Effective Estimation of a Multi-Session Speaker Model using Information on Signal Parameters

Konstantin Simonchik¹, Andrey Shulipa¹, Timur Pekhovsky¹

¹ Department of Speaker Verification and Identification, Speech Technology Center Ltd., St. Petersburg, Russia

(simonchik,shulipa,tim)@speechpro.com

Abstract

The paper deals with the problem of estimation an optimal i-vector based speaker voice model using several sessions of his or her voice recordings, each of which has different signal parameters: speech duration and SNR. Our aim is to minimize inter-session variability so as to achieve minimal EER in the task of speaker recognition.

We examine the influence of the main signal parameters on inter-session variability and propose a model for multi-session i-vector estimation based on minimizing inter-session variability.

Index Terms: speaker recognition, multisession, i-vector.

1. Introduction

The results of the two latest NIST Speaker Recognition Evaluation (SRE) competitions demonstrate that currently the best art-systems for speaker recognition are systems based on the i-vector/PLDA-approach [1], [2], as well as those based on MFCC/JFA [3].

However, the condition of using a multi-session speaker model at the last NIST SRE 2012 competition uncovered a problem with using i-vector point estimation. The problem was that many sessions of the speaker enroll had very short recordings of his or her speech, and consequently very unreliable i-vectors are extracted from them, in contrast to long recordings. PLDA-estimation [2] makes it possible to work with a multi-session enroll in a natural way, but it regards all i-vectors of its sessions with the same accuracy, which was the reason for the degradation of PLDA systems at the competition.

The easiest way of building a multi-session speaker model for a speaker whose session durations are highly variable is combining the GMM-UBM model statistics calculated for each session. That is easy to implement both for building a multi-session JFA model [4] and for extracting an enroll i-vector from all the speaker’s sessions using a i-vector extractor. The drawback of this method can be indicated immediately: the longest session will be the one to contribute the most to the process of the i-vector extraction.

In this paper we propose an alternative approach to solving the problem posed at NIST SRE 2012. Like the authors of [5] we treat the i-vector as a random vector. However, in contrast to [5], we will directly calculate the posterior covariance matrix reflecting the uncertainty of the i-vector point estimation, instead of doing it within PLDA.

Moreover, such a direct approach will allow us to significantly widen the original task, so as to take into account not only utterance duration variation but also other signal characteristics as well, such as noise level. So in this paper we aim to examine the influence of not only pure speech durations but also SNR values on inter-session variability with respect to a voice model using i-vectors.

We propose a model that provides optimal estimation of an i-vector averaged over all sessions, which has minimal variability with respect to the measured signal parameters of each session and can further be efficiently used in PLDA estimation of verification.

2. Multi-Session Model

In this section we propose our representation of the model of inter-session noise of the i-vector, as well as a way of parameterizing it.

2.1. I-vector noise model

In this paper we are dealing with modeling a speaker’s voice by means of representing it in an i-vector space [6].

Let the speaker be represented by R sessions, with an i-vector built for each of them. We suppose that the i-vector of the k-th session \( \rho_k \) has a Gaussian distribution with the center at the point \( \overline{\rho} \) and a covariance matrix \( \Sigma_k \):

\[
\rho_k = \overline{\rho} + \xi_k,
\]

where \( \xi_k \in N(0, \Sigma_k) \) is the noise model that depends on different factors, such as speech duration T, SNR, reverberation, frequency response of the channel, etc, as well as the noise of the model itself. Since in this paper we are only dealing with verification in telephone channels, we assume that reverberation effects and variability in frequency response of communication channels do not influence inter-session variability of the i-vector very much, so we concentrate on speech duration and SNR only.

We make the assumption that these factors are independent and represent the noise model for the k-th session as a sum of components:

\[
\xi_k = \xi_k^{(T)} + \xi_k^{(SNR)} + \xi_k^{(N)},
\]
\[ \sum_{k} = \sum^{(T)} + \sum^{(SNR)} + \sum^{(N)}, \]  
(1)

where \( \xi_{k}^{(T)} \) is the noise component introduced by the duration \( T \) of the speech signal; \( \xi_{k}^{(SNR)} \) is the signal’s SNR; \( \xi_{k}^{(N)} \) is the residual noise component of the verification system; \( \sum^{(T)} \), \( \sum^{(SNR)} \) and \( \sum^{(N)} \) are the corresponding covariance matrices for each noise component.

This set of parameters is not accidental. These are the parameters that not only have a great influence on the quality of audible speech but can also be measured with a relatively high precision.

### 2.2. Multi-session i-vector model

We propose an estimation of a multi-session speaker i-vector \( \omega_{M} \) based on a linear weighted combination of i-vectors \( \omega_{k} \) that are calculated independently for each session \( k = 1 \ldots R \).

In this case, the weighted linear combination of i-vectors for each session is as follows:

\[ \omega_{M} = \sum_{i=1}^{R} \alpha_{k} \omega_{k}, \]  
(2)

where the coefficients of the linear combination satisfy the following condition:

\[ \sum_{i=1}^{R} \alpha_{k} = 1 \]

Since \( k \)-th session \( \omega_{k} \) has a Gaussian distribution,

\[ \omega_{k} = \bar{\omega} + \xi_{k}, \quad \xi_{k} \sim N(0, \Sigma_{k}), \]

then weighted linear combination of i-vectors \( \omega_{k} \) has a Gaussian distribution also

\[ \omega_{M} = \bar{\omega}_{M} + \xi_{M}, \quad \xi_{M} \sim N(0, \Sigma_{M}), \]

where expectation of a multi-session speaker i-vector \( \bar{\omega}_{M} \) and covariance \( \Sigma_{M} \) are defined as:

\[ \bar{\omega}_{M} = \bar{\omega} \]  
(3)

\[ \Sigma_{M} = \sum_{k=1}^{g} \alpha_{k}^{2} \Sigma_{k} \]  
(4)

We will consider the estimation of the multisession i-vector effective if it yields minimal inter-session variability. For the sake of simplicity, we will consider it to be the trace of the matrix \( \Sigma_{M} \). Then the coefficients \( \{ \alpha_{k} \}_{k=1..g} \) will be the solution of the following:

\[ Tr \left( \sum_{k=1}^{g} \alpha_{k}^{2} \Sigma_{k} \right) \rightarrow \text{min}; \]

\[ \sum_{k=1}^{g} \alpha_{k} = 1. \]  
(5)

It is easy to demonstrate that \( \{ \alpha_{k} \}_{k=1..g} \) can be found:

\[ \alpha_{k} = \frac{1}{Tr \Sigma_{k}} \left( \frac{\xi_{M}}{Tr \Sigma_{j}} \right) \]  
(6)

As evident from the obtained formulae, there exists an optimal solution for the combination of i-vectors that correspond to separate utterances of the speaker; this solution provides the minimal variability of the multi-session model i-vector.

### 2.3. Estimation of inter-session variability

As demonstrated in (6), in order to obtain an effective evaluation for the multi-session i-vector it is necessary to evaluate the inter-session covariance matrix trace for each session. The evaluation is supposed to be performed using the knowledge of the signal parameters: SNR, reverberation time and pure speech duration. However, in theory the task of calculating the influence of all these parameters on inter-session covariation of the i-vector is not solved, and it is probably not feasible due to complicated processing of the original signal, first of all at the stage of calculating speech features. For this reason, in this paper we propose to perform an experimental estimation of this dependence for a specific speaker recognition system, and then to parametrize the dependence of the inter-session covariance matrix trace in the following form:

\[ Tr \Sigma_{k} \approx f_{1}(SNR) + f_{2}(SNR) + Tr \Sigma^{(N)}, \]

where \( T \) is the duration of the speech signal, and \( f_{1}, f_{2} \) are approximating functions.

### 3. Experiments and Discussion

#### 3.1. Speaker recognition system

In our experiments we used the telephone i-vector based speaker recognition system developed by Speech Technology Center, Ltd (STC) for participation in the NIST SRE 2012 [7].

We used speech signal preprocessing, VAD-segmentation and MFCC extraction procedures described in [7]. 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 [8] for the cepstral coefficients.

We used a gender-independent UBM with 2048-component GMM, obtained by standard ML-training on the telephone part of the NIST’s SRE 1998-2010 datasets (English language, male gender).

In our study we used 1385 training speakers in total. We also used a diagonal, not a full-covariance GMM UBM.

The i-vector extractor was trained on 18403 telephone recordings from the NIST 1998-2010 comprising 1385 male speakers’ voices (English only). PLDA system was trained on the telephone data from the NIST 1998-2010, male gender only.
3.2. Experiments for estimating inter-session variability

The experiments used part of the development set of NIST SRE 2004 speech database. This database has a relatively high recording quality: SNR of about 40dB, speech duration over 100 seconds, no significant distortions in the data transmission channel. We selected \( S = 101 \) male speakers with 5 recording sessions each.

The first experiment consisted of varying speech signal duration and measuring inter-session variability. The SNR for each session was about 40dB, and reverberation time was less than 0.3 seconds. Thus, the influence of SNR and reverberation time factors on inter-session covariation was minimized. To simplify the experiment we considered their values to be 0. As the estimation of the average i-vector of each speaker for each time duration value we used the i-vector of the \( s \)-th speaker averaged over his longest sessions:

\[
\bar{\omega}_s = \frac{1}{R} \sum_{r=1}^{R} \omega_{r,s}.
\]

The inter-session covariance matrix was estimated as:

\[
\Sigma^{(T,N)} = \frac{1}{SR} \sum_{s=1}^{S} \sum_{k=1}^{R} (\omega_{r,s}^{(T)} - \bar{\omega}_s)(\omega_{r,s}^{(T)} - \bar{\omega}_s)^T.
\]

It is important to note that in this experiment inter-session covariance matrix is equal to the sum of covariance matrices for the factors of speech duration \( T \) and the residual noise component introduced by the verification system.

\[
\Sigma^{(T,N)} = \Sigma^{(T)} + \Sigma^{(N)}.
\]

The observed dependence graph \( Tr\Sigma^{(T,N)} \) is given in Figure 1.

\[
\text{Figure 1. Dependence of the matrix trace of inter-session variability of the i-vector on pure speech duration.}
\]

Its values are well approximated by the following fractional function:

\[
Tr\Sigma^{(T,N)} = \frac{\beta_1}{T + \beta_2} + Tr\Sigma^{(N)},
\]

where \( T \) is speech duration in seconds. For the speaker recognition system under discussion, the parameter values are calculated as follows: \( \beta_1 = 7.355; \beta_2 = 8.151 \); the value of the residual noise component \( Tr\Sigma^{(N)} = 0.3403 \).

To sum up, in the first experiment we established a dependence of inter-session variability on speech duration.

The second experiment involved varying the SNR value of the speech signal and measuring inter-session variability. Speech duration for each session was not less than 60 seconds, and reverberation time was less than 0.3 seconds. This means that the influence of speech duration and reverberation time was minimized. Noisy sessions were modeled by adding different types of real noises that were selected randomly: pink noise, white noise, street noise, office noise, etc (8 noise types in total).

We used the averaged speaker i-vector from the first experiment as an estimate for the average i-vector of each speaker given a constant pure speech duration. We calculated the covariance matrix using the following formula:

\[
\Sigma^{(SNR,N)} = \frac{1}{SR} \sum_{s=1}^{S} \sum_{k=1}^{R} (\omega_{r,s}^{(SNR)} - \bar{\omega}_s)(\omega_{r,s}^{(SNR)} - \bar{\omega}_s)^T.
\]

In the second experiment it was assumed that the inter-session covariance matrix was equal to the sum of covariance matrices for SNR and the residual noise component introduced by the verification system:

\[
\Sigma^{(SNR,N)} = \Sigma^{(SNR)} + \Sigma^{(N)}.
\]

The observed dependence graph \( Tr\Sigma^{(SNR,N)} \) and the approximation result are given in Figure 2. As demonstrated in the experiment, the values \( Tr\Sigma^{(SNR,N)} \) are well approximated by the following function:

\[
Tr\Sigma^{(SNR,N)} = \beta_3 \tanh(\beta_4 \cdot SNR + \beta_5) + \beta_6 + Tr\Sigma^{(N)},
\]

Where the parameter values for the speaker recognition system under discussion, are calculated as follows: \( \beta_3 = 0.4323; \beta_4 = -0.06946; \beta_5 = 0.3751; \beta_6 = 0.5079 \), the value of the trace of the covariance matrix of residual variability \( Tr\Sigma^{(N)} \) was taken from the first experiment.

\[
\text{Figure 2. Dependence of the matrix trace of inter-session variability of the i-vector on SNR.}
\]

To sum up, in the second experiment we established a dependence of inter-session variability on SNR. The total dependence of inter-session variability on the two factors (speech duration and SNR) for the system under discussion can be represented as follows:
\[ T \Sigma = \frac{\beta_1}{T + \beta_2} + \beta_3 \tanh(\beta_4 \cdot SNR + \beta_5) + \beta_6 + T \Sigma^{(N)}, \]

where the coefficients \( \beta_1, \beta_2, ... \beta_6 \) were found above.

### 3.3. Experiments using a speech database with variable signal parameters

For the experiments we used the speech database described in 3.2. The goal of the experiments was to compare different approaches to combining model speaker sessions into one model which would yield minimal verification error. Verification was performed using two approaches: the PLDA system described in 3.1, and the same i-vector system but with a cos kernel [9].

In the first experiment we examined only the influence of the SNR parameter on building a multi-session model. To build a speaker model, we took the speaker’s first three sessions, each 60 seconds long and with random SNR values from 0 to 40dB. The SNR distribution in the model sessions was uniform. As test sessions we took the 2 remaining sessions, each with 25 seconds of speech and a fixed SNR of 20dB.

The results of the first test are presented in Table 1.

Table 1. Experimental results for model sessions with constant 60 second speech duration and SNR varying between 0 and 40dB. Duration of test sessions 25 seconds, SNR 20dB.

<table>
<thead>
<tr>
<th>Method of multisession model estimation</th>
<th>EER, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLDA</td>
<td></td>
</tr>
<tr>
<td>Cos-kernel</td>
<td></td>
</tr>
<tr>
<td>Combination of GMM statistics</td>
<td>1.39</td>
</tr>
<tr>
<td>i-vector averaging</td>
<td>1.18</td>
</tr>
<tr>
<td>Proposed method</td>
<td>0.96</td>
</tr>
</tbody>
</table>

In the second experiment both SNR values and durations of model sessions were varied. The SNR values were generated randomly from 0 to 40dB and speech duration was 3 to 60 seconds.

Table 2. Experimental results for model sessions with speech duration varying from 3 to 60 seconds and SNR varying between 0 to 40dB. Duration of test sessions 25 seconds, SNR 20dB.

<table>
<thead>
<tr>
<th>Method of multisession model estimation</th>
<th>EER, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLDA</td>
<td></td>
</tr>
<tr>
<td>Cos-kernel</td>
<td></td>
</tr>
<tr>
<td>Combination of GMM statistics</td>
<td>4.15</td>
</tr>
<tr>
<td>i-vector averaging</td>
<td>4.19</td>
</tr>
<tr>
<td>Proposed method</td>
<td>3.83</td>
</tr>
</tbody>
</table>

As shown in Tables 1 and 2, the proposed method of effective estimation of a multi-session speaker model using linear weighting is superior in quality to methods using combination of statistics or i-vector averaging both for the PLDA verification method and for the cos kernel method.

The experiments were performed on the database which was used to estimate the approximation parameters of inter-session variability.

In order to test the efficiency of our approach in the case when the development database is different from the test database, we performed an additional experiment. The test database and the protocol of comparisons were taken from NIST SRE 2012.

Since the variance of SNR and speech duration is very small for the NIST SRE 2012 database, to test our method we added noise to the files used for building multisession speaker models to achieve SNR from 0 to 40 dB; pure speech duration in the files was varied from 3 to 60 seconds.

This way it was possible to artificially increase the range of intersession variation, since the original recordings were all approximately the same with respect to speech quality and duration. As in the previous case, testing was performed on recordings of male speakers in a telephone channel. The experimental results are given in Table 3. We performed experiments with the original NIST SRE 2012 but did not find any advantage of the proposed method over the averaging method. Our explanation is that in the original NIST SRE 2012, the method works in the tails of the curves in Figures 1, 2.

To summarize, experiments with the NIST SRE 2004, 2012 speech database demonstrate the effectiveness of the proposed approach to multi-session model estimation using several speaker utterances.

Table 3. Experimental results for NIST SRE 2012. Pure speech duration on the model recordings varied from 3 to 60 seconds, SNR was 0-40 dB.

<table>
<thead>
<tr>
<th>Method of multisession model estimation</th>
<th>EER, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLDA</td>
<td></td>
</tr>
<tr>
<td>Combination of GMM statistics</td>
<td>11.89</td>
</tr>
<tr>
<td>i-vector averaging</td>
<td>11.54</td>
</tr>
<tr>
<td>Proposed method</td>
<td>11.20</td>
</tr>
</tbody>
</table>

### 4. Conclusions

In this paper we proposed a new approach to estimating a multi-session speaker i-vector for PLDA verification systems, taking into account not only the duration variation between model sessions but also the SNR variation between these sessions.

We demonstrated the advantage of this approach over the methods of session statistics combination and simple averaging of i-vectors in building a multi-session model.

The proposed method makes it possible to easily address any other signal characteristics left unexamined in this paper.

In the future we are planning to develop our approach by adding estimates of other signal characteristics (such as reverberation time) and taking into account the frequency response of the transmission channel.

### 5. References


1608