Incorporating Duration Information into I-Vector-Based Speaker-Recognition Systems

Boštjan Vesnicer, Jerneja Žganec-Gros

Alpineon d.o.o.
Ulica Iga Grudna 15
SI-1000 Ljubljana
Slovenia

Simon Dobrišek, Vitomir Štruc

Faculty of Electrical Engineering
University of Ljubljana
Tržaška cesta 25
SI-1000 Ljubljana, Slovenia

Abstract

Most of the existing literature on i-vector-based speaker recognition focuses on recognition problems, where i-vectors are extracted from speech recordings of sufficient length. The majority of modeling/recognition techniques therefore simply ignores the fact that the i-vectors are most likely estimated unreliably when short recordings are used for their computation. Only recently, were a number of solutions proposed in the literature to address the problem of duration variability, all treating the i-vector as a random variable whose posterior distribution can be parameterized by the posterior mean and the posterior covariance. In this setting the covariance matrix serves as a measure of uncertainty that is related to the length of the available recording. In contract to these solutions, we address the problem of duration variability through weighted statistics. We demonstrate in the paper how established feature transformation techniques regularly used in the area of speaker recognition, such as PCA or WCCN, can be modified to take duration into account. We evaluate our weighting scheme in the scope of the i-vector challenge organized as part of the *Odyssey, Speaker and Language Recognition Workshop 2014* and achieve a minimal DCF of 0.280, which at the time of writing puts our approach in third place among all the participating institutions.

1. Introduction

The area of speaker recognition has made significant progress over recent years. Today, recognition systems relying on so-called i-vectors have emerged as the de-facto standard in this area. Most of the existing literature on i-vector-based speaker recognition focuses on recognition problems, where the i-vectors are extracted from speech recordings of sufficient length. The length of the recordings is predefined by the speech corpus used for the experimentation and typically does not drop below a length that would cause problems to the recognition techniques. In practical applications, however, speaker recognition systems often deal with i-vectors extracted from short recordings, which may be estimated less reliably than i-vectors extracted from recordings of sufficient length.

The problem of duration variability is known to be one of importance for practical speaker-recognition applications and has also been addressed to a certain extent in the literature in the context of i-vector-based speaker-recognition systems, e.g., [1], [2], [3], [3], [4], [5], [6], [7], [8], [9], [10].The most recent solutions of the duration-variability problem (e.g., [5], [6], or [7] do not treat i-vectors as point estimates of the hidden variables in the eigenvoice model, but rather as random vectors. In this slightly different perspective, the i-vectors appears as posterior distributions, parameterized by the posterior mean and the posterior covariance matrix. Here, the covariance matrix can be interpreted as a measure of the uncertainty of the point estimate that relates to the duration of the speech recording used to compute the i-vectors.

In this paper we propose a slightly different approach and try to compensate for the problem of duration variability of the speech recordings through weighted statistics. Typically, feature-transformation techniques commonly used in the area of speaker recognition, such as principal component analysis (PCA) or within-class covariance normalization (WCCN) estimate the covariance matrices and sample means by considering the contribution of each available i-vector equally in the statistics, regardless of the fact that the i-vectors may be estimated unreliably. To address this point, we associate with every i-vector a weight that is proportional to the duration of the speech recording from which the i-vector was extracted. This weight is then used to control the impact of a given i-vector to the overall statistics being computed. The described procedure can be applied to any feature transformation technique and results in duration-weighted techniques that should lead to better estimates of the feature transforms.

We evaluate the proposed weighting scheme in the scope of the i-vector challenge (IVC) organized by NIST as part of the *Odyssey, Speaker and Language Recognition Workshop 2014*. The goal of the challenge is to advance the state-of-technology in the area of speaker recognition by providing a standard experimental protocol and pre-computed i-vectors for experimentation. Based on the data provided by the challenge, we show that it is possible to apply the proposed weighting scheme to supervised as well as unsupervised feature-transformation techniques and that in both cases performance gains can be expected. With our best performing (duration-weighted) system we managed to achieve a minimal decision-cost-function (DCF) value of 0.280, which puts our approach in third place among the participating institutions (and in seventh place individually out of 98 partici-

This work was supported in parts by the national research program P2-0250(C) Metrology and Biometric Systems, the European Union's Seventh Framework Programme (FP7-SEC-2011.20.6) under grant agreement number 285582 (RESPECT), the Eureka project S-Verify (contract No. 2130-13-090145) and by the European Union, European Regional Fund, within the scope of the framework of the Operational Programme for Strengthening Regional Development Potentials for the Period 2007-2013, contract No. 3330-13-500310 (eCall4All). The authors additionally appreciate the support of COST Actions IC1106 and IC1206.

pants) at the time of writing.

Before we conclude this section, let us summarize the contributions of this paper:

- we propose a novel weighting scheme to address the problem of variable durations of the speech recordings used to compute i-vectors from,
- we introduce duration-weighted versions of established feature-transformation techniques, namely, PCA and WCCN, and
- we present a detailed experimental assessment of the proposed duration-weighted techniques and benchmark them against state-of-the-art speaker-recognition techniques submitted for evaluation at the 2014 i-vector challenge organized by NIST.

The rest of the paper is structured as follows. In Section 2 we briefly survey the state-of-the-art in the field of speaker recognition and introduce all the techniques relevant for the remainder of the paper. In Section 3 we present our duration-based weighting scheme and show how it can be applied to established feature-transformation techniques used regularly in the field of speaker recognition. In Section 4 we describe the i-vector challenge, its goals, the experimental data and performance metrics used to measure the recognition performance of the participating systems. We assess the proposed weighting scheme in Section 5 and conclude the paper with some final comments in Section 6.

2. Prior work

I-vectors represent low-dimensional feature representations of variable length speech. The i-vector extraction procedure can be seen as an extension of the well-known GMM-UBM modeling of the short-time acoustic features [11], where each speech utterance is represented by the (MAP-adapted) parameters[1] of the UBM model. The main difference between the i-vector extraction procedure and the GMM-UBM modeling approach is that instead of the classical MAP algorithm, an i-vector extractor uses a generalized version of the same algorithm, which takes the dependence of the parameters into account. The algorithm is — depending on the context — known by different names like eigenvoice MAP or total variability modeling and is in fact a slightly modified version of the classical factor analysis[2]. Moreover, the algorithm is a special case of the joint factor analysis [12], which tries to model speaker and channel variability in the supervectors' space. To avoid the complications that arise from the fact that the dimension of supervectors is usually very large, the total variability model takes a different approach and does not try to disentangle the speaker and channel effects by itself, but postpones this task to the subsequent steps.

Two of the most frequently used classification methods in i-vector-based speaker recognition are the cosine similarity [13] and probabilistic linear discriminant analysis (PLDA), independently developed for face [14], [15] and speaker recognition [16]. Since its introduction, the PLDA model has been extended in different ways, e.g. the underlying Gaussian assumption have been relaxed [16], the parameters of the model have been treated as random variables [17] and an extension to the mixture case has been proposed as well [18].

[1]The mean vectors of individual Gaussian components can be stacked on top of each other, forming the so called supervectors.

[2]The modification is needed due to the fact that the parameters of the GMM are not directly observed and should be treated as latent variables.

Before given to the classifier, i-vectors are usually preprocessed in various ways. Common preprocessing methods include whitening (PCA), linear discriminant analysis (LDA) and within-class covariance normalization (WCCN), which can be applied in combination. Another important preprocessing step is length normalization, as it turns out [19] that length normalization brings the i-vectors closer to a normal distribution and therefore provides for a better fit with the assumptions underlying Gaussian PLDA.

3. Duration-based weighting

3.1. Introduction

In this section we introduce our duration-dependent weighting scheme. We assume that the front-end processing of the speech recording has already been conducted and that all we have at our disposal is a set of extracted i-vectors and a single item of metadata in the form of the duration of the recording from which a given i-vector was extracted [20]. Under the presented assumptions the solutions to the problem of duration variability that treat the i-vectors as random variables characterized by a posterior distribution, such as those presented in [5], [6], or [7], are not applicable.

Most feature-extraction (or feature-transformation) techniques used in conjunction with i-vector-based speaker-verification systems (e.g., PCA, WCCN, NAP, etc.) rely on estimates of the first- and second-order statistics to compute the feature transforms. Given some training i-vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i \in \mathbb{R}^m$ and $i = 1, 2, \ldots, n$, the first- ($\mathbf{f}$) and second-order ($\mathbf{S}$) statistics are defined as:

$$\mathbf{f} = \sum_{i}^{n} \mathbf{x}_i \tag{1}$$

and

$$\mathbf{S} = \sum_{i}^{n} \mathbf{x}_i \mathbf{x}_i^T, \tag{2}$$

where ^{T} denotes the transpose operator. Based on these statistics it is straight forward to compute the sample covariance matrix ($\boldsymbol{\Sigma}_s$) and sample mean $\boldsymbol{\mu}_s$, which are at the heart of many feature extraction techniques:

$$\boldsymbol{\mu}_s = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i = \frac{1}{n} \mathbf{f}, \tag{3}$$

$$\boldsymbol{\Sigma}_s = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_s)(\mathbf{x}_i - \boldsymbol{\mu}_s)^T = \frac{\mathbf{S}}{n} - \boldsymbol{\mu}_s \boldsymbol{\mu}_s^T, \tag{4}$$

Note that in the case, where all the training vectors $\mathbf{x}_i$ belong to the same class (i.e., to the same speaker), the above equations represent the class-conditional mean and the class-conditional covariance matrix. In the remainder we will limit our discussion on the presented definitions of the sample mean and the covariance matrix. Note, however, that the same reasoning can be applied to any statistics computed from $\mathbf{f}$ and $\mathbf{S}$.

The definitions of the covariance matrix and sample mean given in Eqs. (4) and (3) assume that all the training vectors $\mathbf{x}_i$ ($i = 1, 2, \ldots, n$) are equally reliable and are, therefore, given equal weights when computing the mean and covariance matrix. While such an interpretation of the equations is (most likely) valid if the training vectors are computed from speech recordings of sufficient length, this may not be true if some of the vectors are extracted from short recordings. In this case, some

of the training vectors are unreliable and should not contribute equally to the computed statistics.

To account for the above observation we propose to use weighted statistics instead of the statistics in Eqs. (1) and (2), where the weight associated with the i-th sample is defined by the duration of the recording from which the vector was extracted. To formalize our weighting scheme, let us assume that each of the available training vectors $\mathbf{x}_i$ also has an associated data instance t_i, defining the duration from which the vector was extracted ($i = 1, 2, \ldots, n$). Based on this additional data, we can define duration-weighted versions of zero ($\mathbf{T}_d$), first- ($\mathbf{f}_d$) and second-order ($\mathbf{S}_d$) statistics:

$$T_d = \sum_{i=1}^{n} t_i, \tag{5}$$

$$\mathbf{f}_d = \sum_{i=1}^{n} t_i \mathbf{x}_i, \tag{6}$$

$$\mathbf{S}_d = \sum_{i=1}^{n} t_i \mathbf{x}_i \mathbf{x}_i^T, \tag{7}$$

and consequently, a duration-weighted sample mean and covariance matrix:

$$\boldsymbol{\mu}_d = \sum_{i=1}^{n} \frac{t_i}{T_d} \mathbf{x}_i = \frac{1}{T_d} \mathbf{f}_d, \tag{8}$$

$$\boldsymbol{\Sigma}_d = \sum_{i=1}^{n} \frac{t_i}{T_d} (\mathbf{x}_i - \boldsymbol{\mu}_d)(\mathbf{x}_i - \boldsymbol{\mu}_d)^T = \frac{\mathbf{S}_d}{T_d} - \boldsymbol{\mu}_d \boldsymbol{\mu}_d^T, \tag{9}$$

Note that all the presented statistics are reduced to their non-weighted versions if the speech recordings, from which the training vectors are extracted, are of the same length. If this is not the case, the presented weighting scheme gives larger emphasis to more reliably estimated i-vectors. In the remainder, we present modifications of two popular feature-transformation techniques based on the presented weighting scheme, namely, principal component analysis and within-class covariance normalization. We first briefly describe the theoretical basis of both techniques and then show, how they can be modified based on the presented statistics.

3.2. Principal component analysis

Principal component analysis (PCA) is a powerful statistical learning technique with applications in many different areas, including speaker verification. PCA learns a subspace from some training data in such a way that the learned basis vectors correspond to the maximum variance directions present in the original training data [21]. Once the subspace is learned, any given feature vector can be projected into the subspace to be processed further or to be used with the selected scoring procedure. In state-of-the-art speaker-verification systems the feature vectors used with PCA typically take the form of i-vectors, which after processing with the presented technique are fed to a scoring technique, based on which identity inference is conducted.

Formally PCA can be defined as follows. Given a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^m$ containing in its columns n training vectors $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$, PCA computes a subspace basis $\mathbf{U} \in \mathbb{R}^{m \times d}$ by factorizing of the covariance matrix $\boldsymbol{\Sigma}$ of the vectors in $\mathbf{X}$ into the following form:

$$\boldsymbol{\Sigma} = \mathbf{U} \boldsymbol{\Lambda} \mathbf{U}^T, \tag{10}$$

where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_d], \mathbf{u}_i \in \mathbb{R}^m$ denotes an orthogonal eigenvector vector matrix (i.e., the projection basis) and $\boldsymbol{\Lambda} = diag\{\lambda_1, \lambda_2, \ldots, \lambda_d, \}$ stands for a diagonal eigenvalue matrix with the eigenvalues arranged in decreasing order. Note that if $\boldsymbol{\Sigma}$ is full-rank the maximum possible value for the subspace dimensionality is $d = n$, if the covariance matrix is not full-rank the upper bound for d is defined by the number of non-zero eigenvalues in $\boldsymbol{\Lambda}$. In practice, the dimensionality of the PCA subspace d is an open parameter and can be selected arbitrarily (up to the upper bound).

Based on the computed subspace basis, a given feature vector $\mathbf{x}$ can be projected onto the $d-$dimensional PCA subspace using the following mapping:

$$\mathbf{y} = \mathbf{U}^T (\mathbf{x} - \boldsymbol{\mu}), \tag{11}$$

where $\mathbf{y} \in \mathbb{R}^d$ stands for the PCA transformed feature vector.

Commonly, the above transformation is implemented in a slightly different form, which next to projecting the given feature vector $\mathbf{x}$ into the PCA subspace, also whitens the data:

$$\mathbf{y} = (\mathbf{U} \boldsymbol{\Lambda}^{-1/2})^T (\mathbf{x} - \boldsymbol{\mu}). \tag{12}$$

Note that with standard PCA the covariance matrix $\boldsymbol{\Sigma}$ and sample mean $\boldsymbol{\mu}$ in Eqs. (10), (11) and (12) are computed based on non-weighted statistics, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_s$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_s$. If the duration-weighted statistics are used instead, i.e., $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_d$ and $\boldsymbol{\mu} = \boldsymbol{\mu}_d$, we obtain a modified version of PCA, which takes duration into account when computing the subspace basis.

3.3. Within-class covariance normalization

Within-Class Covariance Normalization (WCCN) is a feature transformation technique originally introduced in the context of Support Vector Machine (SVM) classification [22]. WCCN can under certain conditions be shown to minimize the expected classification error[3] by applying a feature transformation on the data that as a result whitens the within-class scatter matrix of the training vectors. Thus, unlike PCA, WCCN represents a supervised feature extraction/transformation technique and requires the training data to be labeled. In state-of-the-art speaker verification systems, the feature vectors used with WCCN typically represent i-vectors (or PCA-processed i-vectors) that after the WCCN feature transformation are subjected to a scoring procedure.

Typically WCCN is implemented as follows. Consider a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbb{R}^m$ containing in its columns n training vectors $\mathbf{x}_i$, for $i = 1, 2, \ldots, n$, and let us further assume that these vectors belong to N distinct classes[4] $C_1, C_2, \ldots, C_N$ with the j-th class containing n_j samples and $n = \sum_{j=1}^{N} n_j$. WCCN computes the transformation matrix based on the following Cholesky factorization:

$$\boldsymbol{\Sigma}_w^{-1} = \mathbf{L} \mathbf{L}^T, \tag{13}$$

where $\mathbf{L}$ and $\mathbf{L}^T$ stand for the lower and upper triangular matrices, respectively, and $\boldsymbol{\Sigma}_w^{-1}$ denotes the inverse of the within-class scatter matrix computed from the training data.

Once computed, the WCCN transformation matrix $\mathbf{L}$ can be used to transform any given feature vector $\mathbf{x}$ based on the following mapping:

$$\mathbf{y} = \mathbf{L}^T \mathbf{x}, \tag{14}$$

where $\mathbf{y} \in \mathbb{R}^m$ stands for the transformed feature vector.

[3] on the training data

[4] Note that for the weighted case, presented in the remainder, we actually assume that the classes contain pairs of feature vectors and associated duration-data instances, i.e., $(\mathbf{x}_i, t_i)$.

Commonly, the within-class scatter matrix $\mathbf{\Sigma}_w$ is computed based on class-conditional (i.e., speaker-conditional) first- ($\mathbf{f}_j$) and second-order ($\mathbf{S}_j$) statistics. The expressions for computing these statistics for the j-th class C_j are defined as:

$$\mathbf{f}_j = \sum_{\substack{i=1 \\ \mathbf{x}_i \in C_j}}^{n_j} \mathbf{x}_i, \text{ and } \mathbf{S}_j = \sum_{\substack{i=1 \\ \mathbf{x}_i \in C_j}}^{n_j} \mathbf{x}_i \mathbf{x}_i^T, \tag{15}$$

which results in the following within-class scatter matrix:

$$\begin{aligned} \mathbf{\Sigma}_{ws} &= \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n_j} (\mathbf{S}_j - \frac{1}{n_j} \mathbf{f}_j \mathbf{f}_j^T) \\ &= \frac{1}{N} \sum_{j=1}^{N} \frac{1}{n_j} \sum_{\substack{i=1 \\ \mathbf{x}_i \in C_j}}^{n_j} (\mathbf{x}_i - \boldsymbol{\mu}_j)(\mathbf{x}_i - \boldsymbol{\mu}_j)^T, \end{aligned} \tag{16}$$

where $\boldsymbol{\mu}_j$ denotes the class conditional mean for the j-th class.

For the weighted-version of WCCN relying on our duration-dependent weighting scheme the class-conditional zero (T_{dj}), first- ($\mathbf{f}_{dj}$) and second-order ($\mathbf{S}_{dj}$) statistics are defined as:

$$T_{dj} = \sum_{\substack{i=1 \\ t_i \in C_j}}^{n_j} t_i, \text{ where } T_d = \sum_{j=1}^{N} T_{dj}, \tag{17}$$

$$\mathbf{f}_{dj} = \sum_{\substack{i=1 \\ \mathbf{x}_i \in C_j}}^{n_j} t_i \mathbf{x}_i, \text{ and } \mathbf{S}_{dj} = \sum_{\substack{i=1 \\ \mathbf{x}_i \in C_j}}^{n_j} t_i \mathbf{x}_i \mathbf{x}_i^T, \tag{18}$$

where we assume that $(\mathbf{x}_i, t_i) \in C_j$.

With these definitions the weighted within-class scatter matrix $\mathbf{\Sigma}_{wd}$ can be defined as:

$$\begin{aligned} \mathbf{\Sigma}_{wd} &= \sum_{j=1}^{N} \frac{1}{T_d} (\mathbf{S}_{dj} - \frac{1}{T_{dj}} \mathbf{f}_{dj} \mathbf{f}_{dj}^T) \\ &= \sum_{j=1}^{N} \frac{T_{dj}}{T_d} \sum_{\substack{i=1 \\ \mathbf{x}_i \in C_j}}^{n_j} \frac{t_i}{T_{dj}} (\mathbf{x}_i - \boldsymbol{\mu}_{dj})(\mathbf{x}_i - \boldsymbol{\mu}_{dj})^T, \end{aligned} \tag{19}$$

where $\boldsymbol{\mu}_{dj}$ denotes the duration-weighted class conditional mean for the j-th class.

Similar to the PCA case, factorizing the inverse of the standard within-class scatter matrix (i.e., $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}_{ws}^{-1}$) based on Eq. 14 results in the classical implementation of WCCN, while using the weighted version (i.e., $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}_{wd}^{-1}$) results in the modified duration-weighted implementation of WCCN.

4. The I-vector challenge

We evaluate the feasibility of the proposed duration-weighted scheme in the scope of the i-vector challenge (IVC) organized by NIST as part of the Odyssey, Speaker and Language Recognition Workshop 2014. In this section we provide some basic information on the challenge, present the experimental protocol and define the performance metric used to assess the recognition techniques.

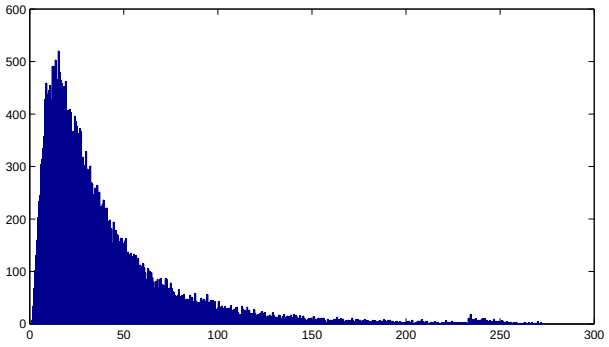

Figure 1: Histogram of recording durations. The histogram was computed from the durations corresponding to the i-vectors in the IVC development set.

4.1. Challenge description

The single task of IVC is that of speaker detection, i.e., to determine whether a specified speaker (the target speaker) is speaking during a given segment of conversational speech. The IVC data is given in the form of 600-dimensional i-vectors, divided into disjoint development and evaluation sets. The development set consists of 36,572 (unlabeled) i-vectors, while the evaluation set consists of 6,530 target i-vectors belonging to 1,306 target speakers (5 i-vectors per speaker) and 9,643 test i-vectors of a unknown number of speakers. Note that no explicit information is provided on whether the 1,306 speakers are distinct or not. Hence, it is possible that some of the target identities are duplicated.

The experimental protocol of IVC defines that a total of 12,582,004 experimental trials need to be conducted, where each trial consists of matching a single i-vector from the 9,643 test vectors against a given target model constructed based on the five target i-vectors belonging to the targeted speaker. It should be noted that — according to the rules [20] — the output produced for each trial must be based (in addition to the development data) solely on the training and test segment i-vectors provided for that particular trial, while the i-vectors provided for other trials may not be used in any way. The main characteristics of the experimental protocol are summarized in Table 1.

The durations of the speech segments used to compute the i-vectors for IVC are sampled from a log-normal distribution with a mean of 39.58 seconds (see Fig. 1, where a histogram of the duration from the development data is presented). This suggests that methods that take the uncertainty of the i-vectors due to duration variability into account should be effective in the challenge. However, since the only information provided with each i-vector is the duration of the speech recording used to compute the corresponding i-vector, techniques exploiting the posterior covariance, such as [5], [6], [7], are not feasible. Nevertheless, we expect that performance improvements should be possible by augmenting the information contained in the i-vectors with duration information in one way or another.

4.2. Performance metrics

In order to establish the performance of the given recognition technique, the file containing the scores for all trials needs to be uploaded to the IVC website. Each registered participant is allowed to upload up to 10 submission per day. The overall performance of the submitted techniques is measured in terms of

Table 1: Characteristics of IVC experimental protocol. The symbol *n/a* stands for the fact that the information is not available.

Data set	# i-vectors	#speakers	quality	# trials
development set	36,572	*n/a*	arbitrary	
evaluation set - target vectors	6,530	1,306	telephone speech	12,582,004
evaluation set - test vectors	9,643	*n/a*	arbitrary	

the minimal value of the decision cost function (DCF) obtained over all thresholds, where the DCF for a given threshold t is computed as:

$$\text{DCF}(t) = \frac{\text{\# misses}(t)}{\text{\# target trials}} + 100\frac{\text{\# false alarms}(t)}{\text{\# non-target trials}} \qquad (20)$$

Note that the minimial DCF value (minDCF) is the only performance metric returned by the on-line system and is, therefore, also the only metric reported in our experiments. When assessing the performance of a submitted recognition system only 40% of the trials are used, while the remaining 60% are withheld for calculating the official results at the end of the challenge. As a consequence, the final performance of our best performing system may differ in other reports on the 2014 i-vector challenge from what is reported here.

5. Experiments and results

5.1. Experimental setup

The experiments presented in the remainder are conducted in accordance with the experimental protocol defined for the i-vector challenge and presented in Section 4.1. The processing is done on a personal desktop computer using Matlab R2010b and the following open source toolboxes:

- the PhD toolbox [23], [24][5], which among others features implementations of popular dimensionality-reduction techniques;
- the Bosaris toolkit [25][6], which contains implementations of score calibration, fusion and classification techniques;
- the Liblinear library (with the Matlab interface) [26][7], which contains fast routines for training and deploying linear classifiers such as linear SVMs or logistic-regression classifiers.

All the experiments presented in the next sections can easily be reproduced using the above tools and functions.

5.2. Experiments with PCA

Our duration-dependent weighting scheme is based on the assumption that not all the available i-vectors are computed from speech recordings of the same length and are, therefore, not equally reliable. If the i-vectors are computed from recordings of comparable length, the weighting scheme would have only little effect on the given technique, as similar weights would be assigned to all the statistics and the impact of the weighting would basically be lost. On the other hand, if the i-vectors are computed from speech recordings of very different lengths, our weighting scheme is expected to provide more reliable results, as more reliable i-vectors are given larger weights when computing statistics for the given speaker-verification technique. Considering the histogram of durations presented in Fig. 1 we expect that our weighting scheme should provide some benefits in terms of performance.

Table 2: *Effect of the proposed weighting scheme on the baseline system defined for IVC. The Table shows minDCF values achieved by the baseline and weighted baseline systems as returned by the web-platform of the IVC as well as the relative change (in%) in the minDCF value, achieved with the weighting.*

Technique	Baseline	Weighted baseline	minDCF_{rel}
Score	0.386	0.372	3.63%

To assess our weighting scheme we first implement the baseline technique defined for the i-vector challenge and use the baseline performance for comparative purposes. Note that IVC defines a PCA-based system used together with cosine scoring as its baseline. Specifically, the baseline system consists of the following steps [20]

- estimation of the global mean and covariance based on the development data,
- centering and whitening of all i-vectors based on PCA (see Eq. 12),
- projecting all i-vectors onto the unit sphere (i.e., length normalization: $\mathbf{x} \leftarrow \frac{\mathbf{x}}{\sqrt{\mathbf{x}^T\mathbf{x}}}$),
- computing models by averaging the five target i-vectors of each speaker and normalizing the result to unit L_2 norm, and
- scoring by computing inner products between all models and test i-vectors.

In our first series of experiments, we modify the baseline system by replacing the PCA step (second bullet) with our duration-weighted version of the PCA. We provide the comparative results in terms of the minDCF values in Table 2. Here, the last column denotes the relative change in the minDCF value measured against the baseline:

$$\text{minDCF}_{rel} = \frac{\text{minDCF}_{base} - \text{minDCF}_{test}}{\text{minDCF}_{base}}, \qquad (21)$$

where minDCF_{base} stands for the minDCF value of the baseline system and minDCF_{test} stands for the minDCF value achieved by the currently assessed system.

Note that the proposed weighting scheme results in a relative improvement of 3.63% in the minDCF value over the baseline. This result suggests that a performance improvement is possible with the proposed weighting scheme, but a more detailed analysis of this results is still of interest. For this reason we examine the behavior of the baseline and weighted baseline

[5] http://luks.fe.uni-lj.si/sl/osebje/vitomir/face_tools/PhDface
[6] https://sites.google.com/site/bosaristoolkit/
[7] http://www.csie.ntu.edu.tw/ cjlin/liblinear/

Table 3: *Effect of excluding samples from the development set of the IVC data on the performance of the baseline and weighted baseline systems. The exclusion criterion is a threshold on the duration of the recording used to compute the i-vectors. The Table shows minDCF values as returned by the web-platform of the IVC.*

Exclusion criterion	$< 10s$	$< 15s$	$< 20s$	$< 25s$
Baseline	0.385	0.381	0.379	0.377
Weighted	0.372	0.371	0.371	0.371

techniques with respect to a smaller development set, where i-vectors computed from shorter recordings are excluded from the estimation of the global mean and covariance. Based on this strategy, we construct four distinct development sets with the first excluding all the i-vectors with the associated duration shorter than 10s, the second excluding all the i-vectors with the associated duration shorter than 15s, the third excluding all the i-vectors with the associated duration shorter than 20s, and the last excluding all i-vectors with the associated duration shorter than 25s. The baseline and weighted baseline technique are then trained on the described development sets. The results of this series of experiments are presented in Table 3.

Note that by excluding vectors from the development set, the baseline technique gradually improves in performance as more and more of the unreliable i-vectors are excluded from training. Continuing this procedure would clearly turn the trend around and the minDCF values would start getting worse, as too much information would be discarded. The weighted baseline system, on the other hand, ensures minDCF values comparable to those that were achieved when the entire development set was used for the training. This result again suggests that duration variability is addressed quite reasonably with the proposed weighting scheme.

5.3. Experiments with WCCN

In the next series of experiments we assess the performance of WCCN-based recognition systems. As a baseline WCCN system, we implement a similar processing pipeline as presented for the IVC baseline technique in the previous section, but add an additional step, which after whitening with PCA also whitens the within-class covariance matrix using WCCN. All the remaining steps of our WCCN-based baseline stay the same including length normalization, model construction and scoring. Whenever using the weighted version of WCCN we also use the weighted version of PCA in the experiments.

To further improve upon the baseline, we implement a second group of WCCN-based systems, where the cosine-based scoring procedure is replaced with a logistic-regression classifier and the length normalization is removed from the processing pipeline. With this approach all five target i-vectors of a given speaker are considered as positive examples of one class, while 5,000 i-vectors most similar to the given target speaker[8] are considered as negative examples of the second class. Based on this setup a binary classifier is trained for each target speaker, resulting in a total of 1,306 classifiers for the entire IVC data.

[8]Here, the similarity between the target vectors and the development vectors is measured by means of the IVC baseline system. Note that 5,000 negative examples are used to speed up experimentation. Our best results were achieved with the entire development set as counter-examples.

Before we turn our attention to the experimental results, it has to be noted that unlike PCA, which is an unsupervised technique, WCCN represents a supervised feature transformation techniques, which requires that all i-vectors comprising the development data are labeled. Unfortunately, the development data provided for the i-vector challenge is not labeled nor is the number of speakers present in the data known. To be able to apply supervised algorithms successfully we need to generate labels in an unsupervised manner by applying an appropriate clustering algorithm [27], [28]. Clustering will, however, never be perfect in practice, so the errors (utterances originated from the same speaker can be attributed to different clusters or utterances from different speakers can be attributed to the same cluster) are inevitable. Although there exists some evidence that labeling errors can degrade the recognition performance (seen as a bending of the DET curve), it is not completely obvious how sensitive different methods are with respect to those errors.

Since the selection of an appropriate clustering technique is (clearly) crucial for the performance of the supervised feature transformation techniques, we first run a series of preliminary experiments with respect to clustering and elaborate on our main findings. The basis for our experiments is whitened i-vectors processed with the (PCA-based) baseline IVC system. We experiment with different clustering techniques (i.e., k-means, hierarchical clustering, spectral clustering, mean-shift clustering, k-medoids and others), using different numbers of clusters and different (dis-)similarity measures (i.e., Euclidian distances and cosine similarity measures). The results of our preliminary experiments suggest the cosine similarity measure results in i-vector labels that ensure better verification performance than the labels generated by the Euclidian distance (with the same number of clusters). Despite the fact that several alternatives have been assessed, classical k-means clustering ensures the best results in our experiments and was, therefore, chosen as the clustering algorithm for all of our main experiments [9]. Based on our preliminary experiments, we select the k-means clustering algorithm with the cosine similarity measure for our experiments with WCCN and run it on the development data. We set the number of clusters to 4,000, which also ensured the best results during our preliminary experimentation.

The results of the WCCN-based series of experiments are presented in Table 4. Here, the relative change in the minDCF value is measured against the WCCN baseline. The first thing to notice is that with cosine scoring the WCCN-baseline systems (weighted and non-weighted) result in significantly worse minDCF values. However, when the scoring procedure is replaced with a logistic-regression classifier, this changes dramatically. In this situation, the WCCN-based system becomes highly competitive and in the case of the weighted system result in a minDCF value of 0.294. All in all, the weighting scheme seems to ensure a consistent improvement over the non-weighted case of around 3%. For the sake of completeness we need to emphasize that the best score we managed to achieve with a PCA-based system, when using a logistic-regression classifier was 0.326.

As a final remark, it needs to be stressed that the perfor-

[9]It is also worth noting, that the cluster labels generated with the k-means clustering algorithm were also used in conjunction with different PLDA-based models, i.e., the models presented in [16], [15] and [29], but different from WCCN no improvements over the baseline were achieved, regardless of the classifier used. This seems to suggest that feature transformation techniques, such as WCCN, are less susceptible to labeling errors than PLDA-models. However, more research would be needed to further validate this observation.

Table 4: *Effect of the proposed weighting scheme on our WCCN-baseline system. The Table shows minDCF values achieved by the baseline and weighted baseline WCCN systems as returned by the web-platform of the IVC as well as the relative change (in%) in the minDCF value, achieved with the weighting.*

Technique	Baseline	Weighted	minDCF_{rel}
Cosine	0.461	0.447	3.04%
Logistic	0.304	0.294	3.29%

mance of the logistic-regression classifier used in our experiments was extremely dependent on the right choice of parameters. Changing the parameters of the classifier only slightly resulted in minDCF values way above 0.3. To arrive at the results presented in Table 4 we needed to include a bias term and set the cost parameter to a relatively large value[10].

5.4. Comparative assessment

For the i-vector challenge we further tuned our best performing recognition system (i.e., the weighted version of our WCCN-system) to achieve even lower minDCF values. After implementing several additional steps we managed to reduce the minDCF value of our system to 0.280 by the time of writing. Specifically, the following improvements were implemented:

- duration was added as an additional feature to the i-vectors to construct 601 dimensional vectors before any processing,
- the clustering was improved by excluding clusters with a small fisher-score,
- the entire development set was used as negative examples when training the classifiers, and
- a second set of classifiers was trained on the test vectors and then used to classify the target vectors; the mean score over a given target speaker was then combined with the score computed based on the classifier trained on the target identity[11].

As indicated a couple of times throughout the paper, the best minDCF value we managed to achieve by the time of writing puts our system at third place in the i-vector challenge among the participating institutions. For the final ranking and performance scores the reader is referred to NIST's IVC web-site, where the IVC leader-board can be found: https://ivectorchallenge.nist.gov. However, it should be noted that after the Odyssey paper-submission deadline, we did not make any further improvements to our technique, while other participants probably did, so the ranking presented at the IVC web-site may differ to what is reported here.

6. Conclusions

We have presented a duration-based weighting scheme for feature transformation techniques used commonly in an i-vector based speaker-recognition system. We have applied the scheme on two established transformation techniques, namely, principal component analysis and within-class covariance normalization. We have assessed the duration-weighted techniques in the scope of the i-vector challenge organized by NIST within the Odyssey, Speaker and Language Recognition Workshop 2014 and achieved very competitive results. As part of our future work, we plan to evaluate the possibility of using a similar scheme with probabilistic linear discriminant analysis as well.

[10]The following LIBLINEAR settings needed to be used to produce the results reported in Table 4 for the logistic-regression classifier: '-s 0 -B 1 -c 100000'.

[11]Here, the role of the target and test vectors was simply flipped. Each test vector was used as a positive example of one class, while the development set was used for the negative samples. The target vectors were then classified based on the trained classifiers.

7. References

[1] A. Sarkar, D. Matrouf, P. Bousquet, and J. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification," in *Proceedings of Interspeech*, Portland, OR, USA, 2012. 1

[2] A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of Interspeech*, Florence, Italy, 2011, pp. 2341–2344. 1

[3] T. Hasan, S.O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J.H. Hansen, "Crss systems for 2012 nist speaker recognition evaluation," in *Proceedings of the IEEE Interational Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancoucer, Canada, 2013. 1

[4] M.I. Mandasari, M. McLaren, and D.A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensics application," in *Proceedings of Interspeech*, Florance, Italy, 2011, pp. 21–24. 1

[5] D. Garcia-Romero and A. McCree, "Subspace-constrained supervector PLDA for speaker verification," in *Proceedings of Interspeech*, Lyon, France, 2013. 1, 2, 4

[6] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proceedings of the IEEE Interational Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013. 1, 2, 4

[7] S. Cumani, O. Plchot, and P. Laface, "Probabilistic linear discriminant analysis of i-vector posterior distributions," in *Proceedings of the IEEE Interational Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013. 1, 2, 4

[8] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," *Speech Communication*, vol. 59, no. April, pp. 69–82, 2014. 1

[9] T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen, "Duration mismatch compensation for i-vector based speaker recognition systems," in *Proceedings of the IEEE Interational Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013. 1

[10] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P Dumouchel, "Text-dependent speaker recognition using plda with uncertainty propagation," in *Proceedings of Interspeech*, 2013. 1

[11] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000. 2

[12] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, tech. report crim-06/08-13," 2005, Available online. 2

[13] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques," in *Proceedings of Odyssey*, Brno, Czech Republic, 2010. 2

[14] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, 2007. 2

[15] P. Li, Y. Fu, U. Mohammed, J.H. Elder, and S. J.D. Prince, "Probablistic models for inference about identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, 2012. 2, 6

[16] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proceedings of Odyssey*, Brno, Czech Republic, 2010. 2, 6

[17] J. Villalba and N. Brummer, "Towards fully bayesian speaker recognition: Integrating out the between speaker covariance," in *Proceedings of Interspeech*, Florence, Italy, 2011. 2

[18] M. Senoussaoui, P. Kenny, N. Brummer, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender independent speaker recognition," in *Proceedings of Interspeech*, Florence, Italy, 2011. 2

[19] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proceedings of Interspeech*, Florence, Italy, 2011. 2

[20] NIST, "The 2013-2014 speaker recognition i-vector machine learning challenge," 2014, Available online. 2, 4, 5

[21] F. Mihelič V. Štruc and N. Pavešić, "Combining experts for improved face verification performance," in *Proceedings of the International Electrotechnical and Computer Science Conference (ERK)*, Portorož, Slovenia, 2008, pp. 233–236. 3

[22] A. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in *Proceedings of the IEEE Interational Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, 2006. 3

[23] V. Štruc and N. Pavešić, "The complete Gabor-Fisher classifier for robust face recognition," *EURASIP Advances in Signal Processing*, vol. 2010, pp. 26, 2010. 5

[24] V. Štruc, "The PhD face recognition toolbox: toolbox description and user manual," 2012, Available online. 5

[25] N. Brummer and E. de Villiers, "The BOSARIS toolkit user guide: Theory, algorithms and code for surviving hte new dcf," in *NIST SRE'11 Analysis Workshop*, Atlanta, USA, December 2011. 5

[26] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008. 5

[27] M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 22, no. 1, 2014. 6

[28] J. Žibert and F. Mihelič, "Fusion of acoustic and prosodic features for speaker clustering," in *Proceedings of the 12th Internation Conference on Text, Speech and Dialogue (TSD)*, V. Matoušek and P. Mautner, Eds., Pilse, Czech Republic, 2009, Lecture notes in computer science, pp. 210–217, Springer. 6

[29] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, "A scalable formulation of probabilistic linear discriminant analysis: Applied to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1788–1794, 2013. 6