Effect of multicondition training on i-vector PLDA configurations for speaker recognition

Padmanabhan Rajan, Tomi Kinnunen, Ville Hautamäki

School of Computing, University of Eastern Finland
paddy@cs.uef.fi, tkinnu@cs.uef.fi, villeh@cs.uef.fi

Abstract

The i-vector representation and PLDA classifier have shown state-of-the-art performance for speaker recognition systems. The availability of more than one enrollment utterance for a speaker allows a variety of configurations which can be used to enhance robustness to noise. The well-known technique of multicondition training can be utilized at different stages of the system, including enrollment and classifier training. We also study the effect of mismatched training, averaging and length normalization. Our study indicates that multicondition training of the PLDA model, and if possible the enrollment i-vectors are the most important to achieve good performance in noisy evaluation data.

Index Terms: Speaker verification, i-vector, PLDA, multicondition training

1. Introduction

The i-vector representation followed by probabilistic linear discriminant analysis (i-vector PLDA framework) has become state-of-the-art in speaker verification systems over the past few years. There has been active research in improving the basic i-vector-based system proposed in [1]. When speech utterances are represented as i-vectors, the speaker verification problem is simply to determine if the i-vectors have the same speaker information or not. In a typical speaker verification trial, there are two i-vectors. One i-vector represents the enrollment utterance of a given speaker, and the other represents a test utterance. If the verification system determines that the speaker information in the two i-vectors are the same, then both the utterances are deemed to come from the given speaker.

In the recently concluded NIST Speaker Recognition Evaluation (SRE) 2012, a given target speaker had *multiple* utterances (hence multiple i-vectors) available for enrollment. This has led to several interesting possibilities of how to utilize these multiple i-vectors effectively. In this paper, we study various ways of utilizing the multiple enrollment i-vectors. The motivation is to determine the effect of various system level configurations on the verification accuracy. We also experiment with utilizing the enrollment i-vectors by emphasizing the amount of speech present in them.

Recent studies have looked at various configurations to the i-vector PLDA framework. These include studies on utterance length, including the effect of short utterances [2], and mismatched or variable utterance duration [3, 4]. The effect of noise on modern verification systems has been reported in [5]. Studies on the effect of using multiple speech sources, including multicondition training have been studied in [6, 7]. An interesting study with respect to the SRE 2012 evaluation and the population used in PLDA training has been reported in [8]. Most of these studies have looked at the effect of variations beginning with the estimation of i-vector hyperparameters (i. e. the i-vector extractor.) On the other hand, in this paper, our focus is mostly on the enrollment stage and PLDA training; the i-vector hyperparameters are unchanged.

We systematically study the various configurations possible when utilizing multiple enrollment i-vectors. These include:

1. The use of original versus noisy training utterances
2. The effect of applying multicondition training
3. Averaging scores or scoring using averaged i-vectors
4. The effect of i-vector length normalization
5. The effect of taking weighted average of enrollment i-vectors

In particular, we address the above questions using a corpus consisting of millions of test trials with simulated additive noise [9]. In addition to state-of-the-art PLDA classifier, a high-performance voice activity detector, (described in [10]), is adopted for the experiments.

2. I-vector PLDA system

In this section, we give an overview of the i-vector PLDA system utilized for the studies in this paper.

2.1. The i-vector representation

The i-vector representation [1] is a fixed-length representation of speech utterances, which usually consist of variable number of feature vectors. In this representation, the samples of the utterance are transformed in successive steps into a vector in $\mathbb{R}^D$, where D is the i-vector dimension. Cepstral coefficients extracted from the speech utterance are represented in terms of zero- and first-order Baum-Welch statistics, with respect to a universal background model (UBM.) The supervector representing these statistics are projected into a D-dimensional *total variability space* (the matrix of basis vectors for this space is the i-vector extractor.) Compared to the supervector, the total variability space is of much lower dimension (typically between 400 to 600.) The UBM and the i-vector extractor are estimated from appropriate training corpora. Methods to estimate the i-vector extractor and the i-vectors are presented in [11, 12].

2.2. PLDA model

Probabilistic linear discriminant analysis (PLDA) has been applied successfully to specify a generative model of the i-vector representation [13, 14]. A speaker- and recording-specific i-vector $\mathbf{w}_{s,r}$ can be represented as

This work was supported by Academy of Finland (proj. 253120)

$$\mathbf{w}_{s,r} = \mathbf{m} + \mathbf{S}\mathbf{x}_s + \mathbf{C}\mathbf{y}_{s,r} + \epsilon_{s,r} \quad (1)$$

where the i-vector represents the rth recording of the sth speaker. Here, $\mathbf{m} + \mathbf{S}\mathbf{x}_s$ is the speaker-dependent part, and $\mathbf{C}\mathbf{y}_{s,r} + \epsilon_{s,r}$ is a session dependent part. $\mathbf{m}$ is a global offset, $\mathbf{S}$ is a set of basis vectors for the speaker subspace, representing the *between-speaker* variability, $\mathbf{C}$ represents the channel subspace, representing the *within-speaker* variability and ϵ represents the remaining residual variability. The latent variables $\mathbf{x}_s$ and $\mathbf{y}_{s,r}$ are assumed to have standard normal distributions, and ϵ is assumed to follow a Gaussian distribution with zero mean and diagonal covariance.

Given two i-vectors $\mathbf{w}_1$ and $\mathbf{w}_t$, the PLDA framework forms the verification score s by determining the likelihood ratio of them having the same or different $\mathbf{x}_s$ in equation 1 [13]

$$s = \frac{p(\mathbf{w}_1, \mathbf{w}_t|H_1)}{p(\mathbf{w}_1|H_0)p(\mathbf{w}_t|H_0)} \quad (2)$$

where the hypothesis H_1 indicates that both i-vectors come from the same speaker, and H_0 indicates they come from different speakers.

3. Experimental configuration

Practical speaker verification systems have to work in diverse conditions, including noisy and reverberant environments. Our experiments will focus on noisy conditions. Performance is reported in terms of equal error rate (EER) and MinDCF, with cost parameters (see [15]) $C_{\text{miss}} = 10$, $C_{\text{fa}} = 1$ and $P_{\text{tgt}} = 0.01$.

3.1. I4U corpus

As part of the pre-evaluation activity for the NIST SRE 2012, the I4U consortium[1] developed a dataset based on previous years' NIST corpora. The presence of noisy test data encourages the use of *multicondition training* [7, 16]. The I4U dataset is thus designed with multicondition training in mind. With the objective of performing system fusion and calibration, the dataset consisted of a development part (I4U DevSet) and an evaluation part (I4U EvalSet). Since our studies do not involve system fusion, our experiments are done on the EvalSet portion. In the EvalSet, the train data and the test data are drawn from the SRE 2006, 2008 and 2010 corpora. The data had multiple channels, including telephone, microphone and interview data, as determined from the keys released by NIST. In addition to the utterances used as-such from these corpora (henceforth termed *original utterances*), noisy versions of each utterance were generated using FaNT[2]. For each train utterance, two noisy versions at 6 dB and 15 dB signal-to-noise ratio (SNR) were generated using HVAC (heating, ventilation and air-conditioning) and crowd noise. The number of enrollment utterances for target speakers varied from 3 to 108, with an average of 19 per speaker. Statistics about the I4U EvalSet are provided in Table 1. More details about the I4U dataset is provided in another paper submitted to Interspeech 2013 [9].

3.2. i-vector PLDA system description

The i-vector PLDA system used for our studies uses a standard Mel frequency cepstral coefficient (MFCC) front-end with 30 ms frame size and 15 ms shift. The MFCCs were obtained using a 27-channel mel-frequency filterbank followed by RASTA filtering, adding delta and double deltas, frame dropping using VAD and utterance level cepstral mean and variance normalization (CMVN), in this order. The 1024-mixture UBM was trained with data from NIST 2004, 2005, 2006 and 2008 SRE, whereas the i-vector extractor from NIST 2004, 2005, 2006, Fisher and Switchboard data. The i-vector dimension was 600, and gender-dependent hyperparameters were built. In our experiments, as in [14], we assume that the residual term in the PLDA formulation (Equation 1) has full covariance and hence omit the channel subspace. The PLDA model consisted of 200 dimensions for the speaker subspace.

[1]The I4U consortium consists of nine universities and research institutes.

[2]FaNT - Filtering and Noise Adding Tool. Available: http://dnt.kr.hsnr.de/download.html

Table 1: Statistics about data in the I4U EvalSet, used for the experiments in this paper.

	Male	Female
Num. train spk.	763	1155
Num. test spk.	804	1102
Num. enroll seg.	29961	43119
Num. test seg.	21837	28548
Num. tgt. trials	15483	20763
Num. non-tgt. trials	16646148	32952177

4. Studies on different configurations

The focus of our experiments is the on effect of using multiple enrollment utterances (which contain both original and noisy versions) on the i-vector PLDA framework. Another possibility of using noisy data is for training the UBM and the i-vector extractor. However, training these hyperparameters is time consuming, and hence is not pursued in this study. Our experiments focus on the enrollment phase and the PLDA training, which are computationally inexpensive.

4.1. The use of original versus noisy data

Matched environment conditions between train and test data usually result in improved classifier performance, when compared to the mismatched case. In realistic scenarios, we cannot assume prior information about the type of noise which will be encountered. Classical speaker verification systems, such as the GMM-UBM system, show much degradation in performance when there is mismatch. So it is interesting to see the effect of clean versus noisy training, in the i-vector PLDA framework.

The results of speaker verification experiments using various cases of original and noisy data (for the female case) is tabulated in Table 2. The first column gives the data used for target speaker enrollment and PLDA training. In this set of experiments, multiple enrollment i-vectors are averaged into a single i-vector. Verification performance is reported for each SNR in the test data.

From Table 2, it can be seen that matched environment conditions do not necessarily give the best performance for the i-vector PLDA system. From columns 1 and 2, we see that even for noisy enrollment data, the (relatively) cleaner original test data gives good verification accuracy. Similar observations were made in [5]. But unlike in [5], observing the first row of Table 2, we cannot conclude that if at least one of the utterances in a trial is clean, performance will improve. The best performance is obtained when all three environments (as shown in the last row) is used for enrollment and PLDA training. This leads

Table 2: Effect of using original or noisy training utterances for enrollment and PLDA training. Only female case is shown. SNR-matched case is shown shaded. Performance is in terms of EER (MinDCF.)

Enroll and PLDA	Test data		
	Orig.	15 dB	6 dB
Orig. only	0.67 (0.26)	1.40 (0.59)	4.13 (1.79)
15 dB only	0.85 (0.38)	1.09 (0.43)	2.32 (0.90)
6 dB only	1.34 (0.68)	1.30 (0.52)	1.87 (0.73)
Orig.+15dB+6dB	0.73 (0.21)	0.92 (0.25)	1.41 (0.51)

us to the effect of multicondition training.

4.2. Multicondition training

For multicondition training, we have access to multiple versions of the training data, which reflects possible distortions expected during evaluation. For the I4U dataset, each enroll utterance also has two noisy versions, at 6dB SNR and 15 dB SNR. Following [7], *pooled multicondition training* is done to estimate the PLDA hyperparameters. Thus, this model assumes that all of the N enrollment i-vectors $\mathbf{w}_{s,r}$ $1 \leq r \leq N$ for a given speaker are generated by the same hyperparameters in Equation 1.

The availability of multicondition training during enrollment and/or PLDA training gives the four possibilities tabulated in Table 3.

Table 3: Effect of multicondition training on enrollment, PLDA training, or both. 'MC' stands for multicondition training. The performance is in terms of EER (MinDCF.)

Enroll MC	PLDA MC	Male	Female
No	No	2.25 (0.87)	2.23 (0.92)
Yes	No	1.58 (0.62)	2.06 (0.83)
No	Yes	1.10 (0.37)	1.26 (0.53)
Yes	Yes	1.06 (0.34)	1.3 (0.50)

As expected, multicondition training improves verification performance. Multicondition during both enrollment and PLDA training gives the best performance, although the difference between the rows 3 and 4 is not much. Hence, for the rest of the experiments, the configuration in row 4 is used.

5. Averaging and length normalization

When multiple i-vectors $\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_N$ are available for enrollment, and $\mathbf{w}_t$ is the test i-vector, the scoring function can be expressed as follows:

$$s = \frac{p(\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_N, \mathbf{w}_t | H_1)}{p(\mathbf{w}_1, \mathbf{w}_2, \ldots \mathbf{w}_N | H_0) p(\mathbf{w}_t | H_0)} \tag{3}$$

Although the PLDA model specified in [13] can directly score multiple enrollment i-vectors as above, it is simpler to score the averaged i-vector,

$$\mathbf{w}_{\text{avg}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{w}_i \tag{4}$$

where N is the number of enrollment i-vectors for a given target speaker. Another way to utilize the multiple enrollment i-vectors is to score each of them individually, and average the scores,

$$\text{scr} = \frac{1}{N} \sum_{i=1}^{N} \text{score}(\mathbf{w}_i, \mathbf{w}_{\text{test}}) \tag{5}$$

where score() is the PLDA scoring function.

Length normalization [17] is usually applied to i-vectors to make them more Gaussian. It has been shown that this helps in reducing mismatch between train and test i-vectors and results in increased recognition accuracy without using the more complicated heavy-tailed PLDA models [17, 18]. With averaging i-vectors or scores, length normalization can be applied before or after the averaging.

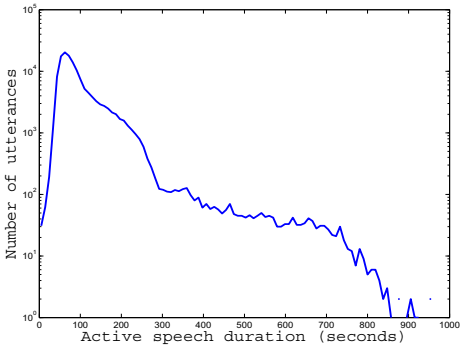

Figure 1: Histogram (in log scale) of the amount of active frames in all utterances of the I4U dataset.

Table 4: Effect of averaging and length normalization (LN) in terms of EER (MinDCF.)

Averaging	LN	Male	Female
No length normalization			
Avg. i-vec	None	1.58 (0.58)	2.08 (0.75)
Avg. score	None	1.71 (0.69)	2.22 (0.88)
With length normalization			
Avg. i-vec	Before	1.06 (0.34)	1.30 (0.50)
Avg. i-vec	After	1.11 (0.43)	1.43 (0.71)
Avg. score	Before	1.25 (0.48)	1.52 (0.66)

The effect of applying length normalization, i-vector averaging or score averaging is tabulated in Table 4. As can be seen, length normalization improves performance considerably. Length normalizing the enrollment i-vectors and then averaging them into a single enrollment i-vector seems to give the best result.

6. Using weighted averages

Different enroll utterances typically have different lengths, and the averaged enroll i-vector $\mathbf{w}_{\text{avg}}$ in Equation 4 may not accurately emphasize longer utterances (presumably the ones with more speech content.) The histogram (in log scale) of the amount of active frames in the I4U dataset is shown in Figure 1, and shows considerable variation in utterance duration.

To mitigate this, we also experiment with a weighted average i-vector, where the weights reflect the amount of speech present in the utterance. The number of active frames for each utterance, as determined in the VAD step [10] can be effectively used for this purpose. Hence we replace Equation 4 with

$$\mathbf{w}_{\text{avg}} = \sum_{i=1}^{N} \alpha_i \mathbf{w}_i \tag{6}$$

where the weight α_i is determined as

$$\alpha_i = \frac{k_i}{\sum_j k_j} \quad 1 \leq i, j < N \tag{7}$$

where k_i is the number of active frames for enrollment utterance i and j is the index for summation.

The VAD algorithm typically returns different numbers of active frames for an original utterance and its two noisy versions. Since this may not reflect the amount of active speech in the noisy utterance, we use the value from the original for the corresponding noisy versions. Hence, for a given speaker, the set consisting of an original enrollment utterance and its two noisy versions get the same weight. This is done for each set, with different sets being assigned different weights.

Table 5: Effect of weighted averaging of enrollment i-vectors. Performance in terms of EER (MinDCF.)

Weighting used	Male	Female
No weights used (baseline)	1.06 (0.34)	1.30 (0.50)
Weight based on number of active frames of original utterances	1.16 (0.61)	1.39 (0.61)

The performance obtained with weighted average is presented Table 5. We see that the weighted averaging of the enrollment i-vectors did not bring performance improvement. Although giving less emphasis to i-vectors derived from shorter utterances makes intuitive sense, more sophisticated weighting methods need to be analyzed.

7. Discussion

Several take-home messages stem from the above experiments on the i-vector PLDA framework. First, mismatched train/test conditions do not adversely affect the performance of the system. For noisy training data, clean test utterances give better accuracy than matched noisy test data. Secondly, pooled multicondition training, improves performance considerably, and adds robustness. Our best results were obtained when both the enrollment i-vectors and the PLDA system used multicondition data. It is also interesting to note that the results slightly degraded only slightly when multicondition training was applied to the PLDA model alone. Thirdly, the difference in verification accuracy between averaging the enrollment i-vectors or averaging their scores is relatively minor. I-vector length normalization improves performance, and the point of applying the length normalization (before or after i-vector averaging) is not very critical. Thus, from a practical viewpoint, the most important system configuration needed to handle noisy data is to use length normalized i-vectors, multicondition data for PLDA training, and if possible also for enrollment.

8. Conclusions

In this paper, we experimented with several system-level configurations for the i-vector PLDA speaker verification framework. From our experiments, the importance of multicondition training is seen as the most important configuration for good performance. We also experimented with a heuristic weighted averaging system for the enrollment i-vectors, which did not result in further improvement. Further studies will look at more sophisticated weighting measures.

9. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[2] A. Kanagasundaram, R. J. Vogt, D. B. Dean, and S. Sridharan, "PLDA based speaker recognition on short utterances," in *Proc. Odyssey*, 2012.

[3] A. K. Sarkar, D. Matrouf, P. M. Bousquet, and J. F. Bonastre, "Study of the effect of i-vector modeling on short and mismatch utterance duration for speaker verification," http://www.researchgate.net/publication/. [Online]. Available: http://www.researchgate.net/publication/225280320_Study_of_the_Effect_of_I-vector_Modeling_on_Short_and_Mismatch_Uterance_Duration_for_Speaker_Verification/file/d912f4fd907a71a0f3.pdf

[4] P. Kenny, T. Stafylakis, P. Ouellet, J. Alam, and P. Dumouchel, "PLDA for speaker verification with utterances of arbitrary duration," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013.

[5] M. Mandasari, M. McLaren, and L. D. A., "The effect of noise on modern automatic speaker recognition systems," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012.

[6] M. McLaren and D. van Leeuwen, "Improved speaker recognition when using i-vectors from multiple speech sources," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 5460–5463.

[7] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of Gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4257–4260.

[8] D. van Leeuwen and R. Saeidi, "Knowing the non-target speakers: The effect of the i-vector population for plda training in speaker recognition," in *Proc. Int. Conf. Acoust. Speech, Signal Process.*, 2013.

[9] R. Saeidi and et. al., "I4U submission to NIST SRE 2012: A large-scale collaborative effort for noise-robust speaker verification," in *Proc. Interspeech*, 2013.

[10] T. Kinnunen and P. Rajan, "A practical, self adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7229–7233.

[11] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, 2005.

[12] O. Glembek, L. Burget, P. Matejka, M. Karafiat, and P. Kenny, "Simplification and optimization of i-vector extraction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2011, pp. 4516 –4519.

[13] S. J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc.Int. Conf. Computer Vision*, 2007, pp. 1–8.

[14] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Odyssey*, 2010.

[15] "The NIST year 2010 speaker recognition evaluation plan," 2010, http://www.itl.nist.gov/iad/mig/tests/sre/2010/.

[16] J. Ming, T. Hazen, J. Glass, and D. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 5, pp. 1711–1723, 2007.

[17] D. Garcia-Romero and C. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Proc. Interspeech*, 2011, pp. 249–252.

[18] P. Bousquet, A. Larcher, D. Matrouf, J. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis," in *Proc. Odyssey*, 2012.