INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Duration mismatch compensation using four-covariance model and deep neural network for speaker verification

Pierre-Michel Bousquet, Mickael Rouvier

University of Avignon - LIA, France

pierre-michel.bousquet@univ-avignon.fr, mickael.rouvier@univ-avignon.fr

Abstract

Duration mismatch between enrollment and test utterances still remains a major concern for reliability of real-life speaker recognition applications. Two approaches are proposed here to deal with this case when using the i-vector representation. The first one is an adaptation of Gaussian Probabilistic Linear Discriminant Analysis (PLDA) modeling, which can be extended to the case of any shift between i-vectors drawn from two distinct distributions. The second one attempts to map i-vectors of truncated segments of an utterance to the i-vector of the full segment, by the use of deep neural networks (DNN). Our results show that both new approaches outperform the standard PLDA by about 10 % relative, noting that these back-end methods could complement those quantifying the i-vector uncertainty during its extraction process, in the case of duration gap.

Index Terms: speaker recognition, i-vector, short utterance, duration mismatch, deep neural networks.

1. Introduction

Real-life speaker recognition applications often impose strong constraints on the amount of data available in test speaker models (for example, in applications on mobile phones). This challenge motivated many studies in the field of speaker recognition relying on the so-called i-vector representation of utterance [1, 2, 3, 4, 5, 6, 7, 8, 9] . The most common scenario is the one in which target speakers are well trained with long speech segments but tested on short segments.

Dealing with the particular case of duration mismatch in i-vector based recognition systems faces two main issues. First, the i-vector paradigm [10] assumes that an utterance can be mapped by a low rank total variability factor. This presupposes that a sufficient amount of acoustic data is available for statistics estimation. This requirement is not fulfilled for short duration utterances, in particular in terms of phonetic content [3]. Several studies underscore that shorter segments tend to produce larger covariances, so that i-vector estimates become less reliable [4, 5]. Also, shortening of speech segments can be thought of as noise [3, 7, 8], which it is pointless capturing into a subspace, as done for speaker or domain variability (JFA [11], PLDA [12], IDVC [13]). Second, duration mismatch induces a shift between the distributions of i-vectors that has to be handled for recognition accuracy.

It turns out that state-of-the-art procedures for i-vector based speaker recognition systems (pre-normalization and PLDA modeling) can be used as unsupervised techniques of duration mismatch compensation. On the one hand, it is shown in [2] that learning PLDA parameters with a training set only comprised of long duration utterances yields the best performance for the case of duration mismatch. Moreover, whitening techniques such as within-class covariance normalization [10] and dimensionality reduction techniques as LDA, when trained with long utterances only, tend to map distribution of short utterances to the one of the latter, which is clearly more discriminative. On the other hand, some studies reveal a shift of magnitude (e.g. Table 1 in [3]), which is canceled by length-normalization.

However, these unsupervised techniques are limited. Simple analyses carried out after them show that the shorter the utterances, the larger their within-speaker covariance matrix and the smaller their between-speaker covariance matrix will be. This result recalls the uncertainty in estimating the i-vector of short duration segments, studied in [4, 5, 9]. These studies handle utterances of arbitrary duration. We focus here on the specific case of duration gap between enrollment and test utterances (e.g. more than 30 sec. vs less than 15 sec.). Two supervised techniques of duration mismatch compensation are proposed to complete the benefit of state-of-the-art normalization and modeling. The first one takes into account the shift of distribution during PLDA modeling, then determines a probabilistic relation between them, delivering a log-likelihood ratio specific to enrollment and test mismatch. The second one attempts to map i-vectors of truncated segments of an utterance to the i-vector of the full segment, regarded as the reference, using a non-linear transformation learned by DNN. These approaches are described in section 3. Our experiments and results are presented in section 4.

2. Gaussian PLDA model

Gaussian Probabilistic Linear Discriminant Analysis (PLDA) assumes that an i-vector $\mathbf{w}$ can be additively decomposed as follows:

$$\begin{aligned} \mathbf{w} &= y + \varepsilon \\ y &\sim \mathcal{N}(\mu, \mathbf{B}) \\ \varepsilon &\sim \mathcal{N}(0, \mathbf{W}) \end{aligned} \tag{1}$$

where $\mathcal{N}$ denotes the normal pdf and the latent variable y, only dependent on the speaker, is statistically independent from the residual term ε. If the speaker factor y is not constrained to lie in an eigenvoice subspace, the model is referred to as two-covariance model [14] [1].

The goal of evaluating hypotheses θ_{tar} that two i-vectors $\mathbf{w}_1$, $\mathbf{w}_2$ (assumed independent given the hidden variables) are produced by the same source and θ_{non} that they are produced by different sources reduces to estimating the log-likelihood ratio

[1] which is often preceded by LDA dimensionality reduction to achieve optimum performance [15].

http://dx.doi.org/10.21437/Interspeech.2017-93

score:

$$\begin{aligned} LLR\left(\mathbf{w}_1, \mathbf{w}_2\right) &= \log \frac{P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{tar}\right)}{P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{non}\right)} \\ &= \log \frac{\int \prod\limits_{i=1,2} P\left(\mathbf{w}_i | y\right) P\left(y\right) dy}{\prod\limits_{i=1,2} \int P\left(\mathbf{w}_i | y\right) P\left(y\right) dy} \end{aligned} \tag{2}$$

Using i-vectors modeled by PLDA with Gaussian priors achieves state-of-the-art results, provided that the i-vectors are first whitened and normalized. The most commonly used techniques are within-class covariance matrix standardization and length-normalization (LW) [10, 16, 17, 18]. These techniques are known to make i-vectors more Gaussian and to reduce the shift between training and unknown datasets.

3. Compensation techniques

Compensation can be done at two levels: scoring or feature representation. Section 3.1 describes our extension of PLDA scoring and section 3.2 presents our transformation of the i-vector representation.

3.1. Four covariance model

Given an i-vector $\mathbf{w}_1$ (resp. $\mathbf{w}_2$) extracted from a long (resp. short) duration speech utterance, the two-covariance models for long and short utterances are, for $i = 1, 2$:

$$\begin{aligned} \mathbf{w}_i &= y_i + \varepsilon_i \\ y_i &\sim \mathcal{N}\left(\mu_i, \mathbf{B}_i\right) \\ \varepsilon_i &\sim \mathcal{N}\left(0, \mathbf{W}_i\right) \end{aligned} \tag{3}$$

To compare $\mathbf{w}_1$ and $\mathbf{w}_2$ for speaker verification, the two distributions have to be probabilistically related. We propose to link the speaker's classes by means of their Gaussian latent variables. Hence, the model assumes that y_1 and y_2 can be related as follows:

$$\begin{aligned} y_2 - \mu_2 &= \mathbf{A}\left(y_1 - \mu_1\right) + \eta \\ \eta &\sim \mathcal{N}\left(0, \mathbf{M}\right) \end{aligned} \tag{4}$$

where the full rank matrices $\mathbf{A}$ and $\mathbf{M}$ have to be estimated. We refer to this model as four-covariance model (4-cov).

Given a development dataset of short and long duration utterances from the same speakers and their punctual estimations of speaker factors, multivariate regression [19] allows to estimate a closed-form expression of $\mathbf{A}$, with which the variance of the residual term η is minimal. For our purposes, the covariance matrices can be weighted by the amount of observations per speaker. This solution also maximizes the likelihood of y_2 given y_1 and assuming normal prior for η. Multivariate regression does not ensure normality of the residue η, and relevance of the 4-cov model must be demonstrated by performance in speaker detection tasks.

Under hypothesis θ_{tar}, the likelihood of $\mathbf{w}_1$, $\mathbf{w}_2$ becomes:

$$P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{tar}\right) = \iint P\left(\mathbf{w}_1, \mathbf{w}_2, y_1, y_2 | \theta_{tar}\right) dy_1 dy_2 \tag{5}$$

This likelihood can be written as:

$$\iint P\left(\mathbf{w}_1 | y_1\right) P\left(\mathbf{w}_2 | y_2\right) P\left(y_2 | y_1\right) P\left(y_1\right) dy_1 dy_2 \tag{6}$$

using some conditional independence between these variables. The final expression of the log-likelihood is:

$$\begin{aligned} \log P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{tar}\right) =& c + \frac{1}{2}\mathbf{w}_1^t \mathbf{N}_{11}^{tar} \mathbf{w}_1 + \frac{1}{2}\mathbf{w}_2^t \mathbf{N}_{22}^{tar} \mathbf{w}_2 \\ &+ \mathbf{w}_1^t \mathbf{N}_{12}^{tar} \mathbf{w}_2 + n_1^{tar} \mathbf{w}_1 + n_2^{tar} \mathbf{w}_2 \end{aligned} \tag{7}$$

The scalar c is a constant and

$$\begin{aligned} \mathbf{N}_{11}^{tar} &= \mathbf{W}_1^{-1} \mathbf{R}^t \mathbf{Q}^{-1} \mathbf{R} \mathbf{W}_1^{-1} - \mathbf{W}_1^{-1} + \mathbf{W}_1^{-1} \mathbf{P}^{-1} \mathbf{W}_1^{-1} \\ \mathbf{N}_{22}^{tar} &= \mathbf{W}_2^{-1} \mathbf{Q}^{-1} \mathbf{W}_2^{-1} - \mathbf{W}_2^{-1} \\ \mathbf{N}_{12}^{tar} &= \mathbf{W}_1^{-1} \mathbf{R}^t \mathbf{Q}^{-1} \mathbf{W}_2^{-1} \\ n_1^{tar} &= -\left(a - \mathbf{R}b\right)^t \mathbf{Q}^{-1} \mathbf{R} \mathbf{W}_1^{-1} + 2b^t \mathbf{P}^{-1} \mathbf{W}_1^{-1} \\ n_2^{tar} &= -\left(a - \mathbf{R}b\right)^t \mathbf{Q}^{-1} \mathbf{W}_2^{-1} \end{aligned}$$

where

$$\begin{aligned} \mathbf{M} &= \mathbf{B}_2 - \mathbf{A}\mathbf{B}_1\mathbf{A}^t \\ \mathbf{P} &= \mathbf{W}_1^{-1} + \mathbf{B}_1^{-1} + \mathbf{A}^t \mathbf{M}^{-1} \mathbf{A} \\ \mathbf{Q} &= \mathbf{M}^{-1} - \mathbf{M}^{-1} \mathbf{A} \mathbf{P}^{-1} \mathbf{A}^t \mathbf{M}^{-1} + \mathbf{W}_2^{-1} \\ \mathbf{R} &= \mathbf{M}^{-1} \mathbf{A} \mathbf{P}^{-1} \\ a &= \mathbf{M}^{-1}\left(\mathbf{A}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \\ b &= \mathbf{B}_1^{-1}\boldsymbol{\mu}_1 + \mathbf{A}^t \mathbf{M}^{-1}\left(\mathbf{A}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \\ c &= \boldsymbol{\mu}_1^t \mathbf{B}_1^{-1} \boldsymbol{\mu}_1 + \left(\mathbf{A}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right)^t \mathbf{M}^{-1}\left(\mathbf{A}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\right) \end{aligned}$$

Under hypothesis θ_{non}, the likelihood of $\mathbf{w}_1$, $\mathbf{w}_2$ becomes:

$$P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{non}\right) = \prod\nolimits_{i=1,2} \int P\left(\mathbf{w}_i | y_i\right) P\left(y_i\right) dy_i \tag{8}$$

The explicit solution is:

$$\begin{aligned} \log P\left(\mathbf{w}_1, \mathbf{w}_2 | \theta_{non}\right) =& c + \frac{1}{2}\mathbf{w}_1^t N_{11}^{non} \mathbf{w}_1 + \frac{1}{2}\mathbf{w}_2^t N_{22}^{non} \mathbf{w}_2 \\ &+ n_1^{non} \mathbf{w}_1 + n_2^{non} \mathbf{w}_2 \end{aligned} \tag{9}$$

where c is a constant and, for $i = 1, 2$

$$\begin{aligned} N_{ii}^{non} &= \mathbf{W}_i^{-1}\left(\mathbf{B}_i^{-1} + \mathbf{W}_i^{-1}\right)^{-1} \mathbf{W}_i^{-1} - \mathbf{W}_i^{-1} \\ n_i^{non} &= \left(\mathbf{B}_i^{-1}\mu_i\right)^t \left(\mathbf{B}_i^{-1} + \mathbf{W}_i^{-1}\right)^{-1} \mathbf{W}_i^{-1} \end{aligned}$$

It should be mentioned that the log-likelihood ratio can also be obtained by adapting formulations of equations (18.24) and (18.25) in [12].

3.2. Deep neural network

Another way to deal with duration mismatch consists in estimating a transformation between i-vectors of cut segments of an utterance and the i-vector of the full segment. In [20], a linear transformation, estimated by MMSE linear regression, is proposed for more accurate comparison, when enrollment and test data are both of short duration. In [21, 22], DNN models aim to reduce the nuisance variability by transforming each i-vector into the mean vector of its speaker's class. Inspired by this approach, we propose to use DNN for training a non-linear transform, which is expected to reduce the shift between distributions of short and long utterance i-vector.

Details of this DNN model are depicted in Figure 1. An utterance is cut in n short segments, then their n i-vectors are extracted and used as DNN inputs. The i-vector of the full segment is used as unique DNN desired output ("target") of the

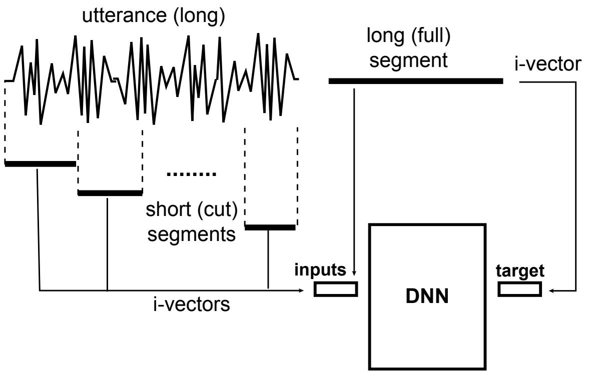

Figure 1: *Illustration of the deep neural network of section 3.2.*

latter n i-vectors (and also used as input, in order to regularize the training phase).

The outputs of this model will be handled as transformed i-vectors, in which the within-session variability has been partially removed. Since many studies revealed the radial distribution of speaker's classes from the origin [10, 23], the loss-function of the DNN is not the mean squared error but the cosine proximity, as proposed in [21]. This loss function does not control the i-vector-length, hence batch normalization [24] is used to regularize each hidden layer, maintaining the mean activation close to 0 and the activation standard deviation close to 1.

4. Experiments

4.1. Evaluation data

To carry out experiments with duration mismatch, an evaluation set is devised by randomly truncating the test utterances into continuous and successive short segments in the female, telephone speech portion of the NIST 2010 extended core condition det 5, so that the durations after voice activity detection of all test utterances lay in the range 500–1500 active frames (5 to 15 seconds of speech). Each trial compares one enrollment i-vector to one of the test cut segment i-vectors, randomly picked up. The experiment is iterated 1000 times, providing distributions of performance measurements. Three performance metrics are reported: the equal error rate (EER), the normalized minimum detection cost function (DCF) with the probability of a target trial set to 0.01 and the cost of misses and false alarms set to 1 (a trade-off between 2008 and 2010 NIST detection costs), and the C_{llr} [25] calibrated on a development set.

4.2. i-vector/PLDA training

Our experiments operate on 20 MFCC parameters (including log-energy) augmented with 20 first (Δ) and 20 second ($\Delta\Delta$) derivatives, providing 60 dimensional feature vectors. A cepstral mean normalization is applied using a sliding window size of 3 seconds. The low-energy frame (corresponding mainly to silence) are removed. The low-energy algorithm is based on thresholding the log-energy and taking the consensus of threshold decisions within a window of 11 frames centered on the current frame. Gender-dependent 2048 full component UBM and total variability matrix of low rank 600 are trained on NIST SRE 2004, 2005, 2006.

A 600 dimensional i-vector extractor is trained using 27213 utterances of more than 30 seconds, from 1625 speakers, female only. The same truncation procedure than for test utterances is done, providing a dataset of 479859 i-vectors of cut segments. For 4-cov and DNN models, LW-normalization and LDA dimensionality reduction (to $r = 100$) are applied, using the long duration i-vector subset for training.

Table 1: *Results on the initial NIST 2010 core condition (female det 5), without duration mismatch.*

long vs long	EER (%)	minDCF	C_{llr}
PLDA	2.05	0.248	0.147

Table 2: *Results obtained by the different systems for experiments with duration mismatch. The last three lines correspond to the proposed approaches.*

long vs short	EER (%)	minDCF	C_{llr}
PLDA (train: overall)	10.25	0.757	0.426
PLDA (train: long only)	7.33	0.650	0.288
MMSE+PLDA	7.13	0.644	0.284
4-cov	6.71	0.611	0.273
DNN+PLDA	6.69	0.620	0.257
DNN+4-cov	6.64	0.613	0.266

4.3. A concern about statistical independence

PLDA modeling assumes that the observations of a training speaker are independent. This hypothesis is not fulfilled for cut segments of a given utterance. To alleviate this bias, PLDA and 4-cov modelings need some modifications. Given a long duration utterance, the weight of its n-size subset of cut segments, equal to n, is replaced by 1 in all the formulas of the EM-ML phase (latent variable statistics, covariance matrices).

4.4. DNN configuration

For DNN model, dropout technique [26] is used, which helps prevent overfitting. The learning rate is decreased at each iteration, allowing quickly learning good weights early and fine tuning them later. The DNN model is comprised of two hidden layers with 1500 units using the sigmoid activation function and a linear output layer with the same number of units than the i-vector size. The network is trained using the Keras algorithms [27].

4.5. Results

Table 1 reports performance on the initial NIST 2010 core condition (female det 5) with full test segments, to better assess degradation of performance caused by short duration test utterances. Results of experiments with duration mismatch are given in Table 2, in terms of average performance. As explained above, experiments are iterated 1000 times. To better compare distributions of performance metrics, Figure 2 shows the box plots of the four systems on which we focus.

Two PLDA based systems are reported as state-of-the-art for our experiments. The first one (row 1 of the Table) uses the overall dataset of available i-vectors, long and short duration, for training. The second one (row 2 of the Table) only uses the long duration i-vector subset. The optimal rank for PLDA speaker subspace is $r = 100$. Comparison of the systems confirms the remark in the introduction, about relevance of learning PLDA model with long duration data only. The slight gain of

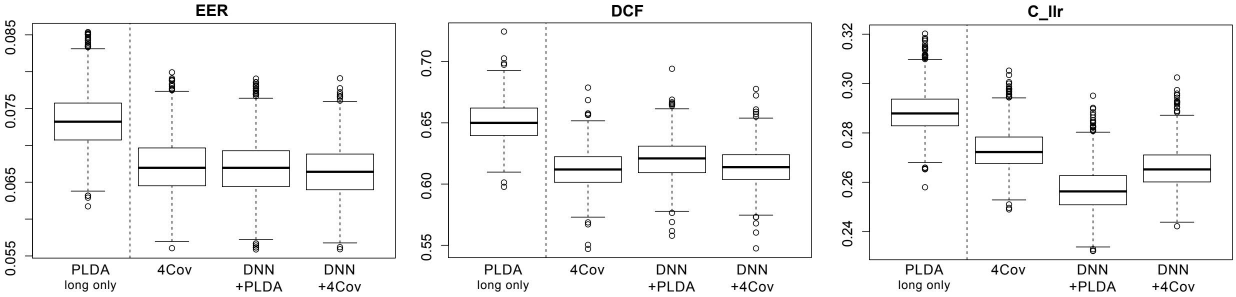

Figure 2: *Comparing distributions of performance: the vertical line separates the best state-of-the-art system and the proposed approaches.*

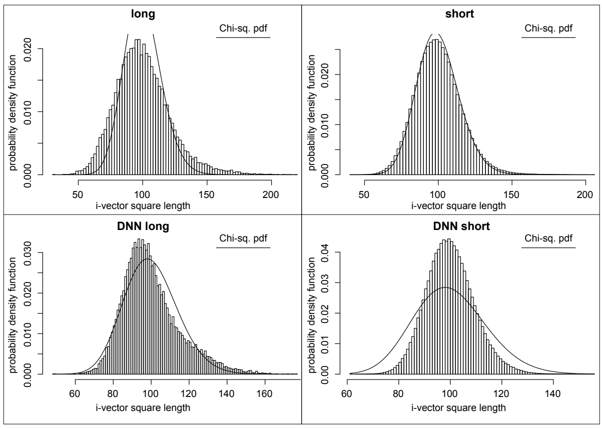

Figure 3: *Histograms of the standardized i-vector square length distribution for long and short duration data, initially and after DNN. The pdf of a χ^2 distribution with 100 degrees of freedom is also depicted.*

performance provided by MMSE approach (row 3) recalls that this method has not been designed for the particular case of duration mismatch.

Row 4 of the Table shows results of the four-covariance model outlined in section 3.1. Compared to the state-of-the-art system of row 2, this model yields a significant gain of performance. The DNN model (DNN+PLDA row 5) outlined in section 3.2 performs similarly to 4-cov model. Analyses of variance (ANOVA) confirm the statistical significance of the gain for both approaches and performance metrics (all p-values of F-test $<$ 2e-16).

The last system combines DNN-based transformation and 4-cov model. Results are similar to those of distinct approaches. This raises an issue, since 4-cov model is designed to take into account the shift of distributions and should be more accurate. This disappointing result may be due to the lack of Gaussianity after DNN-based transformation, limiting the benefit of a too complex Gaussian model. This hypothesis is confirmed by Figure 3, which plots histograms of the standardized i-vector square length distribution for our long and short duration data, initially (when reduced by LDA) and after DNN. If vectors draw a Gaussian distribution, the values are supposed to follow a χ^2 distribution with r d.o.f (pdf also plotted in the Figure). Data after DNN do not match the latter, with slight skewness for long duration data and, above all, a too light tail for short duration data (unlike PLDA, the 4-cov modeling is depending on this distribution). This observation underscores that regarding DNN-output vectors as observations from a probabilistic generative model ignores the process by which they were obtained.

5. Conclusion

The particular case of duration mismatch in speaker recognition relying on the i-vector paradigm, when the amount of data for speaker enrollment is sufficient but the test data is limited to a short duration segment, must tackle two main issues: the expected shift between the distributions of i-vectors has to be taken into account and the lack of information in short duration data involves a growth of nuisance variability that cannot be removed or at least reduced by eigenvoice-like methods.

We explored two approaches for compensating duration mismatch in i-vector based speaker recognition systems. The first one adapts PLDA modeling to the specific case of mismatched data ("four-covariance model") and therefore provides for a better fit with the underlying distributions. The second one relies on DNN to map the i-vectors of cut segments of an utterance to the one of the full segment. These two approaches can be combined, as the first one fits the distributions whilst the second one transforms the i-vector representation. Experiments carried out with these two approaches show that both significantly improve accuracy of detection, in terms of all the performance metrics we tested, but that combining them does not lead to a gain in accuracy. We reveal the lack of Gaussianity after DNN transformation that probably explains this outcome.

The four-covariance model could be an additional scoring process for models embedding the intrinsic ivector uncertainty [4, 5, 9], in the case of duration gap. Future work will test this opportunity. More generally, the four-covariance model is designed to deal with any case of mismatch between i-vectors to compare for speaker verification. Its use for other mismatches (language, channel) should be assessed in future work. The only limitation of this model is the availability of a training set comprising the two conditions for more than r speakers, where r is the i-vector size after dimensionality reduction (100 in our study).

6. References

[1] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector Based Speaker Recognition on Short Utterances," in *International Conference on Speech Communication and Technology*, 2011.

[2] 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 *International Conference on Speech Communication and Technology*, 2012.

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

[4] P. Kenny, T. Stafylakis, P. Ouellet, M. J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *International Conference on Speech Communication and Technology*, 2013.

[5] S. Cumani, P. Laface, and O. Plchot, "On the use of i-vector posterior distributions in probabilistic linear discriminant analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 846–857, 2014.

[6] P. L. S. Martinez, B. G. B. Fauve, A. Larcher, and J. S. D. Mason, "Speaker verification performance with constrained durations," in *2nd International Workshop on Biometrics and Forensics, IWBF*, 2014, pp. 1–6.

[7] A. Nautsch, C. Rathgeb, C. Busch, H. Reininger, and K. Kasper, "Towards duration invariance of i-vector-based adaptive score normalization," in *Odyssey: The Speaker and Language Recognition Workshop*, 2014.

[8] M. I. Mandasari, R. Saeidi, and D. A. van Leeuwen, "Quality measures based calibration with duration and noise dependency for speaker recognition," *Speech Communication*, vol. 72, no. Complete, pp. 126–137, 2015.

[9] S. Cumani, "Fast scoring of full posterior PLDA models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2036–2045, 2015.

[10] N. Dehak, P. 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.

[11] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[12] S. J. D. Prince, *Computer Vision: Models, Learning, and Inference*, 1st ed. New York, NY, USA: Cambridge University Press, 2012.

[13] H. Aronowitz, "Compensating inter-dataset variability in plda hyper-parameters for robust speaker recognition," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2014.

[14] N. Brummer and E. de Villiers, "The speaker partitioning problem," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[15] N. Brummer, J. Villalba, and E. Lleida, "Fully Bayesian likelihood ratio vs i-vector length normalization in speaker recognition systems," in *NIST SRE Analysis Workshop*, 2011.

[16] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *International Conference on Speech Communication and Technology*, 2011, pp. 249–252.

[17] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in *International Conference on Speech Communication and Technology*, 2011, pp. 485–488.

[18] P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, and O. Plchot, "Variance-Spectra based Normalization for I-vector Standard and Probabilistic Linear Discriminant Analysis," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2012.

[19] A. C. Rencher and W. F. Christensen, *Methods of multivariate analysis*. Wiley, 2012.

[20] W. B. Kheder, D. Matrouf, M. Ajili, and J.-F. Bonastre, "Probabilistic approach using joint long and short session i-vectors modeling to deal with short utterances for speaker recognition," in *International Conference on Speech Communication and Technology*, 2016.

[21] G. Bhattacharya, J. Alam, P. Kenny, and V. Gupta, "Modelling speaker and channel variability using deep neural networks for robust speaker verification," in *Workshop on Spoken Language Technology (IEEE SLT)*, 2016.

[22] T. Pekhovsky, S. Novoselov, A. Sholohov, and O. Kudashev, "On autoencoders in the i-vector space for speaker recognition," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, pp. 217–224.

[23] N. Dehak, Z. N. Karam, D. A. Reynolds, R. Dehak, W. M. Campbell, and J. R. Glass, "A channel-blind system for speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2011.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *CoRR*, vol. abs/1502.03167, 2015. [Online]. Available: http://arxiv.org/abs/1502.03167

[25] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech & Language*, vol. 20, no. 2-3, pp. 230–275, 2006.

[26] A. K. I. S. Nitish Srivastava, Geoffrey E Hinton and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. no. 1, pp. 1929–1958, 2014.

[27] F. Chollet, "keras," https://github.com/fchollet/keras, 2015.