Supervised Mixture of PLDA Models for Cross-Channel Speaker Verification

Konstantin Simonchik, Timur Pekhovsky, Andrey Shulipa, Anton Afanasyev

1 Department of Speaker Verification and Identification, Speech Technology Center Ltd., St. Petersburg, Russia
{simonchik, tim, shulipa, afanasyev}@speechpro.com

Abstract
This paper presents a development of previous research by P. Kenny, which deals with using a supervised PLDA mixture of two gender-dependent speaker verification systems under the conditions of gender uncertainty. We propose using PLDA mixtures for speaker verification in different channels. However, in contrast to creating a gender-independent mixture, the optimal decision for training a channel-independent mixture for two channels in our task was mixing three channel-dependent PLDA systems. The experiments conducted on different conditions of NIST 2010 showed the superior robustness of the PDLA system mixture compared to each of its component PDLA subsystems not only in EER value but also in the stability of the decision threshold. The latter fact is very significant for using this approach not just for obtaining a good NIST SRE actual cost but also for commercial applications.

Index Terms: speaker verification, i-vector, length normalization, supervised mixture PLDA

1. Introduction
The results of recent 2010 NIST Speaker Recognition Evaluation (SRE) showed high efficiency of the several approaches based on the use of low-dimensional total-variability vectors (i-vectors) as feature-vectors proposed by Dehak [1]. Particularly, these approaches are effective for resolving the cross-channel problem dealing with differences between telephone channel speech and microphone interview speech that has strong reverberation. Among these methods, the techniques based on the Probabilistic Linear Discriminant Analysis approach (PLDA) [1, 2] are the most promising.

Initially designed for resolving cross-channel (microphone-telephone) problems, the heavy tailed PLDA (PLDA - HT) approach proposed by Kenny [2] suggests that the t-student distribution should describe more adequately the effects of non-Gaussian channels such as the gross channel distortions in microphone interviews. Unfortunately, the excellent efficiency of the PLDA - HT model observed for the homogeneous telephone channel was not observed for microphone speech.

In his next paper [3] Kenny found a better use of the PLDA - HT model for cross-channel verification tasks. The authors applied a heavy-tailed PLDA for the two-pass scheme, specifically the PLDA - HT model was used as the extractor and as the classifier. In fact, in using this scheme, the most important advantage the authors achieve is the effect of space reduction, where the original i-vector from the corresponding space (i.e., telephone-microphone) is transformed into some reducible i-vector of the extended space. The redundancy of the original i-vector components is eliminated, and also the channel effects are reduced. This particular scheme has led to the creation of a well-calibrated system of speaker recognition.

In this paper we refer to another remarkable work of Kenny [8]. In that paper a mixture of PLDA was applied to solve the problem of cross-gender verification. In [8] it was proposed to use supervised training mixtures PLDA that could be trained on both the ML-structure [4, 5] and the Bayesian-structure [6, 7], unlike the well-known classical unsupervised mixtures of Factor Analysis (FA). Unsupervised mixtures make it possible to carry out non-linear modeling of training data without requiring any prior knowledge about the segmentation data, and to perform mixing in the model space.

Unlike the unsupervised mixtures, in [8] training of individual PLDA systems is described. The individual PLDA systems correspond to two genders (components of a mixture). Thus the training of PLDA was performed on the segmented data for each gender. So the PLDA mixture is derived on the scoring stage by soft Bayesian fusioning of PLDA systems evidence. This PLDA mixture shows the superior efficiency, similar to the performance of individual PLDA-components (tested under the training conditions, i.e. one gender). We also assume that these PLDA mixtures demonstrate a good decision threshold stability (though it was not studied in [8]).

In this paper we apply unsupervised PLDA mixtures for the cross-channel verification task in order to deal with the problem of gross channel differences.

The implementation scheme for the mixtures of PLDA components trained on different channels is described in Section 2. Then Section 3 shows the experiments, performed on the NIST 2010. Section 4 presents a conclusion.

2. Supervised Learning of a Mixture of PLDA Models
In this section we show how to generalize a cross-gender mixture of PLDA models [8] to the case of cross-channel trials. Also we define the configuration of PLDA for the components of the mixture in this case.

2.1. PLDA Modeling
In this paper we consider the components of a mixture of PLDA, each of which consists of a single Gaussian analyzer of the speaker factor in the i-vector space.

The only dissimilarity from the original FA [5] is that the s-th training speaker has R(s) training sessions, which in turn is specific for the framework of PLDA-analyzer training. The formulas for updating the hyperparameters of the PLDA-analyzer in EM-algorithm can be defined in two ways: either...
using a direct exact solution for the stationary point where the only difficulty is the inversion of a large partitioned posteriori matrix of hidden variables; or, to avoid this difficulty, using the VBA-inference (Variational Bayesian Analysis) for posteriori hidden variables [2]. As P.Kenny in [2] we chose the latter option.

According to P.Kenny [2], the PLDA model for the F-dimensional input i-vector $w_r(s)$, which corresponds to the r-th session of the s-th speaker, is defined as:

$$w_r(s) = m_0 + Vy + Ux + \xi_r$$

(1)

where $y, x, \xi_r \sim N(0, \Lambda^{-1})$ are hidden speaker factors, channel factors and Gaussian noise, respectively. In this paper, we assume the Gaussian nature of the priors of these variables.

In (1) the model parameters are an $F \times 1$ mean vector $m_0$; a matrix $V$ of dimension $[F \times N_f]$, whose columns are referred to as eigenvoices; a matrix $U$ of dimension $[F \times N_f]$ whose columns are referred to as eigenchannels; and $[F \times F]$ noise precision matrix $\Lambda$, that equal to the inverse full-covariance matrix of the noise covariance matrix $\Sigma$. Henceforth, we will not consider the effects of channels ($U = 0$).

2.2. Channel Modeling

As mentioned in the introduction, according to Kenny, the mixtures of PLDA are implemented at the score level using a Bayesian fusion of scores from each component of the mixture.

Therefore, to build cross-channel models based on mixtures of PLDA, we have to make a generalization of the gender-independent framework [8] for the channel-independent framework. Thus, in the channel-independent PLDA framework [8] the logarithm of the likelihood-ratio is:

$$LLR = \ln \frac{P(y \mid m) \cdot P(w_1 \mid y) \cdot P(w_2 \mid y)}{P(w_1 \mid y, \theta_m) \cdot P(w_2 \mid y, \theta_m)} = \ln \frac{\sum_{m, m'} \sum \theta_m^*, \theta_{m'}^* P(w_1 \mid m, I) \cdot P(m \mid T) \cdot P(w_2 \mid m', I) \cdot P(m' \mid I)}{\sum_{m, m'} \sum \theta_m^*, \theta_{m'}^* P(w_1 \mid m, I) \cdot P(m \mid T) \cdot P(w_2 \mid m', I) \cdot P(m' \mid I)} = \ln \frac{\sum_{m, m'} \sum \theta_m^*, \theta_{m'}^* Q^{(m,m')}}{\sum_{m, m'} \sum \theta_m^*, \theta_{m'}^*},$$

(2)

where indices $m, m' \in \{1, \ldots, M\}$ run over all components of the mixture of PLDA, $M$ is the number of components in the mixture, and target- and imposter-priors are equal for all components:

$$P(m \mid T) = P(m \mid I) = 1 / M$$

$$Q^{(m,m')} = P(m \mid T) \cdot P(m' \mid I) = 1 / M^2.$$  

The most significant difference between the cross-channel and the cross-gender [8] estimates is the fact that now the same speaker in a target trial can be both in telephone and microphone channels, in contrast to the situation in [8] where, for example, males and females cannot be mixed in a target trial.

The numerator and the denominator of equation (2) contain the marginal likelihood ratio, calculated for the separate component of a mixture of PLDA:

$$P(w \mid m) = \int P(w \mid y, \theta_m) \cdot P(y) dy \quad (4)$$

under the assumption that we use only their eigenvoices ($U = 0$).

In this paper we consider a mixture of PLDA with 3 channel components:

$$M = 3, \{ \text{mic, phone, mic/phone} \}.$$  

Where the phone component corresponds to the NIST telephone data, the mic component corresponds to the NIST microphone-interview data, and the mic/phone component corresponds to the mixture of all the data.

3. Experiments and Discussion

3.1. Front-End Processing

In this study we used speech signal preprocessing, VAD-segmentation and MFCC extraction procedures from the GMM-SVM system, developed by Speech Technology Center, Ltd (STC) [9]. The front-end computes 13 mel-frequency cepstral coefficients, as well as the first and second derivatives, to yield a 39 dimensional vector per frame. The derivatives are estimated over a 5-frame context. To obtain these coefficients, speech samples are pre-emphasized, divided into 22ms window frames with a fixed shift of 11ms, and each frame is subsequently multiplied by a Hamming window function.

We also applied a cepstral mean subtraction (CMS) and did not apply Feature Warping [10] for the cepstral coefficients.

3.2. Universal Background Model

We used a gender-independent UBM with 512-component GMM, obtained by standard ML-training on the telephone part of the NIST’s SRE 1998-2008 datasets (all languages, all genders).

In our study we used 4329 training male and female speakers in total. We also used a diagonal, not a full-covariance GMM UBM.

3.3. i-vector extractor

In our cross-channel task we need to use the universal extractor of i-vectors, which performs adequately for both a telephone and a microphone channel. Here, unlike the cross-gender task [8], the big difficulty is the imbalance between the telephone and the microphone datasets. The size of the latter is several times smaller than that of the former.

In this case in [11] the universal extractor of the i-vector was suggested. The universal extractor must be suitable for accurate processing of both microphone and telephone speech. So this extractor is based on separate ML-estimations of two $T$ total variability-matrices. Mathematically, this can be expressed for
the speaker- and channel-dependent supervector $\mu$ with the following:

$$
\mu = \mu_0 + T' \cdot w' + T'' \cdot w''
$$

(5)

In our work, the telephone $T'$-matrix (of dimension 400) was trained on 11256 telephone recordings from the NIST 2002/2003/2004/2005/2006/2008 comprising 1250 male speakers’ voices (English only).

The microphone $T''$-matrix (of dimension 300) was trained on 4705 microphone-interview recordings from the NIST 2005/2006/2008 comprising 203 male speakers’ voices (English only).

This way we solve the problem of a gross imbalance between the phone and the microphone databases. After estimating $T''$ and $T'$, we concatenate them in order to get the $T$ matrix for the phone and the microphone databases. After estimating $T''$ and $T'$.

Discriminant Analysis (LDA) is implemented by combining these three models on the stage of obtaining the PLDA-scoring.

3.4. Channel-PLDA model training

Here we trained three PLDA models, each consisting of a single Gaussian analyzer. Namely, the Phone-PLDA model is trained on 11256 telephone recordings from NIST 2002/2003/2004/2005/2006/2008 comprising 1250 male speakers’ voices (English only). For the voice matrix $V$ the number of columns was chosen as $N_i=150$.

The Mic-PLDA model is trained on 4705 microphone-interviews from the NIST 2005/2006/2008 comprising 203 male speakers’ voices (English only). For the voice matrix $V$ the number of columns was chosen as $N_i=90$.

The problem of the strong imbalance between the phone and the microphone corpora arises again while training a channel-independent PLDA-analyzer ($\text{Chan}_{\text{-ind}}$-PLDA) on the aggregate dataset of the two systems described above.

We solved this problem by taking from the 11256 telephone recordings only 5000 recordings of those speakers who were represented in the microphone corpora, and adding to these all the recordings from the microphone corpora. For the voice matrix $V$ the number of columns was chosen as $N_i=120$.

As mentioned in [8], the mixture (mix-PLDA) model is implemented by combining these three models on the stage of obtaining the PLDA-scoring.

3.5. Channels Test

In this section, we report the evaluation results for the various verification systems. The results correspond to the evaluation tests performed on the male part of the NIST SRE 2010 corpora only: for the interview-interview task (det2), for the cross-channel task (det3), and for the telephone task (det5).

Table 1 leads us to conclude:

First, the best results are demonstrated by the mix-PLDA system: it produces the best EER in all tests and the best min-DCF in most tests. In the det5 telephone test, the mix-PLDA system is slightly inferior to the pure telephone system Phone-PLDA in the min-DCF value (0.458 compared to 0.452), but this difference can be produced by an error caused by the limited number of tests.

Second, in the det2 test, the mix-PLDA system shows a 10% reduction in EER compared to the pure Mic-PLDA system; in the det3 test, there is a 3% EER reduction compared to the channel-independent CI-PLDA system.

Third, the channel-independent CI-PLDA system works satisfactorily in all tests compared to the other pure systems. This system is less efficient than the Mic-PLDA and the Phone-PLDA systems only in the tests that correspond to their training conditions.

In other respects the results of the separate pure PLDA systems are as expected. For instance, in det2 the Mic-PLDA system works better than the Phone-PLDA system, while in det5 the Phone-PLDA system is better than Mic-PLDA.

We have also conducted a special experiment judging the robustness of all the systems under the unified condition $\text{Common det}=1,3,5$. It shows how stable the decision threshold is for each system. For example, the Phone-PLDA system demonstrates a strong variability of decision threshold: the common det=2,3,5 results are worse than the results for any separate condition. In contrast, the mix-PLDA system showed a
good calibration level comparable to that of the CI-PLDA system.
Thus, the test experiments on the male dataset of NIST SRE 2010 show the effectiveness of the proposed approach of supervised training for PLDA mixtures for the cross-channel speaker verification task.

4. Conclusions

In this paper, we proposed to use a mixture of PLDA models to solve the cross-channel problem for speaker verification. The tests on the male part of the NIST SRE 2010 dataset showed that the mixture of PLDA is more effective than verification systems trained on the data from individual channels or on mixed data.

We also showed that a mixture of PLDA models has good calibration.

5. References


