AGNITIO's Speaker Recognition System for EVALITA 2009

Niko Brümmer and Albert Strasheim

AGNITIO

Abstract. AGNITIO's submission for the EVALITA 2009 Speaker Identity Verification Application Track was a fusion of a state-of-the-art JFA system and a new I-Vector system. We briefly describe the two systems and their fusion and then highlight the challenging *cross-channel* conditions of this evaluation, where enrollment was via land-line and verification via mobile, or vice versa. Despite the availability of a matched development database of 10 land-line and 10 mobile calls for each of 60 speakers, the error-rates for cross-channel verifications remained high.

Key words: speaker recognition, EVALITA, joint factor analysis, i-vector.

1 Introduction

The AGNITIO submission for EVALITA'09 SIVAP was a fusion of two subsystems, a *JFA system* and an *I-Vector* system. Our existing JFA system was trained on the Switchboard and Mixer databases and adapted to this evaluation by populating the T-norm cohort with EVALITA development data. The I-Vector system is a new experimental system. Each of the JFA and I-Vector systems has state-of-the-art performance at about 6% EER, on the telephone portion ('DET6') of NIST SRE 2008 [9].

2 EVALITA SIVAP'09 data

In order to understand our development process and evaluation results below, a brief description of the EVALITA data is necessary. The data was collected from an IVR machine, where several speakers called the IVR multiple times. During each call, the speaker responded multiple times to prompts from the IVR. Callers were from all over Italy and spoke Italian.

Every speaker called multiple times from a land-line (probably the *same* land-line every time) and also multiple times from a cellphone (most probably the *same* cellphone every time). That is, different speakers used different telephones, but most speakers probably used only two different telephones.

The EVALITA data was divided into two subsets, which we label the *development data*[1], and the *evaluation data*. There is no overlap in speakers between these subsets.

[1] Referred to as *UBM training data* in the evaluation plan [1]

– The *development data* was released to participants beforehand to use for any development purpose. The development data has for each of 30 male and for each of 30 female speakers, 10 land-line calls and 10 cellphone calls. This gives a total of 20 calls per speaker, or 1200 calls in total.
– A part of the evaluation data, which we label the *tuning data*[2], was also released to participants beforehand for limited use in tuning of calibration parameters.

For more details, see [1]. Our use of the development data is described in the rest of this paper. We used the tuning data only as a development test set and did not use it in any numerical optimizations of system parameters.

3 Common system components

The feature extractor and UBM were common to both JFA and I-Vector systems:

3.1 Features

We used 60-dimensional acoustic features, with a 10ms frame rate, composed of 19 MFCCs plus log energy and augmented by first and second-order deltas. Features were normalized with short-time Gaussianization [7].

3.2 UBM

For the UBM [8], a 2048-component GMM was trained on all of the development data as indicated in Table 1, by using an HTK-style EM algorithm, with component splitting [10], to optimize a maximum likelihood criterion.

4 JFA system

Joint Factor Analysis [2, 3], pioneered by Patrick Kenny [4, 5], is a GMM-based generative modeling technique for speaker recognition. It forms an important part of the state-of-the-art in text-independent speaker recognition, as demonstrated at the recent NIST Speaker Recognition Evaluations from 2004 to 2008—see e.g. [9] and references therein. The *joint* in JFA refers to the fact that both *within* and *between* speaker variabilities are explicitly modeled, while *factor analysis* refers to the technique for representing the covariance matrices associated with these variabilities in very high dimensional spaces.

In our implementation we suppose that a different 2048-component GMM generates the 60-dimensional acoustic feature vector sequence of every call of every speaker. That is, the GMM for call j of speaker i is represented by its *supervector* (the concatenation of mean vectors of the GMM), which is modeled as:

$$\mathbf{m}_{ij} = \boldsymbol{\mu} + \mathbf{V}\mathbf{y}_i + \mathbf{U}\mathbf{x}_j \tag{1}$$

[2] Referred to as *development data* in the evaluation plan [1]

where $\boldsymbol{\mu}$ is the 122880-dimensional UBM supervector; where $\mathbf{y}_i$ is a 300-dimensional speaker-dependent vector; where $\mathbf{x}_j$ is a 100-dimensional call-dependent, but speaker-independent nuisance variable; and where $\mathbf{V}$ and $\mathbf{U}$ are tall thin matrices of rank 300 and 100 respectively.

The JFA system is trained on large quantities of development data, by using EM algorithms to make maximum-likelihood estimates of the parameters $\boldsymbol{\mu}$ (the UBM), $\mathbf{U}$ and $\mathbf{V}$. In Table 1, we show which databases were used to train these different parameters.

Table 1. Development data utilization. I denotes I-Vector system and J denotes JFA system.

	UBM	**V, U**	**W, C, D**	**Znorm**	**Tnorm**	**Snorm**	**Fusion**
SRE'04,'05	I, J	J	I	J			
SRE'06	I, J	J	I				
SWBPh2 Prts 2,3	I, J	J	I				
SWB cell	I, J	J	I				
Fisher English	I, J						
EVALITA dev.	I, J				J	I	I, J

Note significantly, that we did *not* use the EVALITA development data in our estimates of $\mathbf{V}$ and $\mathbf{U}$. It could be expected that this data, which has 20 sessions of each of 60 speakers would have been very good to train inter- and intra-speaker variability parameters, but in our development experiments we found that we could not get significant gain in the more difficult EVALITA cross-channel conditions by including this data. We therefore preferred to exclude this data from JFA training and to use it exclusively for development testing and training of fusion and calibration parameters.

In retrospect, after comparison [11] of our results with other participants, who did use EVALITA development data to train JFA parameters (or equivalents), it is still not clear whether this was a good decision or not. For two of the TS2 cross-channel conditions: G-P (enroll with one GSM call—test over land-line) and 3G-P (enroll with three GSM calls—test over land-line), our system performed best according to both minDCF and actDCF criteria. However, for land-line enrollment and GSM test, other sites did better.

4.1 ZT-norm

The JFA system is gender-independent, but it uses gender-dependent ZT-normalization [3, 6]. We populated the Z and T-norm cohorts of the JFA system as indicated in Table 1. The Z-norm cohorts (one per gender) were selected from the telephone data in NIST SRE'04 and SRE'05. For the T-norm cohort, we found it was better to use EVALITA development data. We created 30 T-norm models per gender, by pooling all of the 20 calls per speaker.

5 I-Vector system

Our *I-Vector* system is based on a recent idea of Najim Dehak [12], although the details of our implementation after the extraction of the 'i-vectors' differs from his. The basic idea is to simplify the JFA model of (1) to the form:

$$\mathbf{m}_{ij} = \boldsymbol{\mu} + \mathbf{W}\mathbf{z}_{ij} \tag{2}$$

where $\mathbf{z}_{ij}$ is a 400-dimensional hidden variable that accounts for both within and between-speaker variability in the observed data and where $\mathbf{W}$ is a tall thin matrix of rank 400, which presumably spans approximately the same subspace of supervector space that $\left[\mathbf{U}\ \mathbf{V}\right]$ would in the full JFA case.

The I-Vector 'extractor' is trained by estimating $\mathbf{W}$ via EM algorithm on a large training data set, in a way very similar to the estimation of $\mathbf{V}$ and $\mathbf{U}$.

Once $\mathbf{W}$ is available, a 400-dimensional i-vector can be extracted from every call i, j, by making a MAP-estimate of $\mathbf{z}_{ij}$. Next, we treat the i-vector as a new 'feature', while ignoring the generative model (2) that helped us to extract it.

Now we build a *new* generative model, again with *within* and *between*-speaker variability. The difference is that we can now model the i-vectors in 400-dimensional space, instead of in the original 122880-dimensional supervector space, so that factor analysis is no longer necessary. We used 400-dimensional multivariate normal distributions for these i-vector variabilities, where the two (full) covariance matrices of these distributions are denoted $\mathbf{C}$ and $\mathbf{D}$. These are estimated with another EM-algorithm on data as indicated in Table 1.

Here we made a similar decision and did not use the EVALITA development data to estimate $\mathbf{C}$ and $\mathbf{D}$.

5.1 S-norm

The I-Vector system is gender-dependent, that is, we use two different systems, trained on male and female data, to respectively process male or female verification trials. The score normalization is also gender-dependent.

In contrast to JFA, the I-Vector system has symmetric scores in the sense that score(train,test) = score(test,train). We found a symmetric normalization, denoted *S-norm*, works better than the asymmetric ZT-norm. S-norm uses a single cohort per gender rather than separate Z and T cohorts. Normalization for a trial, with a given test segment and model, works as follows:

1. Score the test segment against the whole cohort and retain the mean μ_1 and standard deviation σ_1 of these scores.
2. Score the model against the whole cohort and retain the mean μ_2 and standard deviation σ_2 of these scores.
3. raw score = score(test segment,model)
4. Finally:

$$\text{normalized score} = \frac{\text{raw score} - \mu_1}{\sigma_1} + \frac{\text{raw score} - \mu_2}{\sigma_2}$$

The S-norm cohorts were also populated from the EVALITA development data, but here we used each call separately rather than pooled, so there were $20 \times 30 = 600$ members in each gender-dependent cohort.

6 Fusion and calibration

For fusion [15] and calibration [14] of our two sub-systems, we used the logistic regression optimizer of the FoCal Toolkit [13] to perform discriminative optimization of an empirical cross-entropy criterion.

The data used for fusion and calibration training was a large number of same-gender verification trials constructed from the 20×60 calls of the EVALITA development data. We used all possible models of the 6 types prescribed by the evaluation plan [1] and all possible test-segments of type 'TS1' that we could extract from this data. We found in development experiments on independent TS1 and TS2 tests, that training on the short duration TS1 segments gave better results than training on the long duration TS2 or on TS1+TS2.

7 Results and conclusion

In Figure 1 we borrow from the companion paper [11] the DET-curve analysis of the AGNITIO system. We show 24 DET-curves for different subsets of trials, conditioned on test-segment duration, training mode and telephone type. In our work for this evaluation, we were most interested in the *cross-channel* conditions, where enrollment and test were done over different telephone types. These conditions, P-G, G-P, 3G-P and 3P-G are represented with thick curves in the figure, where indeed it shows that these conditions are more challenging.

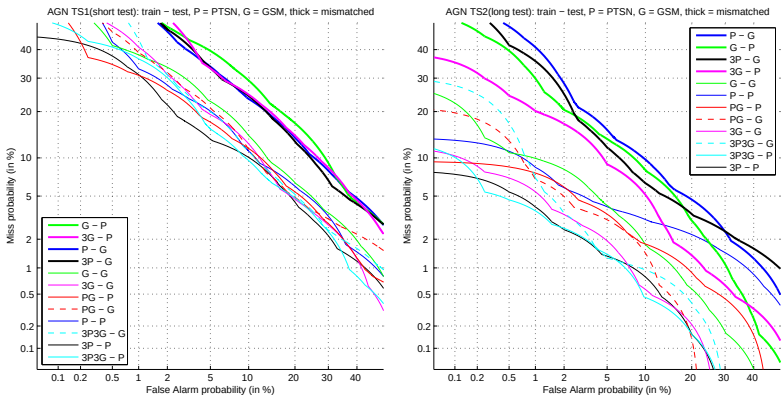

Fig. 1. ROCCH-DET analysis for AGN primary

Apart from the predictable trends as discussed in [11], the most notable feature of this evaluation was the unexpected level of difficulty of the cross-channel trials. Yes, they are more difficult, but from our experience with the cross-channel conditions in the Mixer databases of the NIST evaluations, we had expected better accuracies. Again, as mentioned above in section 4 and as further discussed in [11], all the other systems also found these conditions challenging, even though a considerable amount of apparently matched development data was provided.

In future work we would like to analyze the difficulty of this cross-channel data and explore what may be done to improve it.

References

1. Aversano, G.: EVALITA 2009 - Speaker Identity Verification - Application Track Task Guidelines, http://evalita.fbk.eu/speaker.html
2. Burget, L., Brümmer, N., et al.: Robust Speaker Recognition Over Varying Channels-Report from JHU workshop 2008, http://www.clsp.jhu.edu/workshops/ws08/documents/jhu_report_main.pdf
3. Burget, L., Matejka, P., Hubeika, V. and Cernocky, J.: Investigation into variants of Joint Factor Analysis for speaker recognition. In: Proceedings of Interspeech 2009. Brighton (2009)
4. Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P.: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, issue 4, pp. 1435–1447 (2007)
5. Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P.: A Study of Inter-Speaker Variability in Speaker Verification. IEEE Transactions on Audio, Speech and Language Processing, vol. 16, issue 5, pp. 980-988 (2008)
6. Kenny, P., Dehak, N., Dehak, R., Gupta, V., Dumouchel P.: The Role of Speaker Factors in the NIST Extended Data Task. In: Proceedings of Odyssey 2008: The Speaker and Language Recognition Workshop. Stellenbosch, South Africa (2008)
7. Pelecanos, J., Sridharan, S.: Feature Warping for Robust Speaker Verification. In: Proceedings of A Speaker Odyssey, The Speaker Recognition Workshop (2001)
8. Reynolds, D., Quatieri, T., Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, vol. 10, pp. 19–41 (2000)
9. Martin, A.F., Greenberg, C.S.: NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels. In: Proceedings of Interspeech 2009. Brighton (2009)
10. Hidden Markov Model Toolkit (HTK), http://htk.eng.cam.ac.uk
11. Aversano, G., Brümmer, N., Falcone, M.: EVALITA 2009 Speaker Identification Verification Application Track—Organizer's Report. In: Proceedings of EVALITA 2009. Reggio Emilia, Italy (2009)
12. Dehak, N., et al.: Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In: Proceedings of Interspeech 2009. Brighton (2009)
13. FoCal: Toolkit for Evaluation, Fusion and Calibration of Statistical Pattern Recognizers, http://focaltoolkit.googlepages.com
14. Brümmer, N., Du Preez, J.: Application Independent Evaluation of Speaker Detection. Computer Speech and Language, vol. 20, pp. 230–275 (2006)

15. Brümmer, N., Burget, L., et al.: Fusion of Heterogenous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, issue 7 (2007)