Comparison of speaker recognition approaches for real applications

Comparison of Speaker Recognition Approaches for Real Applications

Sandro Cumani[2], *Pier Domenico Batzu*[1], *Daniele Colibro*[1], *Claudio Vair*[1], *Pietro Laface*[2], *Vasileios Vasilakakis*[2]

[1] Loquendo, Torino, Italy
{First.Lastname}@loquendo.com

[2] Politecnico di Torino, Torino, Italy
{First.Lastname}@polito.it

Abstract

This paper describes the experimental setup and the results obtained using several state-of-the-art speaker recognition classifiers. The comparison of the different approaches aims at the development of real world applications, taking into account memory and computational constraints, and possible mismatches with respect to the training environment. The NIST SRE 2008 database has been considered our reference dataset, whereas nine commercially available databases of conversational speech in languages different form the ones used for developing the speaker recognition systems have been tested as representative of an application domain. Our results, evaluated on the two domains, show that the classifiers based on i-vectors obtain the best recognition and calibration accuracy. Gaussian PLDA and a recently introduced discriminative SVM together with an adaptive symmetric score normalization achieve the best performance using low memory and processing resources.

Index Terms: Speaker Recognition, i-vectors, Joint Factor Analysis, Support Vector Machines

1. Introduction

In recent years, Speaker Recognition Evaluations (SRE) periodically proposed by NIST fostered the improvement of text independent speaker recognition technology. In this context, state-of-the-art technologies were developed with the main goal of optimizing their performance through the minimization of an Actual Detection Cost Function [1] defined by the evaluation rules. This task involves the design of accurate and well calibrated systems, but does not place constraints on their computational and memory requirements (although NIST requires that the system description includes this information). Moreover, development data are usually available, which cover quite well the languages, speaking styles and channels of the evaluation data.

Computational cost and memory requirements are, however, hard constraints in real applications where several recognition channels have to run in real-time on off the shelf hardware. Ever increasing objectives are set by some applications that should be able to process an audio stream tens of times faster than real-time on a single thread to produce a speaker model, or should perform offline tens of thousands speaker comparison tests per second. Thus system design has to account for good tradeoffs between accuracy and costs. Another important issue to be faced is the mismatch between the development and test data often occurring because the speaker recognition system has to operate in previously unseen conditions.

In this work we compare a set of Joint Factor Analysis [2][3][4] and i-vector based classifiers [5][6][7] as perspective candidate technologies for a real application, evaluating their costs and benefits, addressing issues such as the speaker modeling approach, the dimension of the models, the model gender dependency, and the possible mismatch of the development and testing conditions.

Testbeds for our evaluation were the NIST SRE 2008 database and a set of databases of conversational speech in languages different form the ones used for developing the speaker recognition systems.

Using a new score normalization technique and a simple calibration approach that requires a few impostor segments of the application domain, we show that good results can be obtained for the new domain.

The paper is organized as follows: Section 2 summarizes the main modeling approaches that have been compared. Section 3 illustrates the rational for the development and test database selection, and describes their main features. The speaker classifiers and their experimental setup are presented in Section 4. Section 5 introduces a new score normalization procedure. The analysis of the results and our conclusions are given in Section 6 and 7, respectively.

2. Speaker modeling approaches

2.1. Joint Factor Analysis

Joint Factor Analysis (JFA) [2], assumes that an utterance can be modeled by a speaker and channel dependent Gaussian Mixture Model supervector **s**, defined by

$$\mathbf{s} = \mathbf{m} + \mathbf{V} \cdot \mathbf{y} + \mathbf{U} \cdot \mathbf{x} + \mathbf{D} \cdot \mathbf{z} \quad (1)$$

In (1), **m** is the Universal Background Model (UBM) supervector, **V** and **U** are the low rank eigenvoice and eigenchannel matrices defining the speaker and the channel constrained sub-spaces, respectively, and **y, x** are low dimension normally distributed random vectors, usually referred to as speaker and channel factors. The diagonal matrix **D** and its associated common factor vector **z** allow MAP adaptation to be performed in the JFA framework.

In this work we considered a subset of the JFA likelihood computation methods described in [8]: *Linear scoring*, *Channel Point Estimate* and *Integration over Channel Distribution*. These methods are particularly appealing from an application perspective because they compute the likelihoods from the Baum-Welch sufficient statistics, which can be estimated using the UBM in a streaming approach.

2.2. I-vectors

The i-vector approach uses a framework similar to JFA, but solves the problem of intersession compensation in a lower dimensional space [5]. For this purpose, instead of defining different speaker and channel subspaces, this approach estimates a single variability subspace, constrained by a low rank matrix **T**:

$$\mathbf{s} = \mathbf{m} + \mathbf{T} \cdot \mathbf{i} \quad (2)$$

Matrix **T** and the normally distributed random vector **i**, referred to as *i-vector* can be estimated by using the same techniques introduced for JFA.

In this work we considered several i-vector based classifiers including the *LDA WCCN* approach [5], the *Gaussian* and *Heavy-Tailed PLDA* [6], and a recently proposed *Discriminative pair-wise SVM* approach [7], which builds on the PLDA paradigm to train a discriminative system based on Support Vector Machines.

3. Development and evaluation datasets

For evaluating different speaker recognition technologies we had to select a development database, and a test dataset with characteristics similar to the ones often found in application domains. Among the countless variability and mismatches that can occur and taking into account the data availability, our choice was directed to the NIST SRE 2008 database as a development set, and to an evaluation set of commercially available databases in languages not included in the NIST database (Appen).

3.1. SRE08

The SRE08 dataset includes the telephone data of the one conversation train and one conversation test (1conv-1conv core) condition of the NIST SRE 2008 [1]. The average duration of a conversation side is 2.5 minutes. SRE08 has been selected rather than the more recent SRE 2010 dataset, because the former has broader language coverage, while the latter includes English conversations only.

3.2. Appen

The Appen dataset includes 9 two-side conversational telephone speech corpora distributed by Appen Pty. Ltd. [9]. In each corpus the conversations are carried out between 200 native speakers of a given language. The primary use of the Appen databases is language identification of telephone speech. It is, however, possible to use these corpora also for speaker recognition evaluation, because almost all the speakers made two different calls. These calls can be used to create both target and impostor speaker trials. Each Appen conversation side typically lasts 5 minutes, twice as much as the SRE08 segments. On the other hand, all the target speaker's trials are affected by handset and channel mismatch, because the Appen specifications impose that each speaker makes two calls: one from a fixed telephone line, and the other one from a mobile phone.

The Appen databases used in this work include the following languages: Bulgarian, Dutch, Hebrew, Croatian, Italian, European Portuguese, Romanian, Russian and Turkish. One call of each speaker has been randomly selected as an enrollment segment, and the other call is used as a target speaker trial. The set of impostor trials is populated by all the segments having the same language (i.e. belonging to the same corpus) and same gender speakers, not previously selected as enrollment segments. The total number of female target and impostor speaker trials is 810 and 71817, respectively. These numbers increase to 1028 and 117365, respectively, for the male speakers. The trials are evenly distributed among the nine languages.

4. Features and systems description

In this section we first illustrate the two set of features of different dimensions that have been selected for evaluating the tradeoffs between accuracy and computation costs. We detail the procedures and databases that have been used for estimating the knowledge bases common to all the classifiers. Finally we introduce the five classifiers that have been compared in this work.

4.1. Feature Extraction

The first set of features (*45PLP*) is based on Perceptual Linear Predictive parameters, extracted using a 32 ms Hamming window. 19 parameters were computed every 10 ms on Mel spaced power spectrum bands, ranging from 100 to 4000Hz. Feature warping to a Gaussian distribution is performed on a 3 sec sliding window excluding silence frames [10]. We extract 45 PLP parameters: 19 Cepstrals (c0-c18), 19 delta (Δ0-Δ18) and 7 delta-delta ($\Delta\Delta$0- $\Delta\Delta$6).

The second set of features (*25MFCC*) has been selected for small footprint applications. It is based on the standard MFCC parameters, and the Mel spaced bands range from 300 to 3400Hz. 12 Cepstrals (c1-c12) and 13 delta (Δc0-Δc12) were retained.

4.2. UBM

Gender dependent and gender independent, multi-language UBMs, with 512 and 1024 Gaussians, respectively have been trained on approximately 1000 hours of speech data selected from the NIST SRE 2004, 2005 and 2006, LDC Callfriend [11], and Italian, Portuguese and Swedish SpeechDat2 corpora [12]. The models were trained running 10 iterations of an approximation of the EM algorithm, which, for the sake of efficiency, updates for each frame the best Gaussian statistics only.

4.3. Joint Factor Analysis

The Joint Factor Analysis models have been trained following the guidelines of [2], [3] and [4] with some slight variations.

Matrix **V** has been trained on a subset of the NIST SRE 04/05/06 datasets, including at least 3 conversations per speaker. The number of eigenvoices was fixed to 300. Matrix **U** has been trained on the same data used for the eigenvoice training, including both telephone speech and telephone speech acquired through an auxiliary microphone. 100 eigenchannels were extracted and used in our experiments. Matrix **D** estimation it set to values equivalent to the ones obtained by relevance MAP, with relevance factor equal to 16.

4.4. I-vector subspace

The same procedure that allows the eigenvoice matrix **V** to be obtained can be used for estimating the single variability matrix **T**, providing the procedure a supervector per conversation rather than a supervector per speaker. Matrix **T**, which is the prior knowledge for extracting the i-vectors, has been trained using the same dataset employed for estimating the **V** matrix but excluding the filter on the minimum number of segments per speaker. The i-vector dimension was fixed to 400 for all the experiments.

4.5. LDA-WCCN

In the LDA-WCCN approach, intersession compensation is obtained by means of Linear Discriminant Analysis (LDA), where all the i-vectors of the same speaker are associated with the same class. The LDA matrix has been trained using the same dataset of the JFA **V** matrix, including telephone and microphone segments.

Table 1. *Comparison of the performance of different classifiers on two databases.*

System	Male		Female		Male		Female	
	EER %	MinDCF	EER %	MinDCF	EER %	MinDCF	EER %	MinDCF
	45 PLP 1024G GI SRE08				**25 MFCC 512G GI SRE08**			
JFA Linear scoring	**4.05**	**0.252**	6.83	0.367	**5.51**	0.291	**8.36**	0.460
JFA Channel Point	4.05	0.256	6.67	0.358	6.47	0.294	8.79	0.474
JFA Channel Integral	4.05	0.253	**6.57**	**0.353**	5.99	**0.283**	8.53	**0.450**
LDA WCCN	4.21	0.232	6.67	0.300	5.64	0.253	8.52	0.385
Gaussian PLDA	3.59	**0.195**	5.87	0.289	5.15	0.224	7.47	0.375
Heavy Tailed PLDA	3.73	0.199	5.86	0.290	**4.91**	**0.223**	7.39	0.368
SVM	**3.35**	0.205	**5.63**	**0.284**	5.06	0.230	**7.05**	**0.348**
	45 PLP 1024G GI APPEN				**25 MFCC 512G GI APPEN**			
JFA Linear scoring	**4.96**	**0.225**	**5.70**	**0.253**	**5.52**	**0.228**	**5.66**	**0.246**
JFA Channel Point	5.16	0.247	6.17	0.283	5.93	0.233	6.17	0.260
JFA Channel Integral	5.16	0.245	6.17	0.280	5.84	0.230	5.91	0.259
LDA WCCN	4.47	**0.163**	5.19	0.171	5.74	0.188	5.43	0.231
Gaussian PLDA	4.28	0.165	4.69	0.169	4.98	0.185	4.79	0.218
Heavy Tailed PLDA	4.67	0.175	4.69	0.162	5.35	0.199	**4.71**	0.213
SVM	**3.99**	0.166	**4.45**	**0.144**	**4.56**	**0.180**	5.18	**0.197**

The LDA removes the nuisance directions from the i-vectors by reducing the feature dimensions (in our tests from 400 to 200). These speaker features are finally normalized by means of Within Class Covariance Normalization (WCCN) [5][13]. The WCCN transformation was trained on a subset of the LDA training data (NIST SRE06 in our settings).

4.6. Gaussian and Heavy Tailed PLDA

The Gaussian and Heavy Tailed PLDA implementations follow the framework illustrated in [6]. We trained models without the channel factor component using 200 dimensions for the speaker factors. The PLDA models have been trained with the same data used for training the **V** matrix.

It is worth noting that LDA WCCN and Gaussian PLDA training and scoring have been performed by using L_2 normalized i-vectors. This simple rescaling achieves significant reduction of both the Equal Error Rate (EER) and the Detection Cost Function (DCF), whereas the same transformation is not effective for Heavy-Tailed PLDA.

4.7. SVM

In [7] a fast discriminative training procedure for a linear-Gaussian model has been proposed. In this approach, rather than modeling the speaker classes, a SVM binary classifier is trained to classify a pair of utterances as either belonging to the same speaker or to two different speakers. In particular, the observation patterns are i–vectors *pairs*, the SVM "target" class corresponds to "same speaker pair", and the "non–target" class to "different speaker pair". The SVM has been trained with the same data used for PLDA training.

5. Score normalization

It is worth noting that we apply normalization to the scores provided by all the illustrated techniques, even if in the PLDA approach proposed in [6] the scores are intrinsically log-likelihood ratios. Nevertheless, we believe that the normalization of the impostor score distribution (done for example by the classical Z/T-Norms and their combination and variants) is a key factor for real applications, because it allows clients to customize their normalization set for improving the system calibration and accuracy.

The creation of a custom normalization set for normalizing the impostor score distribution is much less demanding than a full log-likelihood ratio (LLR) calibration, because it involves easily obtainable "impostor" trials only. In contrast, a full LLR calibration requires a development set with multiple speakers and multiple segments for each speaker, to set-up the impostor and true-speaker trials needed by the LLR calibration.

We used two normalization techniques: ZT-Norm for the JFA scores, and a new *Adaptive S-Norm* (AS-Norm) for the scores produced by the i-vector based classifiers.

The AS-Norm is derived from the AT-Norm [13], but preserves the symmetrical property of the S-Norm [6]. The matching score s of two i-vectors i_1 and i_2 is normalized according to

$$\frac{1}{2}\cdot\left[\frac{s-\mu_1(N_2)}{\sigma_1(N_2)}+\frac{s-\mu_2(N_1)}{\sigma_2(N_1)}\right] \qquad (3)$$

where μ_1 and σ_1 are the mean and standard deviation of the scores obtained by matching i_1 against a normalization subset N_2 depending of i_2, and the same notation dually applies to the second term in parenthesis. The selection of the normalization subset follows the procedure in [13].

Our normalization set includes 273 male and 348 female segments, selected from different languages conversations of the NIST SRE 04/05/06. It is worth noting that we always performed gender dependent normalization.

6. Results

Table 1 summarizes the speaker recognition performance of the evaluated approaches. The results are provided for the Gender Independent 45 PLP 1024 Gaussians and 25 MFCC 512 Gaussians systems. I-vector systems outperform JFA in all the conditions. The best results, in average, are obtained by the SVM system, both for SRE08 and Appen.

Table 2 shows the results of applying different normalization approaches to the Gender Independent 45 PLP 1024 Gaussians SVM system scores. Raw scores, S-Norm and the proposed AS-Norm techniques are compared, the latter obtaining the best results in nearly all conditions. Similar results were obtained for the other i-vector systems (LDA WCCN and PLDA).

Table 2. *Comparison of score normalization techniques for the 45 PLP 1024G GI i-vector SVM based system.*

System	SRE08 Male		SRE08 Female		Appen Male		Appen Female	
	EER %	MinDCF	EER %	MinDCF	EER %	MinDCF	EER %	MinDCF
raw	4.19	0.230	5.91	**0.279**	4.64	0.181	5.31	0.177
s-norm	3.47	0.213	5.75	0.298	4.18	0.179	4.69	0.174
as-norm	**3.35**	**0.205**	**5.63**	0.284	**3.99**	**0.166**	**4.45**	**0.144**

Table 3 reports the minimum (Min) and actual DFC (Act) for the Appen database. The actual DCF is computed using the threshold that optimizes the DCF on SRE08. Comparing Min and Act DCFs we see that the systems are totally out of calibration. To solve this problem, a very simple technique has been devised. It uses a small development set of speakers segments from the target domain to create impostor trials, which are exploited to normalize the impostor distribution. In particular, the mean and standard deviation of the development trial scores allow rescaling the scores so that the impostor scores are distributed according to the standard normal distribution. The results given in Table 3 refer to a calibration experiment where just 36 random segments of the Appen domain were selected as development set and excluded from the trial set. The results are the average over 10 different random selections. Since the number of excluded trials is negligible the minimum DCF does not change. The ActComp column in Table 3 shows that most of the out of calibration effects can be recovered by this compensation technique using.

Table 3. *System performance on the Appen database*

GI System	Male DCF Min / Act / ActComp	Female DCF Min / Act / ActComp
Linear scoring	0.224 / 2.722 / 0.250	0.254 / 1.670 / 0.268
LDA WCCN	0.163 / 0.895 / 0.186	0.171 / 0.726 / 0.188
Gaussian PLDA	0.165 / 1.154 / 0.182	0.169 / 0.792 / 0.176
SVM	0.166 / 1.017 / 0.179	0.144 / 0.808 / 0.157

Table 4. *Comparison of SVM GI and GD systems*

		Male		Female	
		EER %	MinDCF	EER %	MinDCF
SRE08	GI 25MFCC	5.06	0.230	7.05	0.348
	GI 45 PLP	3.35	0.205	5.63	0.284
	GI Fusion	3.35	0.191	**5.49**	**0.275**
	GD 45 PLP	**3.20**	**0.187**	5.58	0.281
Appen	GI 25MFCC	4.56	0.180	5.18	0.197
	GI 45 PLP	3.99	0.166	4.45	0.144
	GI Fusion	**3.79**	**0.153**	4.32	**0.138**
	GD 45 PLP	4.18	0.163	**4.20**	0.167

Table 4 compares different SVM systems and their fusion: the combination of the small 25 MFCC 512 Gaussians and of the 45 PLP 1024 Gaussians Gender Independent systems, is always effective. Moreover, this fusion often outperforms the more expensive Gender Dependent 45 PLP 1024 Gaussians system. This result is of particular relevance in a product perspective.

7. Conclusions

In this paper we compared several speaker recognition techniques targeting real world applications. We evaluated a 45 PLP 1024 Gaussians configuration and a smaller 25 MFCC 512 Gaussians system, using JFA and i-vector based classifiers. In both cases, i-vector techniques outperform JFA methods. Among the i-vector approaches, Gaussian PLDA using normalized i-vectors, and discriminative pair-wise SVM are the best classifiers in terms of accuracy. For these methods the proposed AS-norm proved to be effective.

The gender dependent (GD) and gender independent (GI) systems have comparable accuracy, whereas slightly better results are obtained by combining a small and a large GI system, using fewer resources than the GD system.

The results on the Appen database show that a system developed for a NIST SRE evaluation can be profitably used for real world application, even on different languages. Application independent calibration remains an open issue. However a simple linear transformation of the scores, estimated on very few utterances in the target domain, succeeded in normalizing mean and variance of the impostor score distribution, greatly improving the actual DCFs.

8. Acknowledgements

We would like to thank Fabio Castaldo for his contributions to our development environment, and Ondřej Glembek from BUT for suggesting us the use of normalized i-vectors. Vasileios Vasilakakis is supported by the FP7/2007-2013 European Programme under grant agreement n.238803.

9. References

[1] Available at http://www.itl.nist.gov/iad/mig/tests/sre

[2] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," IEEE Trans. Audio, Speech and Language Processing, vol. 16, no. 5, pp. 980–988, 2008.

[3] L. Burget, P. Matějka, V. Hubeika, J. Černocký.: Investigation into variants of Joint Factor Analysis for speaker recognition, In: Proc. Interspeech 2009, p. 1263-1266, 2009.

[4] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms - Technical Report CRIM-06/08-13. Montreal, CRIM, 2005.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", in IEEE Transactions on Audio, Speech, and Language Processing, Vol.19, n. 4, pp. 788-798, 2011.

[6] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors", in Keynote presentation, Odyssey 2010, The Speaker and Language Recognition Workshop, 2010.

[7] S. Cumani, N. Brümmer, L. Burget, and P. Laface, "Fast discriminative speaker verification in the i–vector space," in Proc. ICASSP 2011. pp.4852-4855, 2011.

[8] O. Glembek, L. Burget, N. Dehak, N. Brummer, P. Kenny, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," in Proc. ICASSP 2009, pp.4057-4060, 2009.

[9] Available at http://www.appen.com.au

[10] J. Pelecanos, and S. Sridharan, "Feature Warping for Robust Speaker Verification," in Proc. 2001: A Speaker Odyssey, pp. 213-218, 2001.

[11] Available at http://www.ldc.upenn.edu/Catalog.

[12] Available at http://www.speechdat.org/SpeechDat.html.

[13] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-Class Covariance Normalization for SVM-Based Speaker Recognition," in Proc. ICSLP 2006, pp. 1471-1474, 2006.

[14] D. E. Sturim, D. A. Reynolds, "Speaker Adaptive Cohort Selection for T-norm in Text-Independent Speaker Verification", in Proc. ICASSP 2005, pp. 741-744 , 2005.