Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification

A. K. Sarkar, D. Matrouf, P. M. Bousquet and J. F. Bonastre

Universite D’Avignon, LIA, Avignon, France
sarkar.achintya@gmail.com,
{driss.matrouf,pierre-michel.bousquet,jean-francois.bonastre}@univ-avignon.fr

Abstract

It is well known that state-of-the-art speaker verification systems using the i-vector concept perform well when target speakers training and test utterances have the same condition: long-long as per NIST evaluation. In practice, real-life applications impose strong constraints on the amount of data that can be used in training target and test speaker models. Since speaker verification systems based on the i-vector approach need to estimate some statistical parameters, the aim of this paper is to explore methods to train statistical parameters of the classical i-vector system when target speakers are trained and tested on mismatched data durations. Experimental results are shown on NIST 2008 SRE for various durations of target training and test speech segments ranging from long to very short, such as full (average 2.5 minutes), 5 seconds and 10 seconds.

Index Terms: Short segment, i-vector, Length Normalization, PLDA, Speaker Verification

1. INTRODUCTION

The recently introduced i-vector concept by Dehak et al [1] has become the state-of-the-art technique in Speaker Verification (SV). In this concept, speech segments are characterized by vectors which are obtained by projecting them onto a total variability space, \( T \). These vectors are called i-vectors. The total variability space is trained by pooling a large number of speech utterances from many speakers over various channels/sessions. During test, the i-vector of a given utterance is scored against the claimant representative i-vector. Before scoring, i-vectors are conditioned according to the scoring technique. Most commonly, Probabilistic Linear Discriminant Analysis (PLDA) [2, 3, 4] is used in the i-vector domain. PLDA is like Joint Factor Analysis (JFA) [5] but in the i-vector space, and uses single Gaussian instead of Gaussian Mixture Models (GMMs) [6].

Recent studies in [7, 8] have shown that length normalization of the i-vectors before PLDA modeling plays a significant role in speaker verification performance. An interesting performance is shown in [7], where the length of the i-vectors is iteratively normalized. In the iterative process, i-vectors are normalized by using the covariance matrix, \( W \), which is calculated by pooling the i-vectors of the training data, the same as used for training the \( T \) space. The block diagram of a speaker verification system using length normalization with PLDA is illustrated in Fig.1.

It is well established that the i-vector approach shows very good speaker verification performance in NIST protocol, where speech segments are generally long and comparable in duration during training and test. Approximately, each speech file

is 5 minutes long and has an average of 2.5 minutes of active speech. In practice, real-life applications impose strong constraints [9] on the amount of data that can be used in target speakers’ training and test phases. Several conditions can exist based on the amount of target enrollment and test data:

- Both target speakers training and testing are using long utterances (i.e. long speech segments)
- Both target speakers training and testing are using short utterances
- Target speakers are trained with short speech segments and tested on long utterances
- Target speakers are trained with long speech segments and tested on short utterances

The last condition is the most common scenario in real-time applications [9]. For example, in applications of speaker verification systems on mobile phones, users may give several speech samples during enrollment phase. However, in test phase, data is often limited to short test utterances/segments (e.g. 5 seconds).

Now arises the following question: what is the optimal choice of speech segment duration for training \( T \), \( W \) and PLDA parameters when target speakers are enrolled and tested in either matched or mismatched speech segment durations? Two speech segments with different durations but coming from the same speaker in the same session will lead to two different i-vectors. Can we see this difference as a disturbing noise? If yes, we can assume that the shorter segment’s i-vector is a noisy version of the longer one. Generally, in pattern recognition, the best performance is obtained by training the statistical parameters of a model in the same conditions as in which it will be used. Based on this, we can expect that training the PLDA (say) parameters using long segments and use them with short segment i-vectors in test will drastically reduce the performance. Therefore, the main motivation of this paper is to investigate the effect of training statistical model parameters of state-of-the-art speaker verification systems on different durations of target speaker training and testing speech segments.

We also show an ad-hoc fusion technique, where a number of i-vectors for a given speech segment are generated with
respect to different $T$ spaces which are built using different duration of speech segments. Similarly in test, the i-vectors of the test data are scored against the claimant representative i-vectors over different ranks in PLDA. Then, scores are summed and converted into a single value. This method shows significant improvement in system performance, especially when both target training and test phases use short speech segments.

The outline of this paper is as follows: Section 2 describes the concept of i-vector technique. Section 3 describes the post-processing method on i-vectors before PLDA modeling. Section 4 presents the PLDA technique. Experimental setup is described in Section 5. Results and discussion are described in Section 6. Finally, the paper is concluded in Section 7.

2. i-vector concept

Assume that all cepstral vectors are generated by a GMM called the Universal Background Model (UBM) [10]. All speaker models (GMM) are, in fact, derived from the UBM. The concatenation of a given GMM means leads to very high dimension vector, called super-vector. The main idea behind the i-vector concept [1] is that there is subspace $T$ (columns are the basis of subspace) with low dimension (with respect to super-vector space), where the speaker and the channel variabilities are concentrated. The i-vectors are obtained by projecting the speech segments into that $T$ subspace. This can be expressed as follows:

$$M_{seg(CF \times 1)} = m_{CF \times 1} + T_{CF \times R}[w_{R \times 1}]$$

where $w$ is called i-vector. $M$ and $M_{seg}$ are the super-vectors of the speaker independent UBM and speaker adapted model, respectively. $C$ and $F$ are, respectively, the number of mixture in UBM and dimension of the feature vectors.

During training, target speakers are characterized by their i-vectors using his/her speech data as Eq.(1). In test, the i-vector of the test utterance is scored against claimant speaker specific i-vector. Before scoring, i-vectors are conditioned to compensate the session variability. Conditioning techniques are described in next section.

3. Post-processing of i-vector

Recently, several study in [7, 8] showed that i-vector conditioning is important before PLDA modeling. We consider two methods for i-vector conditioning. In first technique [8], i-vectors are normalized to zero mean, unity variance and then methods for i-vector conditioning. In first technique [8], i-vector conditioning is important before PLDA modeling. We consider two methods for i-vector conditioning. In first technique [8], i-vectors are normalized to zero mean, unity variance and then.

$$\hat{w} \leftarrow \frac{W^{-\frac{1}{2}}(w - \bar{w})}{\sqrt{(w - \bar{w}) W^{-1}(w - \bar{w})}}$$

where $W$ and $\bar{w}$ are the within class covariance matrix and mean vector of the training i-vectors respectively in successive iteration. $W$ is estimated using development data set (see section 5). We call it Spherical normalization and denoted by Sph.

In this paper, we show the system performance with this algorithm using two iterations. For more details see [7].

4. Probabilistic Linear Discriminant Analysis (PLDA)

PLDA is a generative modeling technique. It uses JFA framework to decompose i-vector $w$ into four components, as follows:

$$w = \mu + \phi_{[p \times r_s]}y_{[r_r \times 1]} + \Gamma_{[p \times r_c]}z_{[r_c \times 1]} + \epsilon$$

where $\phi$ and $\Gamma$ are rectangular matrices and represents the eigen voice and eigen channel subspace respectively. $y_s$ and $z$ are the speaker factor and channel factor, respectively. Standard normal priors are assumed for $y_s$ and $z$. $\epsilon$ indicates the residual noise.

During testing, the score between two i-vectors is calculated like follows,

$$score(w_1, w_2) = \log \frac{p(w_1, w_2|\theta_{tar})}{p(w_1, w_2|\theta_{non})}$$

where hypothesis $\theta_{tar}$, states that $w_1$ and $w_2$ are from the same speaker and hypothesis $\theta_{non}$ defines that they are from different speakers. Details about the PLDA concept, training of meta-parameter ($\phi$, $\epsilon$), hypothesis $\theta_{tar}$ and $\theta_{non}$ can be found in [2, 3, 4, 7].

5. Experiment Setup

Experiments are performed on NIST 2008 SRE core condition (male speakers and Det 7 condition) as per NIST evaluation plan [11].

50 dimensional Linear Frequency Cepstral Coefficient (LFCC) feature vectors (19 static, 19 $\Delta$, 11 $\Delta\Delta$ and $\Delta$ energy) are extracted from speech signal at frame rate 10 ms with 20 ms Hamming window over frequency band 300-3400 Hz. Then Voice Activity Detection (VAD) is used to remove the less energized/silence frame from the feature vectors. Finally, silence-removed feature vectors are normalized to zero mean and unity variance normalization at utterance level.

For short segment experiments, desired number of energized frames are selected from the long speech utterance after VAD. The feature vectors are then normalized to zero mean and unity variance.

The GMM UBM of 512 mixture components is trained using data from NIST 2004 SRE.

For i-vector extraction, total variability space, $T$, is trained using 12399 speech utterances from 890 speakers (NIST 2004-05, Switchboard II part 1, 2 & 3; Switchboard cellular part 1 & 2, about 15 sessions per speaker). This data set is also used for implementing spherical normalization algorithm and PLDA. In PLDA, channel rank is kept full i.e. equal to the dimension of the i-vector and speaker rank is varied.

The system performs are evaluated in terms of Equal Error Rate (EER) and Minimum Detection Cost Function (MinDCF) as per NIST evaluation plan [11].

6. Results and Discussion

Table 1 shows speaker verification performance of the i-vector based system for different conditions (i.e. duration of speech segments) of $T$, $W$, PLDA training and evaluation data. full, 5s and 5/10s +full terms used in the Tables are described as follows:

- full indicates the whole speech signal which is available in a particular utterance. There are an average 2.5 minutes of active speech per utterance in this case.
5s represents 5 seconds (i.e. 500 energized frames).

5/10s +full indicates cases where both the full utterance as well as truncated speech segments (e.g. 5/10s) are used.

It is important to note that truncated speech segment is generated by taking the desired number of energized frames from a particular utterance and then normalized to zero mean and unit variance. The performance of the respective systems are shown for their optimal value of PLDA rank. For simplicity, PLDA rank is not shown in Tables. Std+ and Sph denote respectively, the i-vector post-processing techniques standardization and spherical normalization.

We can first look at Table 1 (a) which shows the speaker verification performance when target speakers are trained and tested with truncated short speech segments (e.g. 5s/10s). From Table 1 (a), it is observed that training of T, W and φ with short+full (i.e. 5s/10s +full) speech segments shows better performance. Hence, it indicates that it is better to train T, W and φ with both short+full speech segments instead of only full speech utterances, for evaluation of speaker verification systems on short segments. One of the reasons can be that short segment/utterances characterizes the speakers on a space which is far away from the space of full data. Therefore, only full utterances are not able to model the space which is spanned by short segments. On the other hand, since short speech segments can be thought of as noise, they do not contain enough speaker information. Therefore, there can be less variability among the speakers to discriminate them. In case of full segments, it contains enough speaker information. The combination of short+full utterances covers both noise and speaker information, hence showing better performance.

Now if we look at Table 1 (b) which shows speaker verification performance when target speakers are trained using full data and tested on truncated short segments, it is observed that system Std+PLDA yields the best performance when T, W and φ with short+full (i.e. 5s/10s +full) speech segments. In case of system Sph+PLDA gives the best performance when T, W and φ are trained using full utterances. The different performance in the observations of the two systems could be due to the fact that they are different in i-vector conditioning techniques before PLDA modeling. This indicates that when target speech training and testing are respectively, on full and short, it is better to train T, W and φ with full in case of Std+PLDA system and short+full for Sph+PLDA system. However, system performance is marginally different when T, W and φ are trained using full speech segments, when compared to short+full.

Table 1 (c) shows the best system performance which is achieved when T, φ, target training and testing are done on full utterances. It can be due to the reason that large amount of training and testing data contains more speaker related information. In comparison the information in short segments is irrelevant.

We will now present the speaker verification systems performance using the proposed ad-hoc fusion method. In our ad-hoc fusion system, during training, i-vectors are estimated for a given speech segment with respect to different T and φ spaces which are trained using short, full and short+full, respectively. Similarly, in test phase, i-vectors of the test speech segment are scored against the claimant representative i-vectors (obtained during training). This develops several sub-systems depending upon the training data duration used for T and φ spaces. Each sub-system has its own W for i-vector normalization. The scores for the different PLDA projected i-vectors over the sub-systems (based on T spaces) are then fused. The value of the PLDA rank (speaker factor) is varied from 50 to 400 with increments of 50. Motivation behind the above fusion strategy: we observe that sometime fusion of different systems for their best PLDA rank does not give complementary information (not shown in this paper) for the other and hence it does not provide gain in the fused system performance. Therefore, we consider fusing the scores of different projected PLDA i-vectors over a number of systems. It is important to note that i-vectors are [400 x 1] dimensional in our experiments. During fusion, equal weightage is given to all scores.

Tables 2 (a)-(c) compare the performance of speaker verification systems for different conditions of evaluation with or without applying the proposed ad-hoc fusion technique. The fusion system performance is compared with the best system performance obtained in Tables 1 (a)-(c).

In case of full training and test condition of target, we achieved the best performance when T, W and φ were also trained with full utterances/speech segments. Hence, the fusion results are only shown on full training condition of T, W and φ.

From Tables 2 (a)-(c), it is observed that the proposed fusion strategy significantly improves the system performance substantially when both target training and test are on short segments compared to the respective system without fusion. It indicates that different T and PLDA ranks of i-vectors contain

---

**Table 1:** Comparison of EER and MinDCF of i-vector based speaker verification system for estimation of statistical parameters on different training and test conditions on NIST 2008 SRE core condition (male speakers and Det 7).

(a) truncated training and test condition

<table>
<thead>
<tr>
<th>Target train-test</th>
<th>Train (T, W, φ)</th>
<th>% EER (MinDCF)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stdn+PLDA</td>
<td>Sph+PLDA</td>
</tr>
<tr>
<td>5s - 5s</td>
<td>18.75 (0.0765)</td>
<td>18.79 (0.0774)</td>
</tr>
<tr>
<td>5s + full</td>
<td>21.63 (0.0909)</td>
<td>19.81 (0.0882)</td>
</tr>
<tr>
<td>5s + full</td>
<td>18.34 (0.0760)</td>
<td>17.81 (0.0757)</td>
</tr>
</tbody>
</table>

(b) full training and truncated test condition

<table>
<thead>
<tr>
<th>Target train-test</th>
<th>Train (T, W, φ)</th>
<th>% EER (MinDCF)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stdn+PLDA</td>
<td>Sph+PLDA</td>
</tr>
<tr>
<td>full - 5s</td>
<td>17.08 (0.0637)</td>
<td>15.72 (0.0623)</td>
</tr>
<tr>
<td>full</td>
<td>11.01 (0.0521)</td>
<td>9.34 (0.0459)</td>
</tr>
<tr>
<td>full + full</td>
<td>10.26 (0.0439)</td>
<td>10.24 (0.0445)</td>
</tr>
</tbody>
</table>

(c) full training and full test condition

<table>
<thead>
<tr>
<th>Target train-test</th>
<th>Train (T, W, φ)</th>
<th>% EER (MinDCF)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Stdn+PLDA</td>
<td>Sph+PLDA</td>
</tr>
<tr>
<td>full-full</td>
<td>2.50 (0.0155)</td>
<td>2.28 (0.0163)</td>
</tr>
</tbody>
</table>

Stdn=standardization, Sph=spherical normalization
From the observations in Tables 1 and 2, we can summarize the following comments:

- When a speaker verification system is evaluated on short duration of speech segments/ utterances (say, 5/10s) (in Table 1 (a)), it is better to train \( T \), \( W \) and \( \phi \) with short+full speech segments even though target speakers are trained using full (i.e large amount of data).
- If target speakers are enrolled with full speech segments and tested on either short or full (in Tables 1 (a) & (b)), it is better to train \( T \), \( W \) and \( \phi \) long speech segments.
- The proposed fusion strategy significantly improves the system performance, especially for target speakers trained and tested on short segments (in Table 2). This is more useful for real-time applications of speaker verification systems in limited data conditions.

### 7. Conclusion

In this paper, we have studied the effect of training statistical parameters of the state-of-the-art i-vector based speaker verification system when target training and testing data are mismatched in durations. These mismatched durations of the data are commonly faced by all speaker verification systems in real applications due to limited access of data from users. It is observed that short speech segments are optimal for training system model parameters when target speakers are enrolled and tested on short segments. In the case of target speakers training on long speech segments and tested on either full (i.e. long) or short speech segments, it is appropriate to use long speech segments for training system model parameters. Finally, we have also shown an ad-hoc fusion system which significantly improves the performance of speaker verification systems, specially on short evaluation data. In this approach, i-vectors are estimated with respect to different total variability spaces for a given speech segment. Total variability spaces are trained using short, full and short+full speech segments. Then, scores of the claimant and test i-vectors are fused over different PLDA ranks.

### 8. References


