I3A System for Evalita 2009 Speaker Verification Application Evaluation

Jesús Villalba, Eduardo Lleida, Alfonso Ortega, Carlos Vaquero, and Antonio Miguel

Communications Technology Group (GTC), Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain
{villalba,lleida,ortega,cvaquero,amiguel}@unizar.es

Abstract. This paper describes the I3A-University of Zaragoza speaker recognition system in the Evalita 2009 Speaker Identity Verification Application task. Two systems have been submitted for evaluation: a primary system based on Joint Factor Analysis and a secondary system based on classical MAP adaptation without any channel compensation for comparison. We present results for both systems on the development and evaluation datasets and focus on the discussion about the influence of channel compensation, score normalization and calibration on the performance.

Keywords: Speaker Recognition, Speaker Verification, GMM, Joint Factor Analysis, MAP.

1 Introduction

The Evalita 2009 evaluation has the purpose of promoting the research on speech technologies for the Italian language. This evaluation allows evaluating and comparing different approaches on a common database and scoring metric. The Speaker Identity Verification (SIV) for Applications [1] is one of the tasks included in the evaluation campaign. Systems submitted for this task have been evaluated on a "remote authentication by phone" use case scenario where, given a speech recording, the system must accept or reject the identity claimed by the speaker.

This paper is organized as follows: In section 2, we describe both submitted systems. In section 3, we describe the experiments and data used on the system development and the results we have got on the evaluation dataset. Conclusions are presented in section 4.

2 System Description

We have submitted two systems for evaluation. Both systems are based on modeling the distribution of the speaker's speech frames using Gaussian Mixture Models (GMM) [2]. These systems share the same feature extraction, frame selection, score

normalization and calibration procedures but differ in the classification step. The primary system is based in the Joint Factor Analysis theory (JFA) [3] which, in the last years, has become the state of the art approach for channel compensated speaker modeling. The top performing systems participating in the NIST Speaker Recognition Evaluations [4] are based on JFA. As secondary system, we have submitted a system based on classical MAP [2] adaptation without channel compensation with the purpose of evaluating the channel compensation ability of the JFA system compared to a system without channel compensation. In the following sections we present the individual components of the systems.

2.1 Feature Extraction

The front-end extracts 16 MFCC, including C0, from the speech signal using a sliding window of 25 msec. every 10 msec. MFCC filter bank is band limited to fit the telephone channel from 250 to 3600 Hz. First and second derivatives are appended to the feature vector. Voice detection is done based on the comparison of the long term envelope of the signal spectrum to the average noise [5]. Noise is estimated during the stationary periods of the signal. After frame selection, features are short time Gaussianized with a 3 sec. sliding window.

2.2 Classical MAP

Speaker models are derived by Bayesian adaptation [2] from the UBM using a relevance factor of 16. The scoring is done by log-likelihood ratio (LLR) between the probabilities of speech segment given the Target and the UBM models.

2.3 Joint Factor Analysis

Joint Factor Analysis has been implemented following Kenny's recipe [3]. Each recording is modeled by a supervector M=m+Vy+Dz+Ux where m are the UBM means, V is the speaker space matrix, U is the channel space matrix and D is a diagonal matrix that accounts for the speaker variability not included in the speaker space. U and V are trained by ML iterations from the training data described in the next section. Due to the lack of data for training D, we have used its relation with classical MAP as $D^2=\square/r$ with r=20.

For scoring, channel compensation of the test segment is done using MAP point estimates of the channel factors and LLR between Target and UBM model is performed.

2.4 Score Normalization and Calibration

Gender dependent ZT-Norm has been applied to both systems. After ZT-Norm scores have been gender dependent calibrated for optimum DCF using the logistic regression algorithms implemented in the FoCal package [7].

3 Experiments

3.1 Datasets

We have used only the data provided by Evalita organizers for development purposes. This data consist of three separate datasets: ubm, training and development. The ubm and training datasets are made of short conversation excerpts which we have concatenated to form full 1 minute conversations. In this way we have got the concatenated ubm and training datasets.

We have trained gender dependent UBM of 512 Gaussians using the concatenated ubm dataset (600 conversations of 30 speakers by gender). The same data is used for training 20 speaker factors for the JFA system. 50 channel factors have been trained using the ubm dataset without concatenation. We use the non concatenated dataset because the phonetic mismatch between short utterances is bigger. In this way, we can find the directions of intersession variability between training and test better.

Models for 100 target speakers (50 male + 50 female) are trained from the concatenated training dataset. Six training conditions have been considered:

- TC1. "PSTN short": 1 conversation of PSTN call.
- TC2. "GSM short": 1 conversation of GSM call.
- TC3. "PSTN long": 3 conversations of PSTN call.
- TC4. "GSM long": 3 conversations of GSM call.
- TC5. "Mixed short": 1 conversation of PSTN call + 1 conversation of GSM call.
- TC6. "Mixed short": 3 conversations of PSTN + 3 conversations of GSM calls.

The development dataset includes recordings of 32 of the 100 speakers in the training set (16 male + 16 female). Participants must use this dataset for making a list of trials for evaluating and calibrating their systems. We have made a list of 321 target trials and 3200 non target trials. For generating the non target trial list, we have used 100 recordings per target speaker. These recordings are selected randomly from the development dataset belonging to the others 15 speakers of the same gender. Two test conditions are considered:

- TS1. "SHORT": 10 seconds excerpts.
- TS2. "LONG": 30 seconds excerpts.

We have trained models of cohort speakers for T-Norm from the concatenated ubm dataset. The cohorts used for each trial are adapted to the training condition. That means cohorts and target speaker are trained with the same amount of speech and the same kind of channel. In this way we have selected 90 cohort models per gender and training condition.

Segments for Z-Norm are selected from the non concatenated ubm dataset. Segment length is adapted to the test condition in such a way that utterances between 5 and 18 seconds are used for TS1 condition and utterances longer than 18 seconds are selected for the TS2 condition. We select approximately the same number of PSTN and GSM segments for Z-Norm. In this way we have selected 600 segments

per gender for the TS1 condition and 112 male and 107 female segments for TS2 condition.

3.2 Development Results

In Tables 1 and 2, we show the results of both systems on the development trial list. We present results with and without ZT-Norm. We can see, for the MAP system, normalization is of great help to improve the results of both types of channel together. On the other hand, it does not help if we consider both channels separately. The reason to this is that the score range for each test channel is very different before normalization, and EER and minimum DCF thresholds too.

For the JFA system, we have an improvement, considering both channels jointly as well as separately, compared to the MAP system. This is due to the channel compensation done using the channel factors in the JFA model. On the hand, the improvement with ZT-Norm is lower because the score range is more similar across conditions. In the fact, performance with ZT-Norm is poorer in some TS2 conditions. We think this is due to the fact there are fewer segments available for Z-Norm for this condition.

Table 1. Results of the development experiments for TS1 condition.

TC	Test Ch.	JFA (Primary)				MAP (Secondary)			
		No ZT-Norm		ZT-Norm		No ZT-Norm		ZT-Norm	
		EER(%)	DCF	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF
1	ALL	13.4	0.32	12.77	0.31	18.1	0.41	13.3	0.35
	PSTN	7.7	0.14	7.6	0.12	7.6	0.19	7	0.15
	GSM	18	0.4	19.4	0.34	21.3	0.4	20	0.4
2	ALL	16.8	0.3	14	0.33	19.5	0.39	16.2	0.38
	PSTN	19.4	0.3	17	0.34	19.3	0.35	18.8	0.42
	GSM	11.8	0.25	11.3	0.24	9.4	0.28	9.3	0.28
3	ALL	10.5	0.3	9	0.24	15.3	0.355	11.2	0.29
	PSTN	1.75	0.04	2.9	0.03	4.2	0.08	4.1	0.08
	GSM	14	0.37	12.6	0.28	15	0.33	16	0.36
4	ALL	9.9	0.28	8.7	0.27	16.5	0.355	12.7	0.27
	PSTN	14.3	0.31	12.5	0.29	17.5	0.34	15.2	0.31
	GSM	4.6	0.08	3.2	0.26	6.6	0.12	4.7	0.1
5	ALL	7.8	0.19	7.47	0.16	8.3	0.19	7.5	0.18
	PSTN	4.6	0.16	5.2	0.16	6	0.13	6.7	0.15
	GSM	9.7	0.2	8.6	0.16	9.6	0.23	10	0.20
6	ALL	4	0.075	3.4	0.06	5.8	0.15	5.6	0.15
	PSTN	2.9	0.028	3.5	0.042	4.6	0.1	4.7	0.11
	GSM	4.6	0.1	3.3	0.06	7.5	0.18	6.9	0.17

We must take into account that, despite eigen-channels have been trained using the same kind of channels as the test ones, there is yet a big difference between same channel and cross channel trials. Besides, for same channel trials, GSM performs

poorer than PSTN. This means there is yet room for improvement using channel compensation techniques.

Table 2. Results of the development experiments for TS2 condition.

TC	Test Ch.	JFA (Primary)				MAP (Secondary)			
		No ZT-Norm		ZT-Norm		No ZT-Norm		ZT-Norm	
		EER(%)	DCF	EER(%)	DCF	EER(%)	DCF	EER(%)	DCF
1	ALL	9.6	0.250	10.1	0.300	14	0.350	8	0.290
	PSTN	2	0.060	4.6	0.050	3.3	0.075	3.4	0.050
	GSM	11.1	0.320	14	0.360	13.4	0.300	12	0.330
2	ALL	9.5	0.220	10.6	0.260	13	0.300	10.3	0.240
	PSTN	12.9	0.310	15.3	0.300	13.3	0.240	14.5	0.300
	GSM	5.2	0.125	7	0.110	4.7	0.150	5.2	0.140
3	ALL	7.4	0.190	5.3	0.120	12.7	0.290	7.47	0.200
	PSTN	0.13	0.007	1.3	0.130	0.7	0.005	2	0.016
	GSM	9	0.220	8.2	0.150	10	0.230	11.6	0.250
4	ALL	4.9	0.170	5.3	0.170	10.9	0.280	7.5	0.200
	PSTN	8	0.220	7.3	0.210	10.5	0.200	11.5	0.230
	GSM	1.17	0.011	1.7	0.010	1.9	0.025	2.3	0.025
5	ALL	3.7	0.120	4.12	0.110	4.6	0.090	4.6	0.125
	PSTN	2.6	0.080	2.6	0.080	4	0.070	4	0.090
	GSM	3.6	0.095	5.2	0.100	5.2	0.100	4.6	0.150
6	ALL	1	0.026	1.25	0.012	2.8	0.040	2.8	0.040
	PSTN	1.3	0.015	1.33	0.012	1.4	0.040	2.6	0.044
	GSM	1.17	0.025	1.17	0.006	3	0.030	2.9	0.039

3.3 Evaluation Results

In Tables 3 and 4, we present the results on the evaluation dataset. Results by kind of test channel are not included due to lack of space. These are quite close to the ones of the development dataset.

Table 3. Results for TS1 condition.

TC	JFA (Primary)			MAP (Secondary)		
	EER(%)	Min. DCF	Act. DCF	EER(%)	Min. DCF	Act. DCF
1	13.45	0.302	0.323	13.45	0.347	0.356
2	13.71	0.341	0.341	13.48	0.375	0.380
3	8.89	0.212	0.227	10.24	0.302	0.315
4	8.30	0.257	0.268	10.87	0.299	0.302
5	7.93	0.193	0.211	8.39	0.246	0.265
6	4.58	0.140	0.171	7.48	0.197	0.221

We want to emphasize the actual and minimum DCF are close, so the calibration done using the development trials is quite good. Another think we observe in the

development and evaluation results is we the more data available in training and test the more gain we get from the JFA approach. We can see this comparing the gains for TC1 and TC2 with TC6. This is due to the fact the channels factors are estimated more precisely if more data is available.

Table 4. Results for TS2 condition.

TC	JFA (Primary)			MAP (Secondary)		
	EER(%)	Min. DCF	Act. DCF	EER(%)	Min. DCF	Act. DCF
1	8.72	0.204	0.216	8.42	0.241	0.258
2	9.94	0.286	0.310	9.17	0.245	0.258
3	4.05	0.102	0.117	5.35	0.180	0.184
4	5.23	0.159	0.176	6.63	0.188	0.196
5	4.22	0.103	0.119	5.18	0.163	0.179
6	1.97	0.031	0.070	3.68	0.121	0.162

4 Conclusions

We have presented a state of the art speaker verification system based on JFA and its performance on the Evalita 2009 task. We have seen this system is able to improve the error rates on cross channel conditions compared to classical MAP but we can do a great deal yet for improving this. This is especially important for the conditions where less data is available. In these cases, lack of data does not allow estimate precisely the channel factors of the JFA model. We have seen score normalization and calibration is essential for optimum performance across conditions.

Acknowledgments. This work has been supported by the Spanish Government through national project TIN2008-06856-C0504.

References

1 Aversano, G.: EVALITA 2009 - Speaker Identity Verification - Application Track Task Guidelines, http://evalita.fbk.eu/doc/Guidelines_evalita09_SIV-Application_track.pdf
2 Reynolds D., Quatieri T., Dunn R.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, vol. 10, pp. 19--41 (2000)
3 Kenny, P.: Joint Factor Analysis for Speaker and Session Variability: Theory and Algorithms. Technical Report CRIM-06/08-13, Montreal, CRIM (2005)
4 NIST SRE Evaluation http://www.itl.nist.gov/iad/mig/tests/sre/
5 Ramirez J., Segura J., Benitez C., De La Torre A., Rubio A.: Efficient Voice Activity Detection Algorithms Using Long-Term Speech Information. Speech Communication, vol. 42, pp. 271--287 (2004)
6 Pelecanos J., Sridharan S.: Feature Warping for Robust Speaker Verification. In: Proceedings of Proc. Odyssey (2001)
7 FoCal Package, http://www.dsp.sun.ac.za/~nbrummer/focal