Analysis of the Influence of Speech Corpora in the PLDA Verification in the Task of Speaker Recognition

Lukáš Machlica and Zbyněk Zajíc

University of West Bohemia in Pilsen,
Faculty of Applied Sciences, Department of Cybernetics,
Univerzitní 22, 306 14 Pilsen
machlica@kky.zcu.cz,zzajic@kky.zcu.cz
http://www.kky.zcu.cz/en

Abstract. In the paper recent methods used in the task of speaker recognition are presented. At first, the extraction of so called i-vectors from GMM based supervectors is discussed. These i-vectors are of low dimension and lie in a subspace denoted as Total Variability Space (TVS). The focus of the paper is put on Probabilistic Linear Discriminant Analysis (PLDA), which is used as a generative model in the TVS. The influence of development data is analyzed utilizing distinct speech corpora. It is shown that it is preferable to cluster available speech corpora to classes, train one PLDA model for each class and fuse the results at the end. Experiments are presented on NIST Speaker Recognition Evaluation (SRE) 2008 and NIST SRE 2010.

Key words: PLDA, latent space, fusion, supervector, FA, i-vector

1 Introduction

A major progress in the task of speaker recognition was done introducing supervector based techniques. Supervector is in fact a high dimensional feature vector obtained by the concatenation of lower dimensional vectors containing speaker dependent parameters – the most effective turned out to be the parameters related to Gaussian Mixture Models (GMMs) [1]. First attempts to incorporate supervectors into the speaker recognition task utilized Support Vector Machines (SVMs) along with distinct kernel functions [2]. Since GMMs belong to the class of generative models, whereas SVMs are based on the discrimination between classes, the techniques comprising both methods are also known as *hybrid modelling*. Subsequently, additional techniques were added to solve the problem of the change of operating conditions, namely Nuisance Attribute Projection (NAP) [3]. NAP is based on an orthogonal projection, where directions most vulnerable to environment changes are projected out. Since supervectors are of substantially high dimension (tens of thousands), which is often higher than the number of supervectors provided for the training, it is obvious that a

lot of dimensions will be correlated with each other, and that the information about the identity will be contained in a subspace of a much lower dimension. This idea was incorporated into the principle of Joint Factor Analysis (JFA) [4], where the word *joint* refers to the fact that not only the speaker, but also the channel variabilities are treated in one JFA model. However, since experiments in [5] showed that the channel space obtained by JFA does still contain some information about the speaker's identity, JFA was slightly adjusted giving rise to *identity vectors*, or *i-vectors* [6]. The main difference between JFA and i-vectors is that i-vectors do not distinguish between speaker and channel space. They work with a *total variability space* containing simultaneously speaker and channel variabilities, whereas JFA treats both spaces individually.

Parallel to JFA a very similar approach was introduced in the image recognition called Probabilistic Linear Discriminant Analysis (PLDA) [7]. The only difference from JFA is that in PLDA ordinary feature vectors are used instead of GMM based supervectors (for details on the treatment of supervectors in JFA see [4]). Since PLDA is a generative model, it allows to compute the probability that several i-vectors originate from the same source, and thus it is well suited as a verification tool for a speaker recognition system [8]. System presented in this paper will utilize i-vectors (described in Section 3) based on GMM supervectors (see Section 2) with a PLDA model (refer to Section 4) used in the verification phase.

The crucial problem when proposing a speaker verification system composed of modules (e.g. JFA, PLDA) described above is that data from a lot of speakers are required, moreover several sessions have to be available for each speaker in order to train a reliable i-vector extractor and a PLDA model. The problem faced in this paper will address the question whether distinct speech corpora (e.g. Switchboard 1, Switchboard 2, NIST SRE 2004, NIST SRE 2006, etc.) should be pooled together and used to train one PLDA model, or if each corpus should be used individually to train a separate PLDA model. In the latter scenario the results are fused at the end. Experiments can be found in Section 5.

2 Supervector Extraction based on GMMs

At first a Universal Background Model (UBM) has to be trained. UBM is in fact a Gaussian Mixture Model (GMM), however it is trained from a set containing a lot of speakers. The speakers data should match all the conditions, in which the recognition system is going to be used. UBM consists of a set of parameters $\boldsymbol{\lambda}_{\text{UBM}} = \{\omega_m, \boldsymbol{\mu}_m, \boldsymbol{C}_m\}_{m=1}^{M}$, where M is the number of Gaussians in the UBM, ω_m, $\boldsymbol{\mu}_m$, $\boldsymbol{C}_m$ are the weight, mean and covariance of the m^{th} Gaussian, respectively. Let $\boldsymbol{O}_s = \{\boldsymbol{o}_{st}\}_{t=1}^{T_s}$ be the set of T_s feature vectors $\boldsymbol{o}_{st}$ of dimension D belonging to the s^{th} speaker, and

$$\gamma_m(\boldsymbol{o}_{st}) = \frac{\omega_m \mathcal{N}(\boldsymbol{o}_{st}; \boldsymbol{\mu}_m, \boldsymbol{C}_m)}{\sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{o}_{st}; \boldsymbol{\mu}_m, \boldsymbol{C}_m)} \tag{1}$$

be the posterior probability of m^{th} Gaussian given a feature vector $\boldsymbol{o}_{st}$. And let $\boldsymbol{m}_0 = [\boldsymbol{\mu}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_m^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_M^{\mathrm{T}}]^{\mathrm{T}}$ be the supervector composed of UBM means. Then, for each speaker two supervectors are extracted

$$\begin{aligned}
\boldsymbol{n}_s &= \sum_{t=1}^{T_s} \Big([\gamma_1(\boldsymbol{o}_{st}), \ldots, \gamma_m(\boldsymbol{o}_{st}), \ldots, \gamma_M(\boldsymbol{o}_{st})]^{\mathrm{T}} \otimes \mathbf{1}_D \Big) \text{ of size } DM \times 1, \\
\boldsymbol{b}_s &= \sum_{t=1}^{T_s} \left[\gamma_1(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}^{\mathrm{T}}, \ldots, \gamma_m(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}^{\mathrm{T}}, \ldots, \gamma_M(\boldsymbol{o}_{st}) \boldsymbol{o}_{st}^{\mathrm{T}} \right]^{\mathrm{T}} \text{ of size } DM \times 1, \quad (2)
\end{aligned}$$

where $\otimes$ is the Kronecker product, and $\mathbf{1}_D$ is a D dimensional vector of ones. Note that $\boldsymbol{n}_s$ is the supervector containing "soft" counts of feature vectors aligned to Gaussians $1, \ldots, M$, and denoting $\boldsymbol{N}_s$ a diagonal matrix containing $\boldsymbol{n}_s$ on its diagonal, $\boldsymbol{m}_s = \boldsymbol{N}_s^{-1} \boldsymbol{b}_s$ is the new Maximum Likelihood (ML) estimate of supervector $\boldsymbol{m}_0$ given the dataset $\boldsymbol{O}_s$. At last note that the Maximum Aposteriory Probability (MAP) adaptation [9] of means of the UBM according to the given data set $\boldsymbol{O}_s$ expressed in the supervector notation is given as $\boldsymbol{m}_{\text{MAP}} = \tau \boldsymbol{m}_s + (1 - \tau) \boldsymbol{m}_0$ for some relevance factor τ.

3 i-Vector Extraction

The concept of the i-vectors extraction is based on Factor Analysis (FA) extended to handle session and speaker variabilities of supervectors to Joint Factor Analysis (JFA) [4]. Contrary to JFA, different sessions of the same speaker are considered to be produced by different speakers [5]. The generative i-vector model has the form

$$\boldsymbol{\psi}_s = \boldsymbol{m}_0 + \boldsymbol{T} \boldsymbol{w}_s + \boldsymbol{\epsilon}, \;\; \boldsymbol{w}_s \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}), \;\; \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \quad (3)$$

where $\boldsymbol{T}$ (of size $D \times D_w$) is called the total variability space matrix (it contains both the variabilities between speakers and the channel variabilities between distinct sessions of a speaker), $\boldsymbol{w}_s$ is the s^{th} speaker's i-vector of dimension D_w having standard Gaussian distribution, $\boldsymbol{m}_0$ is the mean vector of $\boldsymbol{\psi}_s$, however often the UBM's mean supervector is taken instead as a good approximation (therefore the same notation $\boldsymbol{m}_0$ is used), and $\boldsymbol{\epsilon}$ is some residual noise with a diagonal covariance $\boldsymbol{\Sigma}$ constructed from covariance matrices $\boldsymbol{C}_1, \ldots, \boldsymbol{C}_m$ of the UBM ordered on the diagonal of $\boldsymbol{\Sigma}$.

To train the matrix $\boldsymbol{T}$ two steps are iterated in a sequence. Given a training set of S couples of supervectors $\boldsymbol{b}_s, \boldsymbol{n}_s$, and the diagonal matrix $\boldsymbol{N}_s$ containing $\boldsymbol{n}_s$ on its diagonal, these steps are:

1. use previous estimate of $\boldsymbol{T}$ to extract new i-vectors for all speakers $1, \ldots, S$

$$\boldsymbol{w}_s = (\boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_s \boldsymbol{T})^{-1} \boldsymbol{T}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{b}}_s, \quad (4)$$

2. according to the new i-vectors compute block-wise a new estimate of $\boldsymbol{T}$

$$\boldsymbol{T}_m = \left(\sum_{s=1}^{S} \bar{\boldsymbol{b}}_{sm} \boldsymbol{w}_s^{\mathrm{T}}\right) \left(\sum_{s=1}^{S} \boldsymbol{N}_{sm} \left(\boldsymbol{w}_s \boldsymbol{w}_s^{\mathrm{T}} + \left(\boldsymbol{I} + \boldsymbol{T}^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \boldsymbol{N}_s \boldsymbol{T}\right)^{-1}\right)\right)^{-1}, \tag{5}$$

where $\bar{\boldsymbol{b}}_s = \boldsymbol{b}_s - \boldsymbol{N}_s \boldsymbol{m}_0$ is the centered version of $\boldsymbol{b}_s$ around the mean $\boldsymbol{m}_0$, and the index m in $\boldsymbol{T}_m$, $\bar{\boldsymbol{b}}_{sm}$, $\boldsymbol{n}_{sm}$ (and $\boldsymbol{N}_{sm}$) refers to blocks of $\boldsymbol{T}$, $\bar{\boldsymbol{b}}_s$, $\boldsymbol{n}_s$ (and thus to $\boldsymbol{N}_{sm}$) of sizes $D \times D_w$, $D \times 1$, $D \times 1$ so that $\boldsymbol{T}^{\mathrm{T}} = [\boldsymbol{T}_1^{\mathrm{T}}, \boldsymbol{T}_2^{\mathrm{T}}, \ldots, \boldsymbol{T}_M^{\mathrm{T}}]$, $\bar{\boldsymbol{b}}_s^{\mathrm{T}} = [\bar{\boldsymbol{b}}_{s1}^{\mathrm{T}}, \bar{\boldsymbol{b}}_{s2}^{\mathrm{T}}, \ldots, \bar{\boldsymbol{b}}_M^{\mathrm{T}}]$, $\boldsymbol{n}_s^{\mathrm{T}} = [\boldsymbol{n}_{s1}^{\mathrm{T}}, \boldsymbol{n}_{s2}^{\mathrm{T}}, \ldots, \boldsymbol{n}_M^{\mathrm{T}}]$, respectively. In fact, also $\boldsymbol{\Sigma}$ may be updated in each iteration, for details see [10].

4 Probabilistic Linear Discriminant Analysis (PLDA)

Let us assume that the i-vector extractor (4) was already trained, and that for each feature set $\boldsymbol{O}_s$ of a speaker s one i-vector $\boldsymbol{w}_s$ was extracted. Further, let us assume that several sessions $\{\boldsymbol{O}_{sh}\}_{h=1}^{H_s}$ of a speaker s are available, and that for each set of feature vectors $\boldsymbol{O}_{sh}$ of each session $h = 1, \ldots, H_s$ one i-vector $\boldsymbol{w}_{sh}$ was extracted. Since in the i-vector extraction phase no distinction between session space and speaker space were made a new model in the total variability space will be now described that is going to utilize also the session variabilities.

PLDA is a generative model of the form

$$\boldsymbol{w}_{sh} = \boldsymbol{m}_w + \boldsymbol{F}\boldsymbol{z}_s + \boldsymbol{G}\boldsymbol{r}_{sh} + \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}) \tag{6}$$

where $\boldsymbol{m}_w$ is the mean of $\boldsymbol{w}_{sh}$, columns of $\boldsymbol{F}$ span the speaker identity space, $\boldsymbol{z}_s$ of dimension D_z are coordinates in this space and they do not change across sessions of one speaker, columns of $\boldsymbol{G}$ span the channel space, $\boldsymbol{r}_{sh}$ of dimension D_r are the session dependent speaker factors, and $\boldsymbol{\epsilon}$ is some residual noise with diagonal covariance $\boldsymbol{S}$ and a zero mean. Further restrictions are put on distributions of latent variables $\boldsymbol{z}_s$ and $\boldsymbol{r}_{sh}$, namely that both follow standard Gaussian distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. Hence, $\boldsymbol{w}_{sh} \sim \mathcal{N}(\boldsymbol{m}_w, \boldsymbol{F}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{S})$. It is common and reasonable assumption that $D_z << D_w$ and that $D_z + D_r \approx D_w$. To train the model parameters $\boldsymbol{F}$, $\boldsymbol{G}$ and $\boldsymbol{S}$ one has to solve the system of equations [7]

$$\begin{bmatrix} \boldsymbol{w}_{s1} - \boldsymbol{m}_w \\ \boldsymbol{w}_{s2} - \boldsymbol{m}_w \\ \vdots \\ \boldsymbol{w}_{sH_s} - \boldsymbol{m}_w \end{bmatrix} = \begin{bmatrix} \boldsymbol{F} & \boldsymbol{G} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{F} & \boldsymbol{0} & \boldsymbol{G} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{F} & \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{G} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_s \\ \boldsymbol{r}_{s1} \\ \boldsymbol{r}_{s2} \\ \vdots \\ \boldsymbol{r}_{sH_s} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \\ \vdots \\ \boldsymbol{\epsilon}_{H_s} \end{bmatrix},\ \hat{\boldsymbol{\Sigma}}_{H_s} = \begin{bmatrix} \boldsymbol{S} & \boldsymbol{0} & \ldots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \ldots & \boldsymbol{S} \end{bmatrix}, \tag{7}$$

and when rewritten to a compact form we get

$$\hat{\boldsymbol{w}}_s = \boldsymbol{A}_{H_s} \hat{\boldsymbol{z}}_s + \hat{\boldsymbol{\epsilon}}, \tag{8}$$

where $\hat{\boldsymbol{\epsilon}} \sim \mathcal{N}(\boldsymbol{0}, \hat{\boldsymbol{\Sigma}}_{H_s})$. Matrices $\boldsymbol{A}_{H_s}$, $\hat{\boldsymbol{\Sigma}}_{H_s}$ depend on s through the number of their row- and column-blocks given by the number of sessions H_s of speaker s.

Problem (8) is a standard FA problem, for details on how to solve it see the appendix in [7].

4.1 Verification

In the verification phase two hypotheses are tested [7], namely

- hypotheses $\mathcal{H}_s$ that two i-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ share the same identity,
- hypotheses $\mathcal{H}_d$ that the identity of two i-vectors $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ differs.

The log-likelihood ratio is given as

$$\begin{aligned}\mathrm{LLR}(\boldsymbol{w}_1, \boldsymbol{w}_2) &= \log \frac{p(\boldsymbol{w}_1, \boldsymbol{w}_2|\mathcal{H}_s)}{p(\boldsymbol{w}_1|\mathcal{H}_d)p(\boldsymbol{w}_2|\mathcal{H}_d)} = \\ &= \log \mathcal{N}\left(\begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{m}_w\\ \boldsymbol{m}_w\end{bmatrix}, \begin{bmatrix}\boldsymbol{C}_w & \boldsymbol{C}_F\\ \boldsymbol{C}_F & \boldsymbol{C}_w\end{bmatrix}\right) - \log \mathcal{N}\left(\begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\end{bmatrix}; \begin{bmatrix}\boldsymbol{m}_w\\ \boldsymbol{m}_w\end{bmatrix}, \begin{bmatrix}\boldsymbol{C}_w & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{C}_w\end{bmatrix}\right),\end{aligned} \tag{9}$$

where $\boldsymbol{C}_w = \boldsymbol{F}\boldsymbol{F}^{\mathrm{T}} + \boldsymbol{G}\boldsymbol{G}^{\mathrm{T}} + \boldsymbol{S}$ and $\boldsymbol{C}_F = \boldsymbol{F}\boldsymbol{F}^{\mathrm{T}}$. Note that in this verification scenario we do not care about the form of the decomposition of $\boldsymbol{w}_1$ or $\boldsymbol{w}_2$ (latent variables $\boldsymbol{z}_s$, $\boldsymbol{r}_{sh}$ stay unknown). The question stated is whether two vectors share the same identity given the subspaces generated by $\boldsymbol{F}$ and $\boldsymbol{G}$.

5 Experiments

The question raised is whether all the available data from several distinct corpora should be pooled and used to train one PLDA model (thus find a decomposition of the total variability space based on all the available data), or if it would be more efficient to find a characteristic decomposition of the total variability space for each corpus individually (hence train several PLDA models), score each pair of vectors in relation to each total space decomposition, and finally fuse the obtained scores (we will use linear combination). We believe that the latter case makes the verification more robust since possible undesirable deviations in acoustic conditions of distinct corpora may become less evident. However, individual corpora still have to contain enough data to be able to train a reliable PLDA model.

5.1 Used Corpora

In order to be able to perform reliable tests we utilized corpora: NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard 1 Release 2 and Switchboard 2 Phase 3 for development purposes, and NIST SRE 2008, NIST SRE 2010 were used for calibration of Fusion Coefficients (FCs), and for the evaluation of generality of obtained FCs, respectively. We used only those speakers from development corpora who had more than 4 recorded sessions. Further, the development corpora were divided into 3 classes:

1. NIST040506 – containing 3787 recordings of 465 males of approximately 8 sessions for each male speaker,
2. SW1 – containing 2342 recordings of 211 males of approximately 11 sessions for each male speaker,
3. SW2 – containing 2183 recordings of 216 males of approximately 10 sessions for each male speaker,

Each of the recordings had approximately 5 minutes in duration including the silence. The division into the classes was made in relation to the similarity of corpora determined according to recording conditions given in the LDC Corpus Catalog[1].

In order to train the FCs "short2-short3 trials" from NIST SRE 2008 [11] were utilized, only telephone speech from males was used (648 target speakers and 1535 test speakers) yielding 16968 trials in total. To test the validity of learned FCs "core-core trials" from NIST SRE 2010 [12] were used, and again only telephone speech from males was used (1394 target speakers and 2474 test speakers) yielding 74762 trials in total. The duration of all the test and target recordings in both corpora was approximately 5 minutes including the silence.

5.2 Feature Extraction

The feature extraction was based on Linear Frequency Cepstral Coefficients (LFCCs), Hamming window of length 25 ms was used, the shift of the window was set to 10 ms. 25 triangular filter banks were spread linearly across the frequency spectrum, and 20 LFCCs were extracted, delta coefficients were added leading to a 40 dimensional feature vector. Also the Feature Warping (FW) normalization procedure was applied utilizing a sliding window of length 3 seconds. Right before the FW Voice Activity Detector (VAD), based on detection of energies in filter banks located in the frequency domain, was used in order to discard the non-speech frames. All the feature vectors were at the end down-sampled by a factor of 2.

The number of Gaussians in the UBM was set to 1024. The size of the total variability space matrix $\boldsymbol{T}$ in the i-vector extraction was set to $1024 * 40 \times 800$, thus the latent dimension (dimension of i-vectors) was $D_w = 800$. At last, the dimension of the speaker identity space in the PLDA model was set to $D_z = 100$ and the dimension of the session/channel space was set to $D_r = 800$, thus $\boldsymbol{F}$ was of size 800×100, and the channel matrix was a square matrix of size 800×800. The disproportion between dimensions of speaker and channel subspaces was adopted from [8].

5.3 Results and Analysis

UBM and the i-vector's extractor described in Section 3 were trained on the pooled dataset NIST040506 + SW1 + SW2. Next, three PLDA models were

[1] `http://www.ldc.upenn.edu/Catalog/index.jsp`

trained utilizing subsequently each of the 3 corpora classes. Trials from NIST SRE 2008 were then scored using all 3 PLDA models, and the scores were used to train the fusion coefficients via the linear logistic regression from the FoCal toolkit [13]. Finally, in order to test the validity of learned FCs the same approach was performed with trials from NIST SRE 2010, but the already learned FCs were used in the linear combination of obtained scores. Results are shown in Figure 1 and Table 1, also minimum of the Decision Cost Function (DCF) is reported. In order to compute the value of DCF the cost of missing a target was set to 10, the cost of the false alarm was set to 1, and the probability of seeing a true trial was set to 0.01. These values are adopted from the NIST Speaker Recognition Evaluation (SRE) 2008 [11].

We have trained one PLDA model also from polled corpora NIST040506 + SW1 + SW2 (this was not used in the fusion). Best results are obtained for the fused system in both NIST SRE 2008 and NIST SRE 2010. Note that PLDA trained only on SW2 outperforms all the other PLDA models trained on other corpora (even on the pooled corpora), but the fusion still increases the performance of the speaker verification system. However, in real conditions one can not rely only on one corpus (in this case it would be SW2) performing best on the development set.

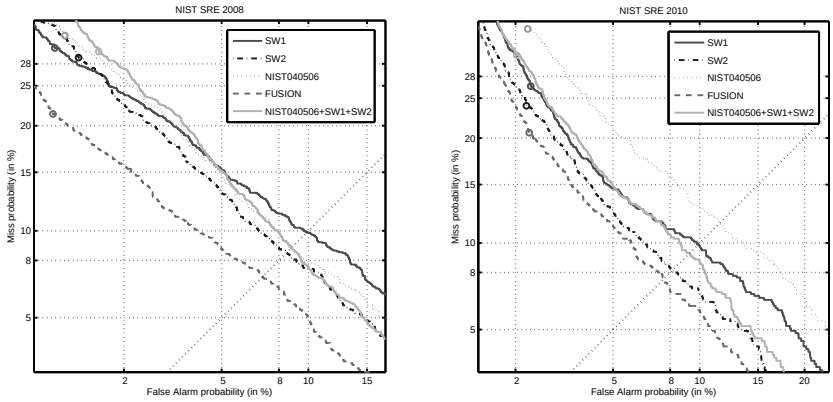

Fig. 1. DET curves for NIST SRE 2008 and NIST SRE 2010. Circles denote points where minDCF occurred.

Table 1. Results are given as EER [%] / minDCF. In the last column of the table also results for PLDA trained on pooled corpora NIST040506 + SW1 + SW2 is given.

	SW1	**SW2**	**NIST040506**	**FUSION**	**pooled**
NIST 2008	9.87/0.034	8.47/0.041	8.83/0.043	**7.06/0.031**	8.78/0.045
NIST 2010	9.89/0.050	8.11/0.046	11.58/0.057	**7.58/0.043**	9.26/0.051

6 Conclusion

Since often the verification conditions are unknown in advance (e.g. in the Speaker Recognition Evaluations (SREs) organized by NIST and other institutions) we cannot count on the use of one specific speech corpus performing best on the development set. It is more convenient to utilize several corpora. We have shown that if the utilized corpora have sufficient amount of data to train reliable PLDA models, it is preferable to train several PLDA models and fuse the results. The verification becomes more robust since the deviations in acoustic conditions of distinct corpora become less evident.

Acknowledgments. This research was supported by the grant of the University of West Bohemia, project No. SGS-2010-054.

References

1. Campbell, W., Sturim, D., Reynolds, D.: Support Vector Machines Using GMM Supervectors for Speaker Verification. In: IEEE Signal Processing Letters, vol. 13, pp. 308311 (2006)
2. Longworth, C., Gales, M.: Parametric and Derivative Kernels for Speaker Verification. In: Interspeech 2007 pp. 310313 (2007)
3. Solomonoff, A., Quillen, C., Campbell,W.: Channel compensation for SVM speaker recognition. In: Odyssey, pp. 5762 (2004)
4. Kenny, P.: Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Tech. report, Centre de Recherche Informatique de Montral (2006)
5. Dehak, N.: Discriminative and Generative Approaches for Long- and Short-term Speaker Characteristics Modeling: Application to Speaker Verification. Ph.D. thesis, École de Technologie Supérieure, Université du Québec (2009)
6. Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis For Speaker Verification. In: IEEE Transactions on Audio, Speech and Language Processing (2010)
7. Prince, S., Elder, J.: Probabilistic Linear Discriminant Analysis for Inferences About Identity. In: IEEE 11th International Conference on Computer Vision, pp. 18 (2007)
8. Matějka, P., Glembek, O., Castaldo, F., Alam, J., Plchot, O., Kenny, P., Burget, L., Černocký, J.: Full-covariance UBM and Heavy-tailed PLDA in I-Vector Speaker Verification. In: ICASSP 2011, pp. 48284831 (2011)
9. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. In: Digital Signal Processing, vol: 10, pp. 19-41 (2000)
10. Patrick, K., Pierre, O., Najim, D., Vishwa, G., Pierre, D.: A Study of Interspeaker Variability in Speaker Verification. In: IEEE Transactions on Audio, Speech and Language Processing, vol. 16, pp. 980-988 (2008)
11. The NIST Year 2008 Speaker Recognition Evaluation Plan, `http://www.itl.nist.gov/iad/mig/tests/spk/2008/sre08_evalplan_release4.pdf`
12. The NIST Year 2010 Speaker Recognition Evaluation Plan, `http://www.itl.nist.gov/iad/mig/tests/spk/2010/NIST_SRE10_evalplan.r6.pdf`
13. Brummer, N.: FoCal: Tools for fusion and calibration of automatic speaker detection systems (2006). http://sites.google.com/site/nikobrummer/focal