor utilizando la base de voces NIST-SRE 2008.

Palabras claves: *señales cortas, compensación de la variabilidad, verificación de locutores, i-vector.*

Introduction

Currently, the necessity of processing speech signals acquired in real uncontrolled environments is growing. This fact imposes new challenges for speaker recognition systems such as the handling of variability factors,

speech duration and emotional state, as well as acoustic distortions, noise and reverberation.

The well-known i-vector speaker representation (DEHAK et al., 2011) does not take into account the speech duration and because that, the performance of the speaker recognition using a cosine similarity measure (CSM) or probabilistic Linear Discriminative Analysis model (PLDA) (PRINCE and ELDER, 2007; KENNY, 2010) decrease quickly when the enrollment or test utterance duration decreases, as shown in (KANAGASUNDARAM et al., 2011; SARKAR et al., 2012; KANAGASUNDARAM et al., 2012).

This problem its very common in biometric and forensic identification by voice. For that reason, some works as (KANAGASUNDARAM et al., 2011; MANDASARI et al., 2011; SARKAR et al., 2012) are based on multi-condition training techniques to compensate the short duration variability in the i-vector framework but not including and not evaluating the full (not short) utterance condition against with full samples. In (SARKAR et al., 2012) only the full utterances are evaluated, showing that the speaker recognition performance decreases when we use multi-condition training regarding the obtained results without multi-condition training.

Previous works as (MANDASARI et al., 2011; HASAN et al., 2013), use the utterances duration as a qualitative measure, to reduce the effect of short duration in speaker verification. More recently in (HAUTAMÄKI et al., 2013), authors proposed a strategy for variability compensation due to short utterances, replacing the Baum-Welch algorithm by the Minimax algorithm (MERHAV and LEE, 1993) to estimate the zero order sufficient statistics. Kenny et al. in (KENNY et al., 2013) proposed the use of the uncertainty propagation to introduce the duration variability into the i-vector. Authors in (KANAGASUNDARAM et al., 2013) have been working to mitigate or reduce the effect caused by short duration samples, proposing a new technique to session variability compensation.

The main goal of our research is to deepen in the analysis of the effect caused by the variability among enrollment and test samples due to different duration in the performance of speaker recognition. As conclusion of this study, we recommend a new method (SUN-LDA2) that incorporate the duration variability information into within-class scatter estimation, to improve the short duration utterances compensation of the speaker verification on i-vector framework. In addition, new approach (IV-DVC) to compensate the duration variability was proposed. This method is based on divide and conquer technique, compensating in different spaces the channel variability and the duration variability. To support our proposed approaches we report experiments on speaker verification evaluation NIST[1] with different utterances duration.

In Section *"Variability compensation method in I-vector framework"* we describe the LDA method and the study of the behavior of the between-classes and within-class variability. In Section *"Short utterances impact*

[1]http://www.nist.gov/itl/

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

on within-class scatter" the two proposed methods and within-class variability analysis are explained. The experimental protocol to evaluate the State of the Art and our methods is develop in *"Experimental set-up"* Section. Experimental results are detailed in *"Results and discussion"* Section and finally we arrived to the conclusions.

Variability compensation method in I-vector framework

State of the Art speaker recognition systems are based on the i-vector representation of speaker utterances and can be defined by the posterior distribution of the hidden variables conditioned to the Baum-Welch statistics extracted from the utterance. I-vector is computed from the one only variability space called Total Variability Space (T) (DEHAK et al., 2011), that simultaneously contains the speaker and session variabilities. These speaker template are represented by

$$M = m + Tw, \tag{1}$$

where m is a super-vector obtained by the concatenation of the UBM centers, that contains the speaker-independent and session information, T is a low rank rectangular matrix and w is a random vector that follow a normal distribution $\mathcal{N}(0, I)$ and represent a speaker into a speaker verification systems, called intermediate vector or i-vector. In equation 1, we assume that the vector M keeps a normal distribution with m and TT' as center and covariance respectively.

Session variabilities are known to be an important factor of performance degradation. Compensating for these variabilities becomes a mandatory part of a modern speaker recognition systems. Some compensation methods, arising from other areas, have been applied with the aim to improve the efficiency in the recognition, one of such method is the Linear Discriminant Analysis (LDA) (RAO, 1948). This technique is a dimension reduction method currently used on the i-vector framework for inter-session variability compensation in the speaker verification, initially proposed by Dehak et. al. in (DEHAK et al., 2011). The principal goal of LDA is to maximize the variance between-classes (S_b) and simultaneously minimize the within-class variance (S_w) of a speakers population:

$$S_b = \sum_{l=1}^{L} (x_l - \bar{x})(x_l - \bar{x})', \tag{2}$$

where L is the speakers number, x_l is the mean of the i-vectors of each speaker and $\bar{x}$ is the global mean vector of the speakers population, and

$$S_w = \sum_{l=1}^{L} \frac{1}{n_l} \sum_{i=1}^{n_l} (x_i^l - x_l)(x_i^l - x_l)', \tag{3}$$

where n_l is the number of utterances for l speaker and x_i^l is the vector of the i-th utterance the l speaker.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

Then the projection matrix A, is a subset of eigenvectors J associated with the largest eigenvalues, which are obtained by optimizing the objective function Fisher:

$$J(v) = \frac{v' S_b v}{v' S_w v}, \tag{4}$$

where v is a given space direction.

Short utterances impact on within-class scatter

Analysis of the short-length utterances variability in the speaker verification is an issue that has been gaining significance, because, this variability affects the speaker discriminative information contained in the i-vectors. LDA algorithm is one of the most common techniques used to reduced the variability introduced by the short utterances. This technique has been modified by several authors in order to face the variability introduced by signals duration. In (KANAGASUNDARAM et al., 2013), authors proposed a duration compensation method (Source and utterance-duration normalized called SUN-LDA) incorporating the duration variability information in the estimation of the between-class scatter matrix by (eq. 9 in (KANAGASUNDARAM et al., 2013)):

$$S_b = \alpha_{full} S_b^{full} + \alpha_{short} S_b^{short}, \tag{5}$$

where α_{full} and α_{short} are the variabilities weight for full and short utterances respectively, and S_b^{full} and S_b^{short} are the between-class scatter matrix of full and short utterances respectively. They not included the duration variability information in the estimation of the within-class scatter, based on the supposition that this inclusion would affect the speaker verification performance.

We consider that both variabilities (S_b and S_w) are of great importance to face the duration variability improving the speaker verification performance. Because the i-vectors obtained from short utterances contain low discriminative information of the speaker, implying a greater within-class scatter. In addition this scatter provokes a shift in the center of the class affecting the real distance between-classes (speakers), causing a greater overlapping between them. In order to involving both variabilities we propose a modification of the method proposed in (KANAGASUNDARAM et al., 2013), and a new method based on divide and conquer paradigm.

Estimation of within-class scatter with duration variability

To reduce variability due to utterances duration a new method called SUN-LDA2 was proposed. This method include not only full utterances information but also the information of the short utterances to estimate the

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

within-class scatter matrix S_w, by:

$$S_w = S_w^{full} + S_w^{short}, \tag{6}$$

where S_w^{full} and S_w^{short} are the within-class scatter matrices of full and short utterances respectively, they are estimated using equation 3. Here we don't used the variability weight proposed in (KANAGASUNDARAM et al., 2013) to compute de S_b using interchangeably the variability matrices, because in our case these weights don't introduces relevant information.

The between-class scatter matrix S_b is calculated using the eq. 5, without the weights. Finally, the projection matrix was obtained using the eigenvectors corresponding to the largest eigenvalues, classic LDA.

Duration Variability compensation using "Divide and Conquer"

Regularly in real conditions, the utterances corresponding to the same speaker are affected by many types of variabilities, some intrinsic as duration or emotion and some extrinsic as channel or noise. The mixture of these two types of variability in a single covariance matrix could be provoking inefficiency in the estimation of the scatter required to mitigate both. Examples of the problem:

- Session variability is inserted into all utterances, therefore when the duration variability is estimated also includes information about the session.
- Covariance matrices containing information about session and duration variability by the sum in S_b (eq. 5) and S_w (eq. 6).
- Multi-condition training techniques (MCLAREN and VAN LEEUWEN, 2011) using databases with different conditions of variability to obtain the covariance matrices, used to obtain S_b (eq. 2) and S_w (eq. 3).

Given this drawback, an interesting variant would be to attack each cause of variability independently, supported in the Divide and Conquer paradigm. So, we propose a new method to duration variability compensation on i-vectors framework, named IV-DVC. The idea behind the method is to compensate the different variabilities in separate spaces. The i-vectors are initially projected to a new space where session variability is mitigated, allowing a correct estimate of the variability with respect to the duration, to reduce the effect caused by the short expressions in the i-vectors.

This method reduces the effect caused by session variability using a projection matrix (A) obtained with LDA algorithm, as described in the section using a development set without duration variability, all samples have

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

long duration (full length). Later on, to mitigate the duration variability, a projection matrix (B) is compute with LDA algorithm. To estimate the between-class (S_b^{dvc}) and within-class (S_w^{dvc}) scatter matrices eq. 7 and 8 were used with a development set with large and short utterances.

$$S_b^{dvc} = \sum_{l=1}^{L}(A'(x_l - \bar{x}))(A'(x_l - \bar{x}))', \tag{7}$$

$$S_w^{dvc} = \sum_{l=1}^{L}\frac{1}{n_l}\sum_{i=1}^{n_l}(A'(x_i^l - x_l))(A'(x_i^l - x_l))', \tag{8}$$

where the parameters have the same definition as in equations 2 and 3.

The CSM between two i-vectors $x1$ and $x2$, when the variability is compensated with the method IV-DVC is:

$$Score_{dvc}(x_1, x_2) = \frac{(B'A'x_1)'(B'A'x_2)}{||B'A'x_1||||B'A'x_2||}. \tag{9}$$

Analysis of between and within class scatter matrices

Therefore, we introduced a visualization tool intended to analyze better the behavior of both variabilities. For this purpose some sessions of five speakers were selected and for each session i-vectors from utterances with different duration, 3, 5, 10, 15, 20 seconds and full duration, were extracted. The LDA projection matrix was trained with three setting, the first one using the method proposed in (KANAGASUNDARAM et al., 2013), the second one was trained using our proposal SUN-LDA2 and the third one, the IV-DVC method was used to reduce the duration variability. To visualize the behavior of the between-class and within-class we use the first two dimension of the i-vectors projected with Principal Component Analysis (PCA).

The Figure 1 shows the distribution of the first two dimension of the i-vectors using SUN-LDA (Figure 1.a), SUN-LDA2 (Figure 1.b) and IV-DVC (Figure 1.c) methods.

The proposed SUN-LDA2 method obtains a new distribution of between-class and within-classes scatter showing a greater reduction of the within-class scatter compared with SUN-LDA. Although this compensation implied a unwanted reduction, of the between-class scatter. The duration compensation with IV-DVC method obtains a new distribution of the classes, showing a similar behavior of between-class and within-class scatter than the SUN-LDA2 method. Nevertheless, experimental results of the speaker verification in section , prove that within-class scatter is more important than between-class scatter if we want compensate the short utterances.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

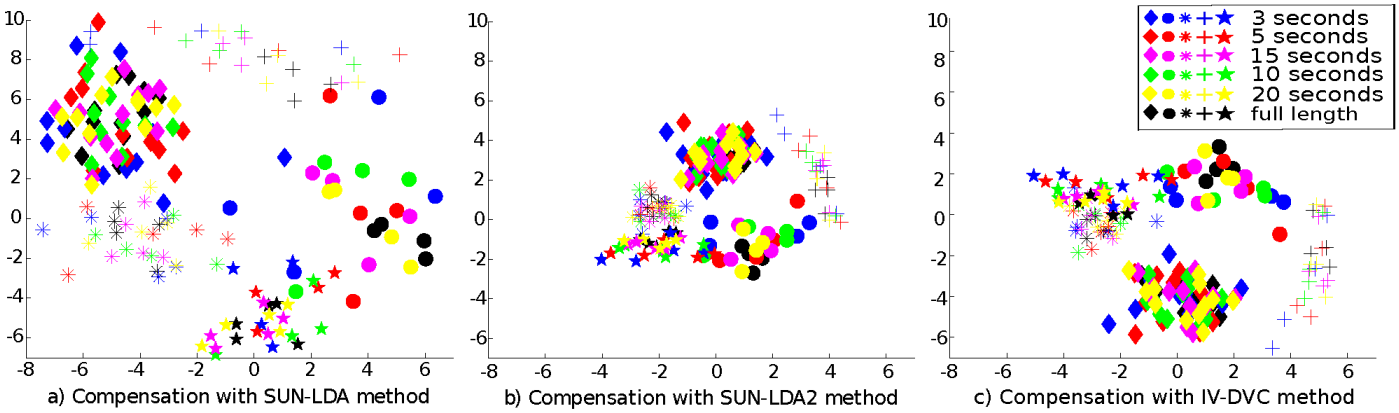

Figure 1. Distribution of the classes obtained after the short utterance compensation methods. Each symbol represent each speaker.

This analysis and the corresponding experimental result confirm our initial idea. The inclusion of variability due to short duration utterances in the estimation of the within-class scatter, far from affecting, reinforced the speaker verification performance.

Experimental set-up

In all the experiments presented in this section, the front end uses 19 linear frequency cepstral coefficients (LFCCs) (DAVIS and MERMELSTEIN, 1980), with energy, delta and delta-delta coefficients, giving a 60-dimensional feature vector. We used NIST 2008 SRE dataset, specifically male telephone sessions, as evaluation set in the speaker verification. To obtain the simulated short utterances we truncate the first 5, 10, 15, and 20 seconds of each sample of the evaluation set. We used NIST SRE-04 and SRE-05 telephone sessions as training data to obtain the gender dependent Universal background model (UBM) (REYNOLDS et al., 2000) with 512 Gaussian components, the total variability T matrix with 400 dimension and the LDA compensation matrix. For multi-condition training with short duration, we truncate the first 3, 5, 10, 15, 20 seconds of each sample of the training data.

The classification with PLDA model represents the State of the Art in the speaker recognition fields. Therefore, we decided to develop some experiments where we combine our proposals of duration compensation with PLDA model as classifier. Different PLDA models, depending on the methods of compensation, were trained using i-vectors resulting from the duration compensation of the simulated short utterances from NIST SRE-04 and SRE-05 telephone sessions.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

Results and discusion

Table 1 shows results obtained for the four baseline and two proposed methods, evaluating the EER of speaker verification performance with different variability compensation methods using LDA to face short duration utterances. The baselines were evaluated using cosine similarity measure (CSM) between i-vectors: IV-LDA: using only the full duration to estimate the session compensation matrix (DEHAK et al., 2011), $IV-LDA_{var}$: using multi-condition training with six set of utterances with different duration and full length to estimate the compensation matrix (KANAGASUNDARAM et al., 2011; SARKAR et al., 2012; MANDASARI et al., 2011), SUN-LDA method proposed in (KANAGASUNDARAM et al., 2013), but evaluating only the duration variability, and Within-class Covariance Normalisation (WCCN[LDA]) method proposed by (DEHAK et al., 2011). In addition the same baseline methods using PLDA model were evaluated.

Table 1. Performance in terms of EER% for different matched duration conditions between target and test. The proposed methods are highlighted

Cosine distance similarity

Approach/Conditions	Full-Full	20-20	15-15	10-10	5-5	Mean
$IV-LDA$	**3,70**	9,56	11,1	13,9	22,1	12,07
$IV-LDA_{var}$	5,50	8,53	10,9	13,2	20,7	11,76
$WCCN[LDA]$	4,55	**8,25**	**10,0**	13,4	**19,7**	11,18
$SUN-LDA$	3,93	9,56	10,7	14,2	22,3	12,13
SUN-LDA2	4,25	9,33	11,1	13,2	21,1	11,80
IV-DVC	4,55	8,42	**10,0**	**12,7**	**19,7**	**11,07**
PLDA model						
	Full-Full	20-20	15-15	10-10	5-5	Mean
$IV-PLDA$	**2,96**	7,97	10,0	14,1	21,4	11,28
$IV-PLDA_{var}$	3,87	8,22	9,33	12,5	18,2	10,42
$WCCN[LDA]+PLDA$	4,27	**7,69**	9,11	11,6	18,0	10,13
$SUN-LDA+PLDA$	4,09	7,86	9,11	11,8	18,8	10,33
SUN-LDA2+PLDA	4,27	7,97	**8,88**	11,8	18,8	10,34
IV-DVC+PLDA	4,32	7,97	8,90	**10,8**	**17,1**	**9,81**

The main result between all experiments was obtained using the proposed method IV-DVC. This proposal obtains the better results in the majority of evaluations (7 of 10) of matched short duration conditions between target and test. IV-DVC obtained an average improvement of 1% using CSM and 3.1% using PLDA score, respect to the second best compensation method WCCN[LDA]. So, the importance of using the Divide and Conquer paradigm to compensate different variabilities in different space, in the same utterance, was demonstrated.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 194-204

The proposed method SUN-LDA2 showed that the information of the intra-class scatter of short utterances is necessary to duration variability compensation, as opposed of the raised in (KANAGASUNDARAM et al., 2013). Hence in the estimation of within-class scatter is very important to incorporate all variabilities between the i-vectors, in our case the duration variabilities. As shown in Table 1 the results of SUN-LDA2 method compared with SUN-LDA reflect an average improvement of 2.7% using CSM and similar efficacy with the PLDA score.

From the scatter analysis section and experimental results obtained with proposed methods aimed at the duration variability compensation, we can raise that is very important to reduce the within-class scatter in speaker verification because this reduction minimizes the overlapping between classes implying an improvement of the classifiers performance.

By other hand, the inclusion of dataset with different short durations (multi-condition training) to obtain the LDA matrices carry out an improvement in terms of EER. However in the full utterances evaluation the performance is worse regard to short utterances evaluation. This problem is due the fitting of the data to short duration conditions for estimation of the variability compensation matrix, so the compensation is biased to the short duration condition. Finally the PLDA score shows a better performance in general than CSM, comparing both sections of the table.

Conclusion

The importance of including the duration variability information in the estimation of within-class scatter matrix was confirmed, attending the graphical analysis and experimental results. This inclusion implied an improvement in terms of EER on experimental results of the proposed methods to face the short utterances in speaker verification, SUN-LDA2 and IV-DVC.

An important contribution of this work is the proposal of a new duration variability compensation method IV-DVC using “Divide and Conquer” paradigm. This method separates the session and duration compensation focusing on mitigate, separately, both variabilities. IV-DVC overcome the efficacy of the rest of the baseline methods facing evaluations with short duration utterances, achieving with PLDA a relative improvement of 3.1% compared with the best reference system WCCN[LDA]. In addition IV-DVC method is the most robust of all evaluated methods, because it shows a minor variance among all duration conditions.

As future work we will continue the research to obtain a variability compensation method able to face short duration utterances without loss in efficiency facing long duration utterances.

References

DAVIS, S. B. and MERMELSTEIN, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 28(4):357–366.

DEHAK, N., KENNY, P., DEHAK, R., P., D., and P., O. (2011). Front-end factor analysis for speaker verification. volume 19, pages 788–798.

HASAN, T., SAEIDI, R., HANSEN, J. H., and Van LEEUWEN, D. A. (2013). Duration mismatch compensation for i-vector based speaker recognition systems. In *Acoustics, Speech and Signal Processing (ICASSP)*, pages 7663–7667.

HAUTAMÄKI, V., CHENG, Y. C., RAJAN, P., and LEE, C. H. (2013). Minimax i-vector extractor for short duration speaker verification. In *In Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 3708–3712. ISCA.

KANAGASUNDARAM, A., DEAN, D., GONZALEZ-DOMINGUEZ, J., SRIDHARAN, S., RAMOS, D., and GONZALEZ-RODRIGUEZ, J. (2013). Improving short utterance based i-vector speaker recognition using source and utterance-duration normalization techniques. In *In Proceedings of the 14th Annual Conference of the International Speech Communication Association*, pages 2465–2469. International Speech Communication Association (ISCA).

KANAGASUNDARAM, A., VOGT, R., DEAN, D. B., and SRIDHARAN, S. (2012). Plda based speaker recognition on short utterances. In *The Speaker and Language Recognition Workshop (Odyssey 2012)*. ISCA.

KANAGASUNDARAM, A., VOGT, R., DEAN, D. B., SRIDHARAN, S., and MASON, M. W. (2011). I-vector based speaker recognition on short utterances. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 2341–2344.

KENNY, P. (2010). Bayesian speaker verification with heavy tailed priors. In *Speaker and Language Recognition Workshop (Odyssey)*.

KENNY, P., STAFYLAKIS, T., OUELLET, P., ALAM, M. J., and DUMOUCHEL, P. (2013). Plda for speaker verification with utterances of arbitrary duration. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 7649–7653. IEEE.

MANDASARI, M. I., MCLAREN, M., and VAN LEEUWEN, D. A. (2011). Evaluation of i-vector speaker recognition systems for forensic application. In *In Proceedings of the 12th Annual Conference of the International Speech Communication Association*, pages 21–24.

MCLAREN, M. and VAN LEEUWEN, D. (2011). Improved speaker recognition when using i-vectors from multiple speech sources. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5460–5463. IEEE.

MERHAV, N. and LEE, C.-H. (1993). A minimax classification approach with application to robust speech recognition. *Speech and Audio Processing,*, 1(1):90–100.

PRINCE, S. and ELDER, J. (2007). Probabilistic linear discriminant analysis for inferences about identity. In *Computer Vision, ICCV 2007.*

RAO, C. R. (1948). The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2):159–203.

REYNOLDS, D., QUATIERI, T., and DUNN, R. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41.

SARKAR, A. K., MATROUF, D., BOUSQUET, P. M., and BONASTRE, J. F. (2012). Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification. In *In Proceedings of the 13th Annual Conference of the International Speech Communication Association.*