INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Probabilistic approach using joint long and short session i-vectors modeling to deal with short utterances for speaker recognition

Waad Ben Kheder, Driss Matrouf, Moez Ajili and Jean-François Bonastre

LIA, University of Avignon, France

Abstract

Speaker recognition with short utterance is highly challenging. The use of i-vectors in SR systems became a standard in the last years and many algorithms were developed to deal with the short utterances problem. We present in this paper a new technique based on modeling jointly the i-vectors corresponding to short utterances and those of long utterances. The joint distribution is estimated using a large number of i-vectors pairs (coming from short and long utterances) corresponding to the same session. The obtained distribution is then integrated in an MMSE estimator in the test phase to compute an "improved" version of short utterance i-vectors. We show that this technique can be used to deal with duration mismatch and that it achieves up to 40% of relative improvement in EER(%) when used on NIST data. We also apply this technique on the recently published SITW database and show that it yields 25% of EER(%) improvement compared to a regular PLDA scoring.

Index Terms: speaker recognition, i-vector, short utterance, duration mismatch, joint modeling.

1. Introduction

Current text-independent speaker recognition systems perform well when enrollment and test data are abundant but their performance suffers greatly when not enough data is provided [1–4]. Such constraint occurs frequently in real applications where it can be difficult to collect enough data since recording conditions cannot always be controlled. One example is speaker authentication in banking applications where users can be reluctant to provide enough speech data particularly at the test phase. Another example is forensic applications where it is really difficult, if not impossible, to collect sufficient data. I-vector based SR systems have been proven to provide an advantageous framework when dealing with short utterances [4] due to its nature of sharing statistical strength among different acoustic regions. Different techniques have been proposed based on this framework to either improve the scoring model by taking into account the duration of segments or exploit phonetic content to achieve more efficient recognition.

The effect of short and mismatched duration utterance modeling was studied in [5] and an ad-hoc score fusion technique was introduced to deal with duration mismatch. In this technique, a test i-vector is projected onto different total variability spaces corresponding to different durations, then the resultant scores are summed and converted to a single value. A different approach was proposed in [6,7] in order to improve the standard G-PLDA model [8] by accounting for the "uncertainty" of the i-vector extraction process. This model, called Full Posterior Distribution PLDA (FP-PLDA), exploits the covariance of the i-vector distribution and improves the recognition performance in presence of duration mismatch by up to 10%.

Alternatively, a range of techniques based on phonetic and prosodic information have also been proposed such as "subregion modeling" [9] where phoneme posteriors are used to partition the acoustic space into subregions modeled by GMMs. Phonetic information present in short utterances is then exploited by scoring test utterance with subregion models. Another "content matching" technique based on phonetic information has been developed in [10] using a DNN where the enrollment data is transformed to be phonetically matched to a given test utterance. A different DNN-based approach has been proposed in [11] to deal with short utterances. In this system, stacked filterbank features are fed to a DNN which is trained as a speaker classifier. Then, the averaged output of the last layer is used as speaker model in test. An improvement by up to 25% is observed when using this model. Finally, an algorithm termed "dual-judgment mechanism" was presented in [12] taking advantage of prosodic features (pitch and formants) in order to improve the decision process in presence of short utterances. Improvement by up to 25% is observed in such system compared to an MFCC-based SR system.

In this paper, we present a new probabilistic approach operating in the i-vector space based on a joint long and short session i-vectors modeling. It aims at improving the quality of short test i-vectors by estimating the corresponding long version using an MMSE (minimum mean square error) estimator. To do so, a large set of i-vectors pairs (long and short) corresponding to the same session are used in the training phase to estimate a joint model. This distribution is then integrated in an MMSE (minimum mean square error) estimator in the test phase to compute an "improved" version of short test i-vectors. This procedure offers two advantages : First, it allows to recover some of the "missing" information in case of short utterances based on a large set of train examples (short/long pairs). Second, it makes the scoring procedure more efficient since training a PLDA model using long utterances can perform poorly on short test utterances [5]. We show that applying this technique on short utterances prior to scoring yields 40% of relative EER improvement compared to a regular PLDA scoring when long utterances are used to train the PLDA. We also apply this technique on the recently published SITW database and show that it yields 25% of EER improvement compared to a regular PLDA scoring.

This paper is structured as follows, Section 2 presents the joint modeling of short and long utterances. Section 3 presents the experimental protocol and Section 4 details the experiments and results achieved using out technique.

2. I-vectors transformation using a joint probability model

It is known that long utterances perform better than the short ones in text-independent speaker recognition tasks [5, 13] since

http://dx.doi.org/10.21437/Interspeech.2016-1302

they convey richer speaker-specific information. But collecting enough data in the real world, either for enrollment or test, can be difficult depending on the application (e.g forensics). Based on this constraint, we propose a transformation operating in the i-vector space that tries to "improve" test i-vectors corresponding short utterances by using a joint model between long and short session i-vectors.

This technique is inspired from the "Stereo Stochastic Mapping" algorithm (SSM) which was first introduced for robust speech recognition [14, 15] then adapted to speaker verification in [16]. In [16], SSM was used to create a mapping between noisy and clean cepstral features based on their joint distribution. In this paper, we use this algorithm in the i-vector space in order to map i-vectors corresponding to short utterances to their "long" versions.

Let's define two random variables x and y representing respectively clean i-vectors corresponding to short and long utterances and let M be the dimension of the i-vector space. We define a third random variable called z as the concatenation of x and y :

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \tag{1}$$

Such variable lie in a $2M$-dimensional space and can be modeled using a mixture of Gaussians :

$$p(z) = \sum_{k=1}^{K} c_k \mathcal{N}(z; \mu_{z,k}, \Sigma_{z,k}) \tag{2}$$

where :

- K is the number of GMM components.
- c_k is the weight of the k^{th} Gaussian.
- $\mu_{z,k}$ corresponds to the mean vector of the k^{th} component.
- $\Sigma_{z,k}$ corresponds to the covariance matrix of the k^{th} component.

This GMM represents the joint distribution between i-vectors corresponding to short and long utterances for each Gaussian k. For each component, it is possible to decompose the mean and covariance matrix as :

$$\mu_{z,k} = \begin{pmatrix} \mu_{x,k} \\ \mu_{y,k} \end{pmatrix} \tag{3}$$

$$\Sigma_{z,k} = \begin{pmatrix} \Sigma_{xx,k} & \Sigma_{xy,k} \\ \Sigma_{yx,k} & \Sigma_{yy,k} \end{pmatrix} \tag{4}$$

Where $\Sigma_{yx,k} = \Sigma_{xy,k}^T$ and $\Sigma_{xy,k}$ models the joint covariance between the two representations (long and short i-vectors).

The problem of transforming a short i-vector y_0 to its long version can be formulated using an MMSE (minimum mean square error) estimator. For a given test i-vector y_0 corresponding to a short utterance, its long version can be estimated as:

$$\begin{aligned} \hat{x} = E[x|y_0] &= \int_x p(x|y_0) x dx = \sum_k \int_x p(x,k|y_0) x dx \\ &= \sum_k p(k|y) \int_x p(x|k,y_0) x dx = \sum_k p(k|y) E[x|k,y] \end{aligned} \tag{5}$$

For each component k, the Schur complement [17] can be used to compute $E[x|k,y]$ as :

$$E[x|k,y] = \mu_{x,k} + \Sigma_{xy,k}\Sigma_{yy,k}^{-1}(y - \mu_{y,k}) \tag{6}$$

The final solution can then be written as :

$$\begin{aligned} \hat{x} &= \sum_k p(k|y) E[x|k,y] \\ &= \sum_k p(k|y)(\mu_{x,k} + \Sigma_{xy,k}\Sigma_{yy,k}^{-1}(y - \mu_{y,k})) \end{aligned} \tag{7}$$

Equation 7 can be re-written as :

$$\hat{x} = \sum_{k=1}^{K} p(k|y_0)(F_k y_0 + g_k) \tag{8}$$

with :

$$F_k = \Sigma_{xy,k}\Sigma_{yy,k}^{-1} \tag{9}$$

$$g_k = \mu_{x,k} - \Sigma_{xy,k}\Sigma_{yy,k}^{-1}\mu_{y,k} \tag{10}$$

In other words, this MMSE-based mapping is a weighted sum of linear functions contributed by each Gaussian component k from the joint GMM distribution $p(x,y)$. The weight is the posterior probability $p(k|y)$ and the linear function is built using the hyper-parameters of each component.

3. Experimental protocol

Our experiments operate on 19 Mel-Frequency Cepstral Coefficients (plus energy) augmented with 19 first (Δ) and 11 second ($\Delta\Delta$) derivatives. A mean and variance normalization (MVN) technique is applied on the MFCC features estimated using the speech portion of the audio file. The low-energy frames (corresponding mainly to silence) are removed.
A gender-dependent 512 diagonal component UBM (male model) and a total variability matrix of low rank 400 are estimated using 15660 utterances corresponding to 1147 speakers (using NIST SRE 2004, 2005, 2006 and Switchboard data). The LIA_SpkDet package of the LIA_RAL/ALIZE toolkit is used for the estimation of the total variability matrix and the i-vectors extraction. The used algorithms are described in [18]. Finally a PLDA-based scoring [8] is applied. The eigenvoice rank the PLDA models is equal to 100 and the eigenchannel matrix is kept full-rank (400). PLDA is preceded by 2 iterations of LW-normalization (spherical nuisance normalization [19]). The equal-error rate (EER) over the NIST SRE 2008 male test data on the "short2/short3" task under the "det7" conditions [20] will be used as a reference to monitor the performance improvements compared to the baseline system.

Short versions of train and test data were generated for different durations : 5s, 10s, 15s, 20s and 30s. This trimming is done in the temporal domain by randomly selecting a continuous portion of speech from the original audio file.

4. Experiments and results

In this section, we start by showing the importance of duration matching in PLDA models training. Then, we apply the proposed technique on short utterances and analyze the amount of data needed to achieve good performances. Finally, we apply it on the SITW database in order to test its performance in real conditions.

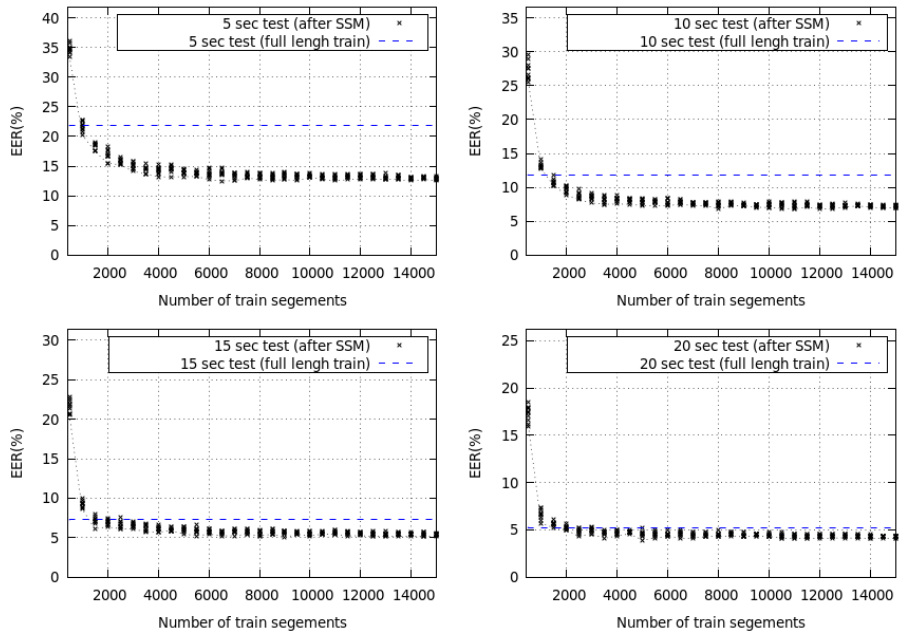

Figure 1: Variation of EER with the amount of i-vectors used to train SSM for 5s, 10s, 15s and 20s test durations (10 measures for each number of segments).

4.1. Effect of train duration on PLDA performance:

For different durations (5s, 10s, 15s, 20s, 30s and full duration), test data are scored using different PLDA models (each one corresponds to a specific duration and the "mixed" model uses train data belonging to all durations). Table 1 shows that matching durations between train and test data improves the system performance compared to a full PLDA (even though long utterances contain more speaker-specific data) which joins the findings of [5].

These results motivate us to transform i-vectors corresponding to short utterances using the technique proposed in Section 2. This transformation would make the scoring with the full-PLDA model more efficient and avoid having to match the durations between train and test in a real SR system (which is impossible for most applications).

4.2. Using the joint i-vectors model for short i-vectors transformation:

For each duration $\mathcal{D} \in \{5s, 10s, 15s, 20s, 30s, \text{full duration}\}$:

1. The i-vectors are extracted for the full-length sessions $\{x_i\}$ and their short versions $\{y_i\}$ of duration $\mathcal{D}$.
2. The distribution $p_{\mathcal{D}}(z)$ is estimated using 20 iterations of the EM algorithm (different number of components are tested for each experiment ; $K \in \{1, 2, 3\}$).
3. Each short enrollment/test i-vector is transformed using Equation 8.

Table 2: Performance of the joint model trained with 15660 pairs of short and long utterances.

	EER			
	Full train	**Joint model 1 Gauss.**	**Joint model 2 Gauss.**	**Joint model 3 Gauss.**
30s	3.59	2.98	3.12	3.25
20s	5.26	4.09	4.69	4.87
15s	7.28	5.21	5.88	6.31
10s	11.84	7.06	8.32	9.35
5s	21.83	13.21	15.32	17.12

Table 2 shows the performance of the joint model compared to a baseline system performance (PLDA trained using long segments). The distribution $p_{\mathcal{D}}(z)$ is trained for each duration $\mathcal{D}$ independently then applied on the corresponding short

Table 1: Effect of the train/test duration on the system performance.

		EER						
		Train speech duration						
		Full	**30s**	**20s**	**15s**	**10s**	**5s**	**Mixed**
Test speech duration	**Full**	***1.59***	2.05	2.49	2.73	3.18	4.56	2.63
	30s	3.59	3.18	***2.96***	3.18	3.87	5.21	3.41
	20s	5.26	4.32	***3.87***	3.87	4.78	5.69	4.55
	15s	7.28	5.92	5.89	***5.50***	5.72	6.54	6.37
	10s	11.84	8.65	7.99	***7.28***	7.75	8.43	9.11
	5s	21.83	17.31	15.91	15.26	13.62	***13.21***	16.40

test i-vectors prior to scoring.

This model improves the performance by up to 40% (1 Gaussian model) and does not require a 2 or 3-components GMM for $p_{\mathcal{D}}(z)$ but uses a large amount of data to be efficient. In the next subsection, we will try to find the minimal amount of data needed to learn the joint distribution while remaining effective.

4.3. Effect of the amount of data used to train $p(z)$ for different durations:

Figure 1 shows the variation of the equal error rate (EER%) with the number of train segments (pairs) to train $p(z)$. In this subsection, we use a 1-Gaussian model for the joint model since this configuration gave the best results in the previous subsection and study the effect of the amount of data used to train each model. It is clear from Figure 1 that 3000 pairs of i-vectors are needed to train $p(z)$ for each duration in order to achieve good results using the joint model.

4.4. Performance on SITW:

In this subsection, we apply this technique on the recently published database SITW [21] to assess the improvement of recognition performance in real conditions. The SITW database contains a short-utterance condition where test utterances have a duration that varies between 15s and 25s. We will test our method on male data. The models used in this experiment are trained from clean NIST train data. Figure 2 shows the distribution of speech durations in the test set corresponding to the clean male 15s-25s condition of the core-core task.

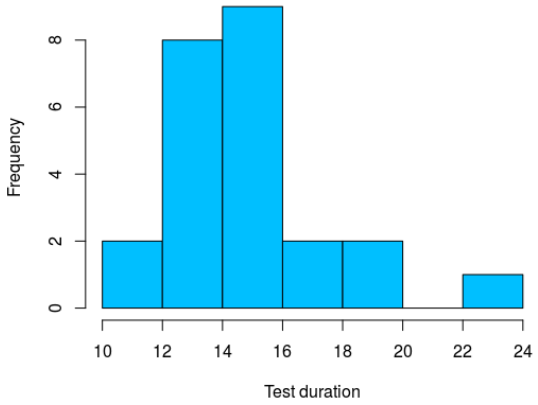

Figure 2: Distribution of speech duration for the male clean 15s-25s condition (the core-core task) in the SITW database.

Table 3 shows the performance of the joint modeling technique on short clean male test sessions of SITW ($< 30s$ of speech duration and $SNR > 20dB$). Different $p_{\mathcal{D}}(z)$ distributions are estimated and applied on all short test i-vectors ($< 25s$ of speech duration). The learned distributions correspond to $\mathcal{D} \in \{30s, 20s, 15s, 10s, 5s\}$ and the "mixed" model uses i-vectors corresponding to all durations.

This model achieves 25% of relative EER improvement compared to the baseline system performance which proves the validity of our technique in real conditions.

Table 3: Performance of the joint model technique on SITW.

EER						
Baseline	**After using the joint model**					
(long train)	**30s**	**20s**	**15s**	**10s**	**5s**	**Mixed**
11.62	9.96	9.53	8.95	9.03	9.53	8.71

5. Conclusion

In this paper, we presented a new probabilistic approach to improve the recognition performance on short utterances based on a joint model of session i-vectors corresponding to long and short utterances. The joint distribution is estimated using a large set of i-vectors pairs (long i-vectors and their short versions generated artificially) then integrated in an MMSE estimator in the test phase to compute an "improved" version of noisy test i-vectors. We tested this algorithm in various configurations and showed it can achieve up to 40% of relative improvement in EER compared to a backend trained using long utterances. Then, we developed a version that can be used to handle multiple durations and tested it on the SITW database and showed a significant gain compared to the baseline system performance.

6. References

[1] M. McLaren, D. Matrouf, R. Vogt, and J.-F. Bonastre, "Applying svms and weight-based factor analysis to unsupervised adaptation for speaker verification," *Computer Speech & Language*, vol. 25, no. 2, pp. 327–340, 2011.

[2] R. J. Vogt, C. J. Lustri, and S. Sridharan, "Factor analysis modelling for speaker verification with short utterances," 2008.

[3] M.-W. Mak, R.-C. Hsiao, and B. Mak, "A comparison of various adaptation methods for speaker verification with limited enrollment data," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1. IEEE, 2006, pp. I–I.

[4] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "I-vector based speaker recognition on short utterances," in *Proceedings of the 12th Annual Conference of the International Speech Communication Association*. International Speech Communication Association (ISCA), 2011, pp. 2341–2344.

[5] 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." in *INTERSPEECH*, 2012, pp. 2662–2665.

[6] S. Cumani, O. Plchot, and P. Laface, "On the use of i–vector posterior distributions in probabilistic linear discriminant analysis," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 4, pp. 846–857, 2014.

[7] S. Cumani, "Fast scoring of full posterior plda models," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 2036–2045, 2015.

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

[9] C. Zhang, D. Wang, L. Li, and T. F. Zheng, "Improving short utterance speaker recognition by modeling speech unit classes."

[10] N. Scheffer and Y. Lei, "Content matching for short duration speaker recognition." in *INTERSPEECH*, 2014, pp. 1317–1321.

[11] E. Variani, X. Lei, E. McDermott, I. Lopez Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.

[12] J. Zhang, J. He, Z. Wu, and P. Li, "Prosodic features based text-dependent speaker recognition with short utterance," in *Computational Intelligence and Intelligent Systems*. Springer, 2015, pp. 541–552.

[13] W. Rao and M.-W. Mak, "Boosting the performance of i-vector based speaker verification via utterance partitioning," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 5, pp. 1012–1022, 2013.

[14] M. Afify, X. Cui, and Y. Gao, "Stereo-based stochastic mapping for robust speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1325–1334, 2009.

[15] J. Du and Q. Huo, "Synthesized stereo-based stochastic mapping with data selection for robust speech recognition," in *Chinese Spoken Language Processing (ISCSLP), 2012 8th International Symposium on*. IEEE, 2012, pp. 122–125.

[16] S. Sarkar and K. Sreenivasa Rao, "Stochastic feature compensation methods for speaker verification in noisy environments," *Applied Soft Computing*, vol. 19, pp. 198–214, 2014.

[17] D. V. Ouellette, "Schur complements and statistics," *Linear Algebra and its Applications*, vol. 36, pp. 187–295, 1981.

[18] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." in *INTERSPEECH*, 2007, pp. 1242–1245.

[19] P.-M. Bousquet, A. Larcher, D. Matrouf, J.-F. Bonastre, and O. Plchot, "Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis." in *Odyssey*, 2012, pp. 157–164.

[20] "The NIST year 2008 speaker recognition evaluation plan," http://www.itl.nist.gov/iad/mig//tests/sre/2008/, 2008, [Online; accessed 15-May-2014].

[21] D. C. A. L. Mitchell McLaren, Luciana Ferrer, "The speakers in the wild (sitw) speaker recognition database," in *Submitted to Interspeech 2016*.