I3A NIST SRE2010 System Description

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

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

Abstract. I3A has submitted several systems to NIST SRE2010 for the core-core, core-10sec, 10sec-10sec and core-summed conditions. The systems are different variants of the GMM-UBM with Joint Factor Analysis approach: JFA-LLR, JFA-SVM, iVectors and the fusion of them. All the systems share the same feature extraction and UBM. The speaker segmentation for the summed condition is performed using a speaker factors based system. We show the results we have got in the evaluation together with some post evaluation experiments focusing on the influence of the development data on the calibration.

Keywords: speaker recognition, NIST, SRE 2010, Joint Factor Analysis (JFA), iVectors, support vector machine (SVM), score calibration

1 Introduction

Every two years NIST conducts a evaluation of speaker recognition systems [1] in order to compare the different state of the art approaches on a commom corpus. Like in previos years, the 2010 task is text independent speaker detection, that is, decide if a given speaker is speaking in a given test segment.

This year, like in 2008, we have data from phonecalls recorded by a telephone channel or by a room microphone channel and from an interview scenario recorded by a room microphone channel. What is new this year is that some of the speech has been collected in a way intended to produce high or low vocal effort speech by the target speaker. Besides, this year the interview segments can have different durations. Finally, there is a new operating point for the core-core conditions that intends to reduce the number of false alarms as much as possible, as we will see later. NIST has made available an optional extended trial list in order to get a robust performance measure in this new operating point.

We have submitted systems based on the well known Joint Factor Analysis Approach [2]. These share the same feature extraction and UBM. The systems submitted for each of the conditions we participate are:

- core-core/coreext-coreext:
  1. Fusion of JFA-SVM and JFA-LLR, both Gender Dependent (GD)
  2. JFA-SVM GD

   3. JFA-LLR GD
- core-10sec/10sec-10sec:
   1. Fusion of JFA-LLR and iVectors, both GD
   2. JFA-LLR GD
- core-summed:
   1. JFA-LLR Gender Independent (GI)

The speaker segmentation for the summed condition is performed using a speaker factors based system.

This paper is organized as follows. Section 2 explains the feature extraction process. Section 3 describes the speaker segmentation algorithm. Sections 4 to 6 describes the classification systems used and the development data used to train the JFA matrices. Section 7 focuses on the calibration procedure. In section 8, we present the results we have got in the evaluation. Finally, in section 9, we draw some conclusions.

2 Feature Extraction

The front-end extracts feature vectors of 20 MFCC including C0 (C0-C19) over a 25 ms hamming window every 10 ms (15 ms overlap), and first and second order derivatives are computed over the feature vector sequence.

Voice Activity Detection (VAD) is performed computing the Long-Term Spectral Divergence (LTSD) of the signal every 10 ms, and comparing it against a threshold as in [3]. For phone calls, where two channels are available, namely channel of interest and reference channel, the reference channel is used for crosstalk removal. For interview segments, NIST provided ASR labels are used for removing the interviewer.

After frame selection, features are short time Gaussianized with a 3 seconds window as in [4].

3 Speaker Segmentation

For the core-summed condition we used a segmentation system to generate two speaker dependent feature vector streams for every test segment.

To perform speaker segmentation, first, we compute a stream of speaker factors of dimension 20 for the given recording. These factors are computed using 12 MFCC with no derivatives (C1-C12) and a GMM of 256 Gaussians. Then, we model the stream with two Gaussians and, considering that every Gaussian belongs to a single speaker, we segment the stream using Viterbi decoding. The system is very similar to that proposed in [5], but we perform the algorithm on the whole segment, rather than in one minute segments. After this first segmentation, we apply two Viterbi re-segmentations using 12 MFCC as features, and GMM as speaker models, using soft-clustering in our second re-segmentation [6].

Segmentation is done over speech frames only. Then, the feature vector stream is separated into two different streams (one per speaker). After that, every stream is Gaussianized separately.

4 JFA-LLR System

This system is a simplified version of [2] with the following configuration.

4.1 Universal Background Models

Gender Dependent (GD) and Gender Independent (GI) Universal Background Models (UBM) of 2048 Gaussians are trained by EM iterations. For this purpose, we have used all the telephone signals in SRE2004, SRE2005 and SRE2006 databases (649 male speakers with 7412 signals and 801 female speakers with 9889 signals).

4.2 JFA Training

JFA Hyperparameters are trained from the previous background models. 300 eigenvoices (v) and 100 telephone eigenchannels (u_{phn}) are trained using telephone data from all the speakers of SRE2004, SRE2005 and SRE2006 databases having, at least, 8 recordings by speaker (530 male speakers with 7398 signals and 731 female speakers with 9938 signals). Another 100 eigenchannels (u_{mic}) are trained using all signals from speakers having far field microphone data in SRE2005 and SRE2006 and 50 speakers (kept out speakers) with interview data from SRE2008 (106 male speakers with 6244 signals and 119 female speakers with 6919 signals). Both eigenchannel matrices are stacked together for the core-core condition. For the other conditions, only the telephone eigenchannel matrix is used. Finally, the remaining speaker variability matrix (d) is trained from the speakers of SRE2004, SRE2005 and SRE2006 having least than 8 recordings by speaker (201 male speakers with 547 signals and 152 female speakers with 668 signals). MAP estimates of speaker and channel factors are fixed for estimating d matrix to speed up the system. The d matrix is used in all the conditions but the 10sec-10sec. All the matrices are trained by EM ML+MD iterations.

4.3 Speaker Enrollment and Scoring

Speakers are enrolled into the system using MAP estimates of their speaker and remaining variability factors (y,z) so the speaker means super vector is given by:

$$M_s = m_{UBM} + vy + dz \ . \tag{1}$$

Trial scoring is performed using first order Taylor approximation of the LLR between the target and the UBM Models like in [7].

$$LLR \approx (vy_{trn} + dz_{trn})^t \Sigma^{-1} (F_{tst} - N_{tst} u x_{tst}) \ . \tag{2}$$

ZTNorm score normalization is applied using telephone data from SRE2004, SRE2005 and SRE2006 (628 male speakers and 858 female speakers with 4 segments by speaker). For the summed condition the maximum score of the two automatic segmented speakers is chosen.

5 JFA-SVM System

5.1 JFA Training

This system shares the same UBM and JFA matrices from the previous one.

5.2 SVM Scoring

Speaker enrollment and scoring is done by an SVM with the following kernel:

$$k(x_1, x_2) = (vy_1 + dz_1)^t \Sigma^{-1}(F_2 - N_2 u x_2) + (vy_2 + dz_2)^t \Sigma^{-1}(F_1 - N_1 u x_1) \ . \quad (3)$$

Background signals for SVM training are chosen from SRE2004, SRE2005, SRE2006 and the kept out speakers from SRE2008. For telephone background segments we have 653 male speakers and 883 female speakers with 4 signals by speaker. For microphone background segments we use 119 males and 106 female speakers with 2 signals by speaker and type of microphone. ZTNorm is applied to the SVM score using the same SVM background segments. For SVM training we use libsvm [8].

6 iVectors System

6.1 Total Variability Space Training

We follow the approach taken in [9]. Total variability space of dimension 400 is trained using the same data as for JFA eigenvoices matrix by ML+MD iterations. Linear Discriminant Analysis (LDA) and Within Class Covariance Normalization (WCCN) are applied to the total variability factors (iVectors) for inter-session compensation reducing the iVector dimension to 200. For this purpose, we use the same data as for the telephone eigenchannels matrix.

6.2 Scoring

Speakers are enrolled into the system using MAP estimates of their channel compensated iVectors. Trials are evaluated by cosine distance between training and testing iVectors. We apply ZTNorm using the same segments as in the JFA-LLR system.

7 Calibration and Fusion

The systems have been calibrated to minimize the NIST performance measure given by the normalized cost function:

$$C_{Det} = C_{Miss} P_{Miss|Target} P_{Target} + C_{FA} P_{FA|NonTarget}(1 - P_{Target})$$

$$C_{Norm} = C_{Det}/C_{Default} \quad C_{Default} = min\{C_{Miss} P_{Target}, C_{FA}(1 - P_{Target})\} \ . \quad (4)$$

In 2010 a new set of cost model parameters has been defined for the core-core conditions. The old values, used in previous evaluations, are kept for the rest of the conditions as shown in Table 1.

Table 1. Cost Function Parameters

	C_{Miss}	C_{FA}	P_{Target}
core-core/coreext-coreext	1	1	0.001
non core	10	1	0.01

7.1 Core-core

Calibration The system output scores are calibrated to mean log-likelihood ratios by linear logistic regression using FoCal package [10]. For the new NIST cost function false alarms count 1000 times more that misses. This means that, in order to have our system well calibrated we need to get a very low number of false alarms. However, according to the Doddington 's "rule of 30" [11], to be 90% confident that the true error rate is ±30% of the true error rate there needs to be at least 30 errors. So, to achieve robust calibration, we need a good estimation of the false alarm error rates and, therefore, a development trial list such as our system produces a fair amount of false alarms for the new operating point. For that purpose, we need a list with millions of non target trials. We have built a trial list using data from SRE2008 including most of the common conditions included in SRE2010 evaluation. This list includes all trials that can be done using all training short, long and follow-up versus all testing short, long and follow-up English SRE2008 segments. We keep out the 50 speakers used in JFA training. In total, we have around 4M male trials and 10M female trials.

Besides, given that false false alarms (targets miss-labeled as non targets) have great influence on the cost, the labeling of the trial list needs to be accurate. On the contrary, we would overestimate the cost and miss-calculate the calibration parameters.

We do multiple condition calibration iteratively until convergence:

1. Gender dependent (male, female)
2. Channel dependent (mic-mic same channel, mic-mic different channel, mic-phn, phn-phn)
3. Length dependent (short-short, long-long, long-short, short-long)

Same-Different Microphone detection We have observed that the same microphone trials present an upwards shifted score distribution compared to the different microphone trials. That makes necessary to calibrate this two kind of trials separately. As far as the microphone information is not provided by NIST, we do automatic same/different microphone detection. For that, we use a iVectors like system. We take the microphone speaker factors corresponding to the u_{mic} matrix as features. We train LDA and WCCN using SRE2008 data from the kept out speakers taking the 12 more discriminative directions. Cosine distance is used for scoring. Scores are normalized using SNorm [12] with microphone segments, again, from the 50 kept out speakers.

The channel detection score is calibrated using FoCal with a same channel prior probability of 0.1. This way, for each mic-mic trials we get a soft probability

of being same or different microphone trial. We use this probability together with FoCal bilinear on the speaker detection calibration.

Fusion The calibrated systems are channel dependent fused using again FoCal package.

7.2 Non core-core

For the non core-core conditions we do gender dependent calibration and fusion using the det7 matching conditions of SRE2008 as development lists.

8 Results

8.1 Core-core

In Figure 1 we show the DET curves for the three systems submitted in the core-core and coreext-coreext conditions. We can see all the systems perform quite similarly having I3A_1 slightly better curves. The system is very robust across the common conditions [1], producing similar curves in most of them whether they include telephone or interview speech. We only notice a clear degradation in the high vocal effort conditions (det6, det7). On the other side, low vocal effort speech (det8, det9) seems not to affect performance getting even better results than the normal vocal effort conditions.

In Table 2 we show the actual and minimum costs for the core-core conditions using the new cost model parameters. There were some labeling errors in the former SRE2008 key provided by NIST. After submitting the SRE2010 results NIST made available some labeling corrections. We compare the results that we get calibrating the systems using the original key and the corrected one on the development. For most of the conditions, but det2, we get better actual costs using the corrected key. This is especially striking for the interview same microphone (det1) and the phonecall microphone (det4, det7, det9) conditions whose improvement allows getting actual costs lower than 1.0.

Comparing the three systems submitted, it is not clear which of them produce better minimum costs. However, I3A_3 seems to have better actual costs. In fact, it is the only one that produces a cost lower than 1.0 on the det1 condition. Therefore, we think it is easier to get a robust calibration across datasets using I3A_3.

We want to explore, too, the impact of the development trial list size on the calibration. For that, we have carried out another calibration using a development list of around 100K male trials and 160K female trials only. The results for I3A_3 system are shown in Table 3. The short list performs better than expected getting reasonably good calibrations in the microphone conditions. However, actual costs in the telephone conditions (det5, det6, det8) are clearly better using the long list.

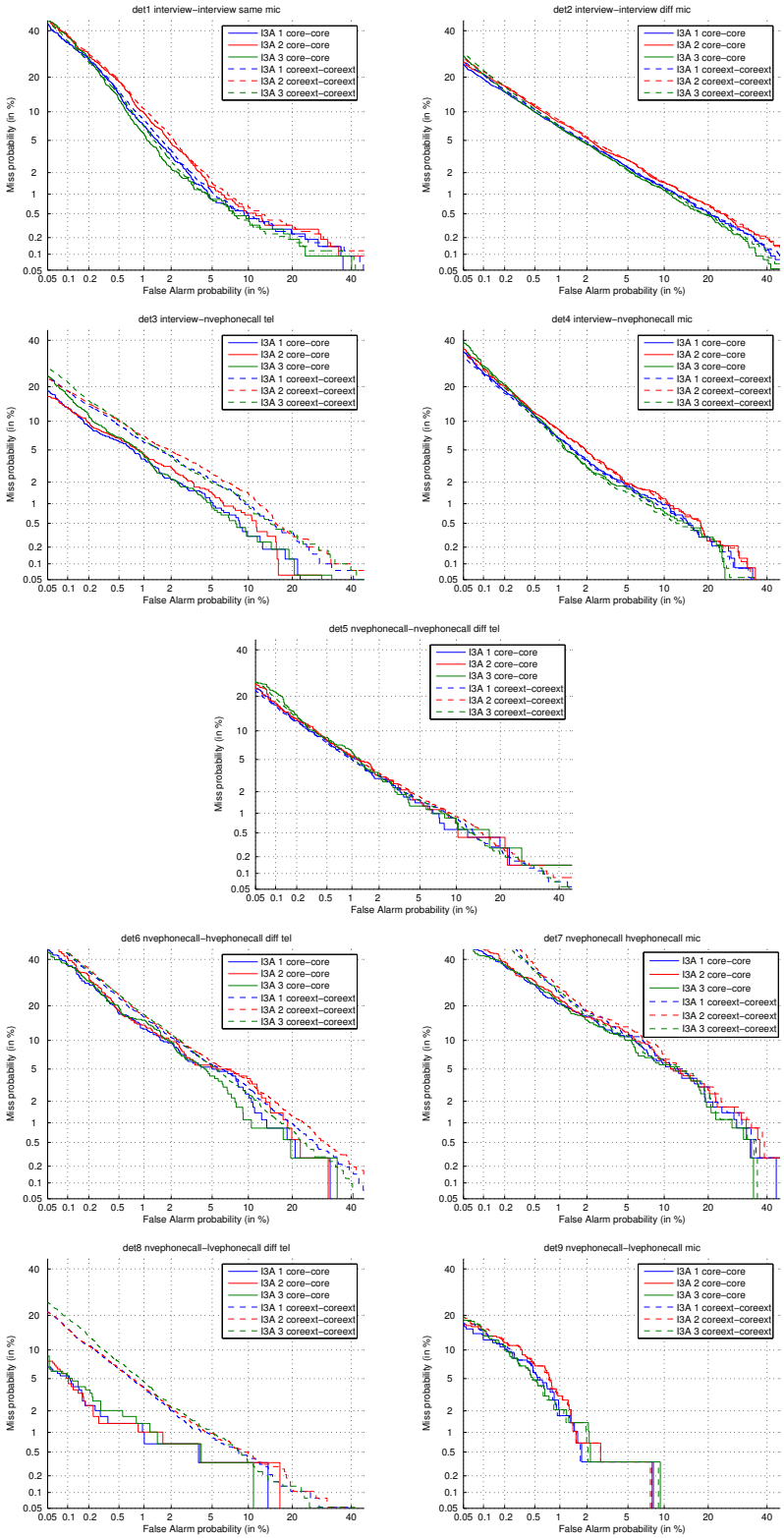

Fig. 1. DET curves for core-core and coreext-coreext conditions

Table 2. actCost/minCost for core-core conditions

actCost/minCost	det1	det2	det3	det4	det5	det6	det7	det8	det9
I3A 1	2.21/**0.70**	**0.49/0.48**	0.61/0.41	1.21/0.62	0.55/0.40	0.84/0.77	0.99/**0.69**	**0.19/0.17**	1.06/0.40
I3A 2	3.12/0.72	0.51/0.50	0.97/**0.35**	1.79/**0.61**	0.81/0.38	0.85/0.80	1.52/0.70	**0.19**/0.18	1.24/0.39
I3A 3	1.18/0.75	0.58/0.57	0.58/0.49	0.82/0.67	0.54/0.50	**0.81**/0.73	0.81/0.75	0.31/0.22	0.73/0.44
I3A 1 recal	1.11/0.75	0.56/0.49	0.51/0.40	0.94/0.65	0.73/**0.35**	0.87/0.81	**0.75**/0.72	0.24/**0.17**	0.59/0.41
I3A 2 recal	1.22/0.73	0.54/0.50	**0.50**/0.40	0.99/0.64	0.73/**0.35**	0.87/0.81	**0.75**/0.71	0.24/**0.17**	0.60/**0.38**
I3A 3 recal	**0.74**/0.71	0.69/0.58	0.59/0.58	**0.70**/0.63	**0.53**/0.49	**0.81/0.72**	**0.75**/0.70	0.33/0.24	**0.43**/0.39

Table 3. actCost/minCost for core-core conditions using different length of trial lists

actCost/minCost	det1	det2	det3	det4	det5	det6	det7	det8	det9
I3A 3 short	0.98/0.73	**0.61**/0.59	**0.59/0.57**	**0.67**/0.65	0.71/**0.45**	0.97/**0.72**	**0.75**/0.71	0.64/0.25	0.51/0.43
I3A 3 long	**0.74/0.71**	0.69/**0.58**	**0.59**/0.58	0.70/**0.63**	**0.53**/0.49	**0.81/0.72**	**0.75/0.70**	**0.33/0.24**	**0.43/0.39**

This year, NIST made available an optional extended list of trials for the core-core condition. Table 4 show the costs of our systems calibrated with the original and corrected SRE2008 key. This list is intended to allow more accurate cost estimations according to the above mentioned "rule of 30" [11]. In spite of that, the costs are quite similar to those of the non extended list. Only det7 and det8 are clearly different. Nevertheless, the shape of the det7 coreext curves, fast rising at the left, suggests that can be same problem with the high vocal effort microphone phonecalls or the SRE2010 key.

Table 4. actCost/minCost for coreext-coreext conditions

actCost/minCost	det1	det2	det3	det4	det5	det6	det7	det8	det9
I3A 1	2.06/**0.69**	**0.51/0.51**	0.66/0.48	1.16/**0.60**	0.54/0.48	0.93/0.90	2.87/**0.99**	0.48/0.46	1.22/**0.40**
I3A 2	3.22/0.72	0.53/**0.51**	0.97/**0.45**	1.75/0.61	0.70/0.49	0.97/0.91	3.73/0.99	0.47/0.45	1.43/0.43
I3A 3	1.08/0.73	0.58/0.57	0.63/0.56	0.74/0.66	0.56/0.52	0.94/0.91	1.88/1.00	0.60/0.53	0.82/0.46
I3A 1 recal	0.97/0.73	0.56/0.53	**0.59**/0.49	0.87/0.64	0.63/**0.47**	0.96/0.91	2.15/**0.99**	**0.45/0.44**	0.73/0.47
I3A 2 recal	1.07/0.71	0.55/0.53	0.60/0.49	0.94/0.62	0.63/**0.47**	0.96/0.91	2.28/**0.99**	**0.45/0.44**	0.76/0.44
I3A 3 recal	**0.71**/0.71	0.69/0.58	0.61/0.59	**0.67**/0.64	**0.53**/0.51	**0.88/0.88**	**1.36/0.99**	0.52/0.48	**0.47**/0.44

8.2 Non core-core

Figure 2 shows the DET curves for the core-summed, core-10sec and 10sec-10sec conditions. Tables 5 and 6 show their costs, that are nicely calibrated. If we compare these results with the official SRE2008 det7 results [13] we realize that our systems greatly outperform the bests of them, what proves the advances done in the last two years. It is worth noticing the det5 core-summed DET is very good compared to core-core one. That is, mainly, due to our speaker segmentation system whose performance is comparable to the best published systems.

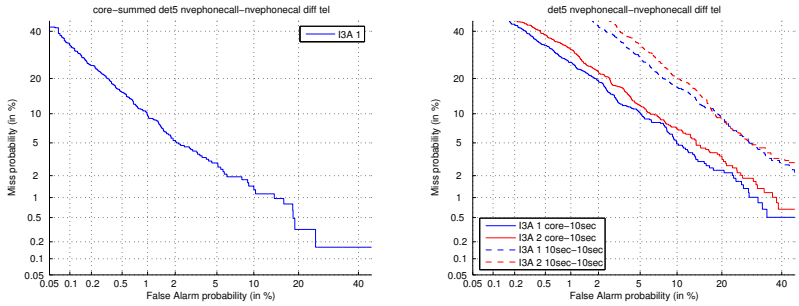

Fig. 2. DET curves for core-summed, core-10sec and 10sec-10sec conditions

Table 5. actCost/minCost for core-summed condition

actCost/minCost	det5
I3A 1	0.20/0.19

9 Conclusions

We have presented state of the art systems to the NIST SRE2010. Our systems present good and robust performance across conditions. Furthermore, our summed and 10sec systems scored among the very best in the evaluation. Systems used in the core-core condition benefits very little from fusion and JFA-LLR presents a better calibration in most of the conditions. On the other side, 10sec systems get an interesting improvement from fusion. We would like to point out that our speaker segmentation system allows us to achieve core-summed results quite near to the ones of the core-core condition.

We have compared the influence of using different development trial lists for system calibration in the new operating point. Results show the importance of using a list with a huge number of non target trials and free of labeling errors. Score shift between same/different microphone conditions makes calibration difficult, which makes necessary to use a channel detector. In the future, we will work to make our systems more microphone independent and to improve our channel detector.

Regarding to the new vocal effort speech present this year, performance markedly degrades for the high vocal effort conditions but is unaffected for the low vocal effort ones.

Table 6. actCost/minCost for core-10sec condition

actCost/minCost	det5 core-10sec	det5 10sec-10sec
I3A 1	**0.36/0.35**	**0.64/0.61**
I3A 2	0.40/0.39	0.69/0.66

10 Acknowledgements

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

References

1. http://www.itl.nist.gov/iad/mig/tests/sre/2010/NIST_SRE10_evalplan.r6.pdf.
2. Patrick Kenny. Joint factor analysis of speaker and session variability : Theory and algorithms - Technical report CRIM-06/08-13, 2005.
3. J. Ramirez, J.C. Segura, C. Benitez, A. de La Torre, and A. Rubio. Voice activity detection with noise reduction and long-term spectral divergence estimation. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages ii–1093–6. IEEE.
4. Jason Pelecanos and Sridha Sridharan. Feature warping for robust speaker verification. In *Oddyssey Speaker and Language Recognition Workshop*, Crete, Greece, 2001.
5. Fabio Castaldo, Daniele Colibro, Emanuele Dalmasso, Pietro Laface, and Claudio Vair. Stream-based speaker segmentation using speaker factors and eigenvoices. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4133–4136, Las Vegas, Nevada, March 2008. IEEE.
6. Doug Reynolds, Patrick Kenny, and Fabio Castaldo. A Study of New Approaches to Speaker Diarization. In *Interspeech 2009*, Brighton, UK, 2009.
7. Ondrej Glembek, Lukas Burget, Najim Dehak, Niko Brummer, and Patrick Kenny. Comparison of scoring methods used in speaker recognition with Joint Factor Analysis. In *ICASSP '09: Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4057–4060, Washington, DC, USA, 2009. IEEE Computer Society.
8. Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001.
9. Najim Dehak, Redah Dehak, Patrick Kenny, Niko Brummer, Pierre Oullet, and Pierre Dumouchel. Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification. In *Interspeech 2009*, Brighton, UK, 2009.
10. Niko Brummer. http://sites.google.com/site/nikobrummer/focalbilinear.
11. G Doddington. The NIST speaker recognition evaluation Overview, methodology, systems, results, perspective. *Speech Communication*, 31(2-3):225–254, June 2000.
12. Niko Brummer and Albert Strasheim. AGNITIO's Speaker Recognition System for EVALITA 2009. In *Poster and Workshop Proceedings of the 11th Conference of the Italian Association for Artificial Intelligence*, Reggio Emilia, Italy, 2009.
13. http://www.itl.nist.gov/iad/mig/tests/sre/2008/official_results/index.html.