I3A NIST SRE2010 System Description

Jesús Villalba, Carlos Vaquero, Eduardo Lleida, Alfonso Ortega, Antonio Miguel
Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{villalba,cvaquero,lleida,ortega,amiguel}@unizar.es

Introduction

I3A has submitted several systems to NIST SRE2010:

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

Feature Extraction

- 19 MFCC + C0 + Δ + $\Delta\Delta$
- VAD comparing the Long-Term Spectral Divergence (LTSD) against a threshold.
  - Crosstalk removal in phonecall using non target channel
  - Interviewer removal using ASR labels
- Short Time Gaussianization with a 3 seconds window.

Summed Channel Speaker Segmentation

- First Viterbi segmentation using 20 speaker factors, calculated using 12 MFCC and 256 Gaussians, as features and modeling each speaker with a Gaussian.
- Two Viterbi re-segmentations using 12 MFCC as features and modeling each speaker with a GMM.
- Softclustering in the second re-segmentation.
- Every output stream is short time Gaussianized separately.

Classification

- 2048 GMM *UBM* trained on SRE04, SRE05, SRE06 telephone data.
- *JFA* Hyperparameters:
  - v (300 speaker factors), u_{phn} (100 channel factors) and d from tel data in SRE04, SRE05 and SRE06
  - u_{mic} (100 channel factors) from mic data in SRE05, SRE06 and 50 speakers from SRE08.
- *JFA-LLR*
  - Speaker mean supervector given by the speaker factors (y,z) MAP estimates:

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

  - Scoring is performed using first order Taylor approximation of the LLR:

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

  - GD ZTNorm using telephone segments from SRE04,SRE05 and SRE06 (2363 male and 3318 female)
- *JFA-SVM*
  - Enrollment and scoring is done by an SVM with the following kernel:

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

  - GD SVM Background and ZTNorm using tel and mic segments from SRE04, SRE05, SRE06 and 50 speakers from SRE08 (4241 male and 5400 female)
- *iVectors*
  - T (400 total variability factors), LDA (200 dimensions) and WCCN trained from telephone data in SRE04,SRE05 and SRE06.
  - Scoring is done by Cosine distance between enrollment and testing iVectors
  - GD ZTNorm using telephone segments from SRE04,SRE05 and SRE06 (2363 male and 3318 female)

Calibration and Fusion

Core-core

- Calibration
  - Dev trial list including most of the common conditions included in SRE2010.
  - All training vs. all testing short, long and follow-up SRE08 English data keeping out the 50 speakers used in JFA training (4M male trials and 10M female trials).
  - Calibration has been done by linear logistic regression using FoCal package.
  - 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
  - iVectors like system using mic channel factors as features.
  - We train LDA (12 dimensions) and WCCN using SRE08 mic data from the 50 keep out speakers.
  - The scores are normalized using SNorm and calibrated to provide a soft decision probability.
- Fusion: The calibrated systems are channel dependent fused.

Non core-core

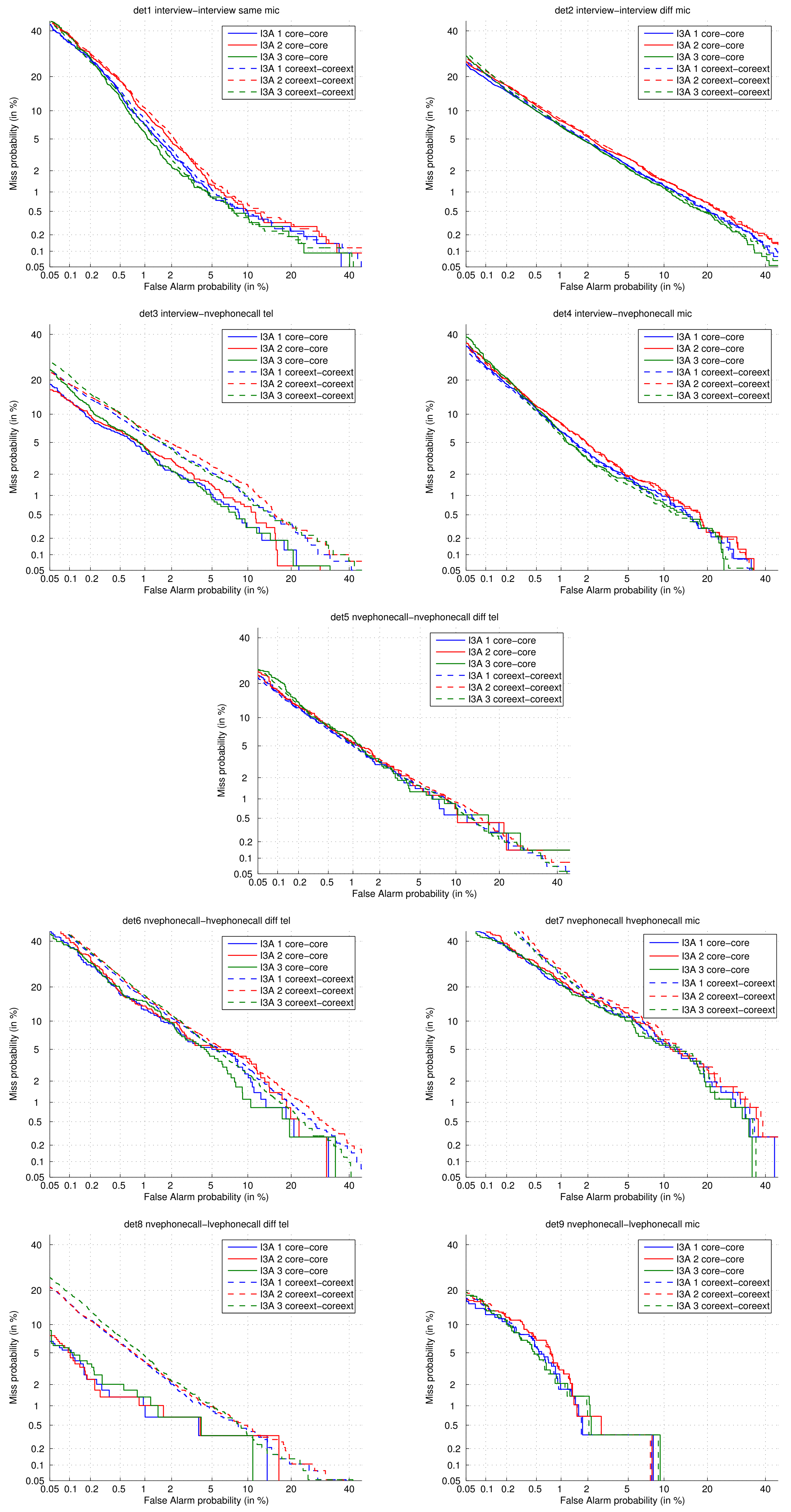
- Gender dependent calibration and fusion using the det7 matching conditions of SRE08.

Acknowledgments

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

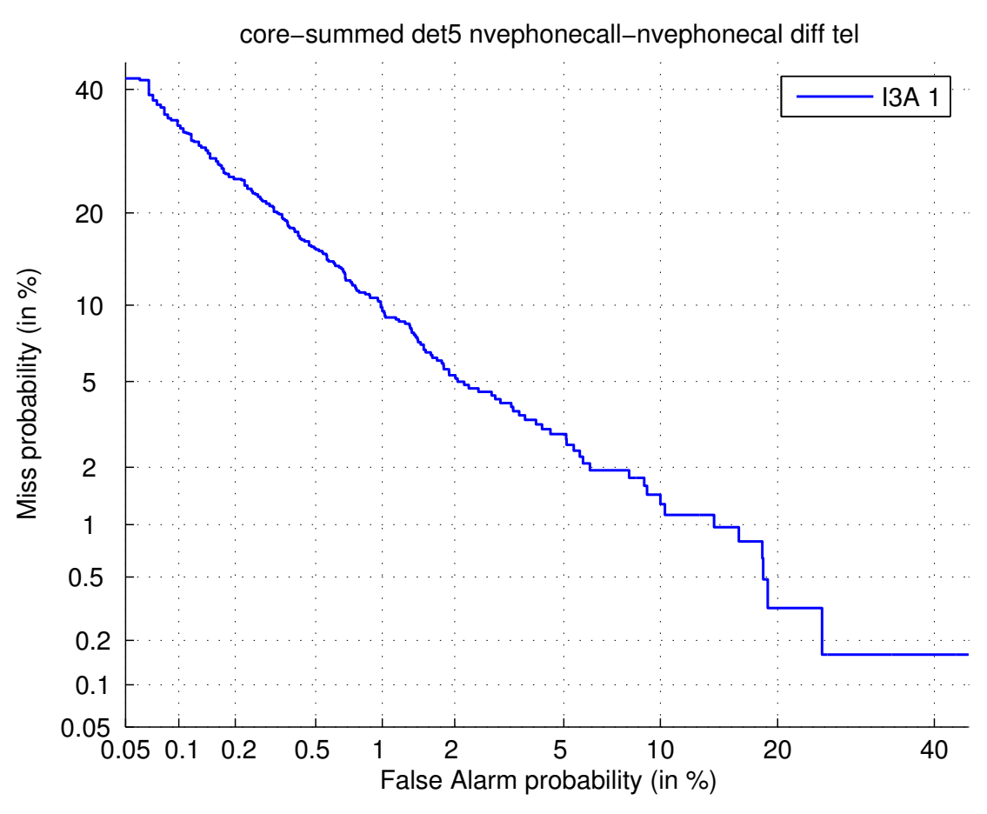
Results

core-core/coreext-coreext



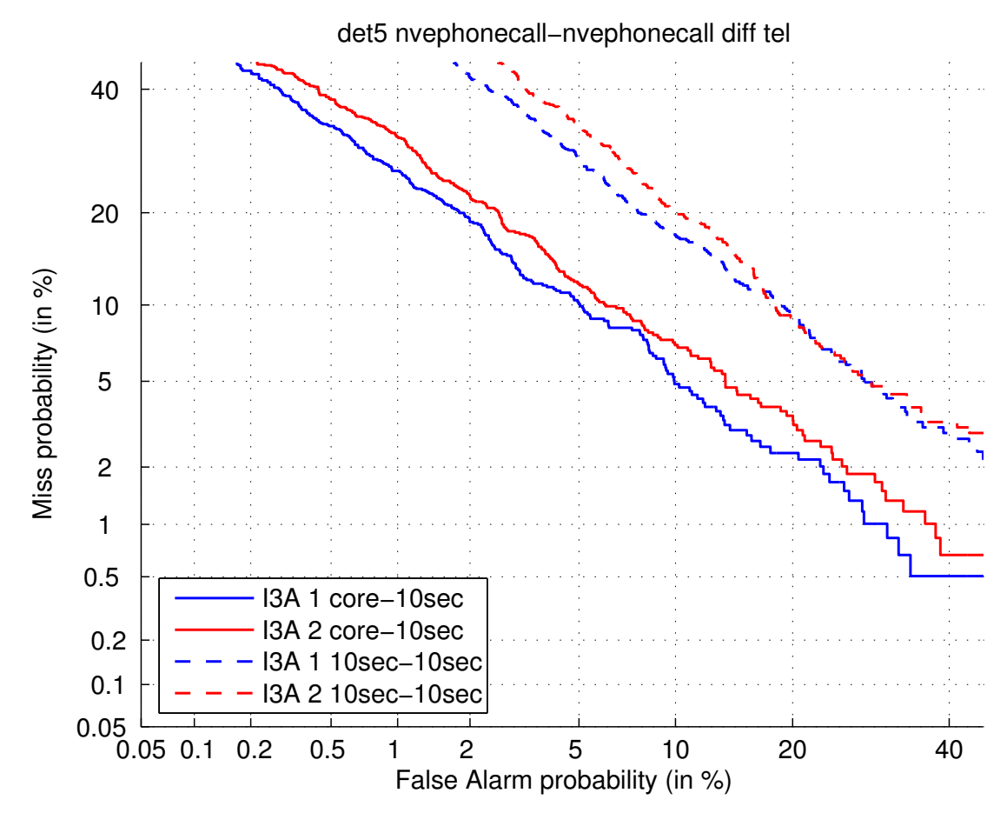
actCost/minCost	det1	det2	det3	det4	det5	det6	det7	det8	det9
I3A 1 core-core	2.21/**0.70**	**0.49/0.48**	0.61/0.41	1.21/0.62	0.55/0.40	0.84/0.77	0.99/**0.69**	**0.19/0.17**	1.06/0.40
I3A 2 core-core	3.12/0.72	0.51/0.50	0.97/**0.35**	1.79/**0.61**	0.81/0.38	0.85/0.80	1.52/0.70	**0.19**/0.18	1.24/0.39
I3A 3 core-core	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 2 core-core 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 core-core 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
I3A 1 coreext-coreext	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 coreext-coreext	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 coreext-coreext	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

core-summed



actCost/minCost	det5
I3A 1	0.20/0.19

core-10sec/10sec-10sec



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

Conclusions

- We have built state of the art systems with good performance in most of the conditions
- Especially good results for summed channel and 10sec conditions.
- SVM system has better minDCF than LLR but it is more difficult to get a robust calibration.
- Need of big number of non target trials without key errors for robust calibration in the new operating point.
  - Improvement in actCost recalibrating after correcting SRE08 key errors.
- Score shift between same/diff mic conditions difficults calibration ⇒ Need channel detector.
- High vocal effort speech degrades performance considerably.