Alp-ULj Speaker Recognition System for the NIST 2014 i-Vector Challenge

Boštjan Vesnicer,* Jerneja Žganec-Gros,* Simon Dobrišek† and Vitomir Štruc†

*Alpineon d.o.o.
Ulica Iga Grudna 15, SI-1000 Ljubljana
{bostjan.vesnicer,jerneja.gros}@alpineon.si

†Faculty of Electrical Engineering
University of Ljubljana
Tržaška cesta 25, SI-1000 Ljubljana
{simon.dobrisek,vitomir.struc}@fe.uni-lj.si

Abstract

I-vectors enable a fixed-size compact representation of speech signals of arbitrary durations. In recent years they have become the state-of-the-art representation of speech signals in text-independent speaker recognition. For practical reasons most systems assume that the i-vector estimates are highly reliable. However, this assumption is valid only in the case when i-vectors are extracted from recordings of sufficient length, but for short recordings the assumption does not hold any more. To address the problem of duration variability we propose a simple duration-based preprocessing weighting scheme that accounts for different reliability of i-vector estimates. We evaluate the proposed approach in the scope of NIST 2014 i-vector machine learning challenge, where we achieved competitive results.

Sistem za prepoznavo govorcev Alp-ULj s prireditve "NIST 2014 i-Vector Challenge"

I-vektorji omogočajo zgoščeno predstavitev govornih signalov poljubne dolžine v obliki vektorjev fiksne razsežnosti. V zadnjih letih so postali ena izmed najuspešnejših tehnologij na področju prepoznave govorcev. Zaradi praktičnih razlogov ponavadi predpostavimo, da je ocena i-vektorjev zelo zanesljiva. Ta predpostavka velja le v primeru, ko i-vektor ocenimo iz dovolj dolgega govornega posnetka, medtem ko je pri posnetkih krajše dolžine ta predpostavka v veliki meri kršena. V prispevku predlagamo posebno metodo predobdelave, v kateri na enostaven način upoštevamo dolžino posnetkov, iz katerih smo i-vektorje ocenili. Predlagano rešitev smo ovrednotili v okviru prireditve "NIST 2014 i-Vector Challenge", na kateri smo dosegli vzpodbudne rezultate.

1. Introduction

The area of speaker recognition has made significant progress over recent years. Today, recognition systems relying on so-called i-vectors, introduced in (Dehak et al., 2011), 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. (Sarkar et al., 2012; Kanagasundaram et al., 2011; Hasan et al., 2013a; Mandasari et al., 2011; Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013; Kanagasundaram et al., 2014; Hasan et al., 2013b; Stafylakis et al., 2013). The most recent solutions of the duration-variability problem, e.g. (Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013) 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

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.

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 NIST 2014 i-vector machine learning challenge (IVC). 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, a 27% relative improvement over the baseline system.

2. Prior work

Two of the most frequently used classification methods in i-vector-based speaker recognition are the cosine similarity (Dehak et al., 2010) and probabilistic linear discriminant analysis (PLDA), independently developed for face (Prince and Elder, 2007; Li et al., 2012) and speaker recognition (Kenny, 2010). Since its introduction, the PLDA model has been extended in different ways, e.g. the underlying Gaussian assumption have been relaxed (Kenny, 2010), the parameters of the model have been treated as random variables (Villalba and Brummer, 2011) and an extension to the mixture case has been proposed as well (Senoussaoui et al., 2011).

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 (Garcia-Romero and Espy-Wilson, 2011) 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

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 (NIST, 2014). 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 (Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013), are not applicable.

The basic step in computing the feature transform for most feature-extraction (or feature-transformation) techniques (e.g., PCA, WCCN, NAP, etc.) is the calculation of the sample mean and scatter (or covariance) matrix. 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 sample mean $\boldsymbol{m}$ and scatter matrix $\boldsymbol{S}$ can be calculated by the following formulas:

$$\boldsymbol{m} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \tag{1}$$

and

$$\boldsymbol{S} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{m})(\mathbf{x}_i - \boldsymbol{m})^{\mathrm{T}}. \tag{2}$$

The definition of the sample mean and scatter matrix in Eqs. (1) and (2) 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 multiply the contribution of each i-vector in Eqs. (1) and (2) by the weight which corresponds to the duration of the recording from which the vector was extracted. This modification gives the following formulas for the weighted mean $\boldsymbol{m}_{\boldsymbol{w}}$ and weighted scatter matrix $\boldsymbol{S}_{\boldsymbol{w}}$:

$$\boldsymbol{m}_{\boldsymbol{w}} = \frac{1}{T} \sum_{i=1}^{n} t_i \mathbf{x}_i \tag{3}$$

and

$$\boldsymbol{S}_{\boldsymbol{w}} = \frac{1}{T} \sum_{i=1}^{n} t_i (\mathbf{x}_i - \boldsymbol{m}_{\boldsymbol{w}})(\mathbf{x}_i - \boldsymbol{m}_{\boldsymbol{w}})^{\mathrm{T}}, \tag{4}$$

where $T = \sum_{i=1}^{n} t_i$.

Note that the presented weighting scheme reduces to the (non-weighted) standard version 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, PCA and WCCN. We first briefly describe the theoretical basis of both techniques and then show, how they can be modified based on the presented statistics.

3.1. 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 (V. Štruc and Pavešić, 2008). 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 $\mathbf{\Sigma}$ of the vectors in $\mathbf{X}$ into the following form:

$$\mathbf{\Sigma} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T, \tag{5}$$

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 $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_d, \}$ stands for a diagonal eigenvalue matrix with the eigenvalues arranged in decreasing order. Note that if $\mathbf{\Sigma}$ 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 $\mathbf{\Lambda}$. 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:

$$\mathbf{y} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}), \tag{6}$$

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:

$$\mathbf{y} = (\mathbf{U}\mathbf{\Lambda}^{-1/2})^T(\mathbf{x} - \boldsymbol{\mu}). \tag{7}$$

3.2. 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 (Hatch and Stolcke, 2006). WCCN can under certain conditions be shown to minimize the expected classification error 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 $C_1, C_2, \ldots, C_N$ 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:

$$\mathbf{\Sigma}_w^{-1} = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{L}$ and $\mathbf{L}^T$ stand for the lower and upper triangular matrices, respectively, and $\mathbf{\Sigma}_w^{-1}$ 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:

$$\mathbf{y} = \mathbf{L}^{\mathrm{T}}\mathbf{x}, \tag{9}$$

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

The weighted version of the WCCN transform can be obtained by replacing the standard withing-class scatter matrix with the weighted one.

4. The I-vector challenge

We evaluate the feasibility of the proposed duration-weighted scheme in the scope of IVC. 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.

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 (NIST, 2014) — 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 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. 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 (Garcia-Romero and McCree, 2013; Kenny et al., 2013; Cumani et al., 2013), 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.

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 (Štruc and Pavešić, 2010; Štruc, 2012)[1], which among others features implementations of popular dimensionality-reduction techniques;
- the Bosaris toolkit (Brummer and de Villiers, 2011)[2], which contains implementations of score calibration, fusion and classification techniques;
- the Liblinear library (with the Matlab interface) (Fan et al., 2008)[3], 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.

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 (NIST, 2014)

- estimation of the global mean and covariance based on the development data,
- centering and whitening of all i-vectors based on PCA (see Eq. 7),
- 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.

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

Table 1: *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%

Table 2: *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

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 1. Here, the last column denotes the relative change in the minDCF value measured against the baseline:

$$\text{minDCF}_{rel} = \frac{\text{minDCF}_{base} - \text{minDCF}_{test}}{\text{minDCF}_{base}}, \qquad (10)$$

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 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 2.

Note that by excluding vectors from the development set, the baseline technique gradually improves in perfor-

mance 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 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.

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 (Senoussaoui et al., 2014). 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. 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.

Table 3: *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%

The results of the WCCN-based series of experiments are presented in Table 3. 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.

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.

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 NIST i-vectormachine learning challenge 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.

7. References

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

S. Cumani, O. Plchot, and P. Laface. 2013. Probabilistic linear discriminant analysis of i-vector posterior distributions. In *Proc. ICASSP*, Vancouver, Canada.

N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny. 2010. Cosine similarity scoring without score normalization techniques. In *Proc. Odyssey*, Brno.

N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet. 2011. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 19(4):788–798.

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

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

D. Garcia-Romero and A. McCree. 2013. Subspace-constrained supervector PLDA for speaker verification. In *Proc. Interspeech*, Lyon, France.

T. Hasan, S.O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J.H. Hansen. 2013a. Crss systems for 2012 nist speaker recognition evaluation. In *Proc. ICASSP*, Vancoucer, Canada.

T. Hasan, R. Saeidi, J. H. L. Hansen, and D. A. van Leeuwen. 2013b. Duration mismatch compensation for i-vector based speaker recognition systems. In *Proc. ICASSP*.

A. Hatch and A. Stolcke. 2006. Generalized linear kernels for one-versus-all classification: application to speaker recognition. In *Proc. ICASSP*, Toulouse, France.

A. Kanagasundaram, R. Vogt, D. Dean, S. Sridharan, and M. Mason. 2011. I-vector based speaker recognition on short utterances. In *Proc. Interspeech*, pages 2341–2344, Florence, Italy.

A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez-Rodriguez, and D. Ramos. 2014. Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques. *Speech Communication*, 59(April):69–82.

P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel. 2013. PLDA for speaker verification with utterances of arbitrary duration. In *Proc. ICASSP*, Vancouver, Canada.

P. Kenny. 2010. Bayesian speaker verification with heavy-tailed priors. In *Proc. Odyssey*, Brno, Czech Republic.

P. Li, Y. Fu, U. Mohammed, J.H. Elder, and S. J.D. Prince. 2012. Probablistic models for inference about identity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):144–157.

M.I. Mandasari, M. McLaren, and D.A. van Leeuwen. 2011. Evaluation of i-vector speaker recognition systems for forensics application. In *Proc. Interspeech*, pages 21–24, Florance, Italy.

NIST. 2014. The 2013-2014 speaker recognition i-vector machine learning challenge. Available online.

S. J. D. Prince and J. H. Elder. 2007. Probabilistic linear discriminant analysis for inferences about identity. In *Proc. ICCV*, Rio de Janeiro, Brazil.

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

M. Senoussaoui, P. Kenny, N. Brummer, and P. Dumouchel. 2011. Mixture of PLDA models in i-vector space for gender independent speaker recognition. In *Proc. Interspeech*, Florence, Italy.

M. Senoussaoui, P. Kenny, T. Stafylakis, and P. Dumouchel. 2014. A study of the cosine distance-based mean shift for telephone speech diarization. *IEEE Transaction on Audio, Speech and Language Processing*, 22(1).

T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P Dumouchel. 2013. Text-dependent speaker recognition using plda with uncertainty propagation. In *Proc. Interspeech*.

F. Mihelič V. Štruc and N. Pavešić. 2008. Combining experts for improved face verification performance. In *Proceedings of the International Electrotechnical and Computer Science Conference (ERK)*, pages 233–236, Portorož, Slovenia.

J. Villalba and N. Brummer. 2011. Towards fully bayesian speaker recognition: Integrating out the between speaker covariance. In *Proc. Interspeech*, Florence, Italy.

V. Štruc and N. Pavešić. 2010. The complete Gabor-Fisher classifier for robust face recognition. *EURASIP Advances in Signal Processing*, 2010:26.

V. Štruc. 2012. The PhD face recognition toolbox: toolbox description and user manual. Available online.