The L^2F - UPC Speaker Recognition System for NIST SRE 2010

Alberto Abad
L^2F, INESC-ID/IST

Jordi Luque
L^2F, INESC-ID/IST
TALP/UPC
Research Center

Javier Hernando
TALP/UPC
Research Center

Isabel Trancoso
L^2F, INESC-ID/IST

alberto.abad@l2f.inesc-id.pt

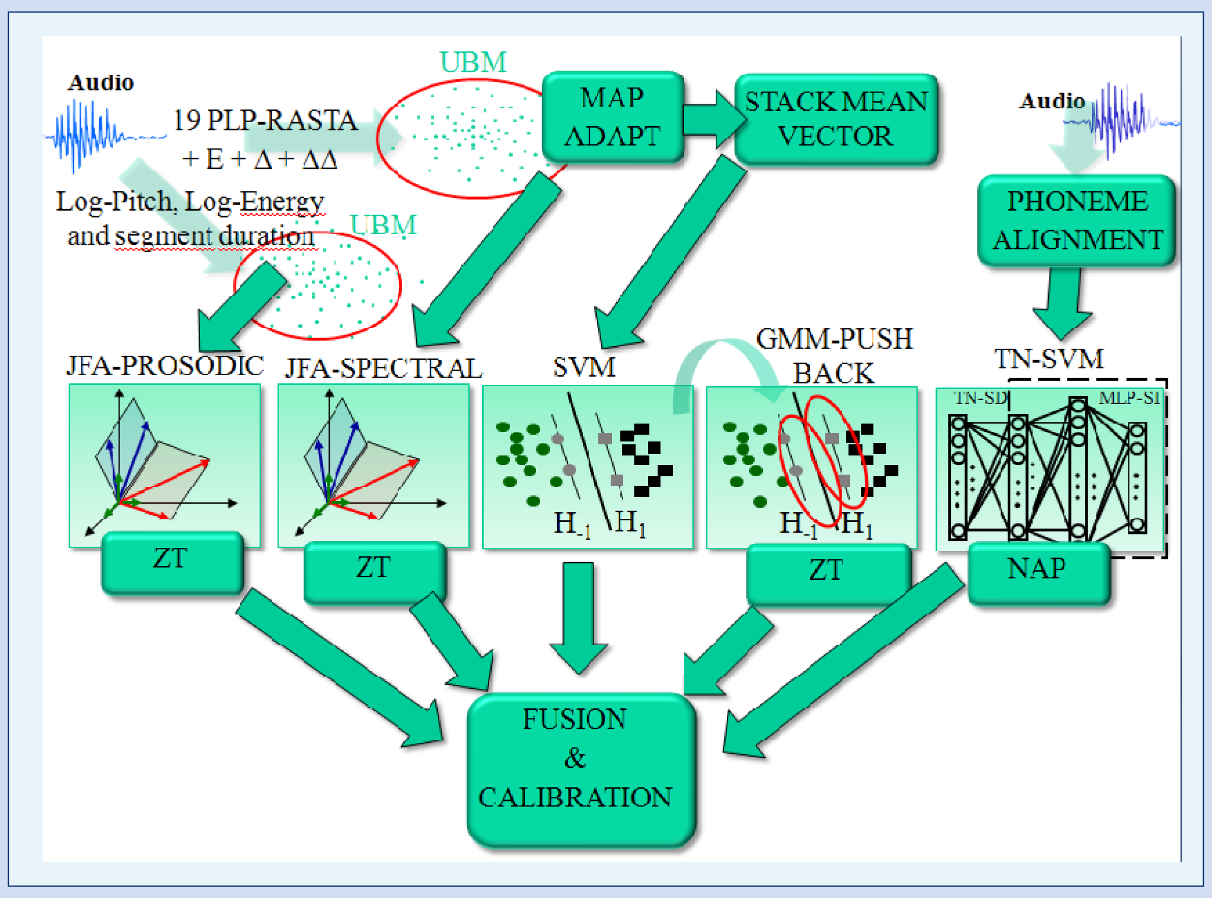
jorge.luque@upc.edu

SRE'10 System Description

Three systems were submitted to core-core condition

→ The **primary system**, fusion of 5 individual:

- **(I)** JFA-SPECTRAL based on PLP with log-RASTA processing
- **(II)** JFA-PROSODIC uses prosodic features
- **(III)** GSV-SVM is the standard supervector approach
- **(IV)** GSV-GMM is the pushing-back version of GSV-SVM (III)
- **(V)** TN-SVM a **novel** system based on features obtained from MLP speaker adaptation



→ Two **alternative systems**

- Fusion of JFA sub-systems: JFA-SPECTRAL (I) + JFA-PROSODIC (II)
- Fusion of sub-systems not using NIST transcription: (I)+(II)+(III)+(IV)

1

Resources

Corpora

- NIST SRE 2004, 2005, 2006 for training
  - → Speaker dependent Universal Background Models (UBM)
  - → ZT normalization
  - → Background Impostor modeling for SVM based systems
  - → Compensation techniques NAP and JFA training
- HUB-4 speech (140h)
  - → Training MLP acoustic models
- NIST SRE 2008 short2-short3 condition for development
  - → Assess the SR system performance

2

Sub-system Descriptions: Commons

→ Speech/non-speech segmentation (I,II,III,IV)

- MLP speech-non-speech based detector combined with simple bi-Gaussian model of the log energy distribution
- Segmentation of the interview segments was post-processed

→ Gender-dependent Universal Background Models (UBMs)

- 72 hours from 870 male speakers and 100 hours from 1200 female
- GMM modeling with 1024 Gaussians and 20 EM iterations
- Two sets of UBMs were trained, for spectral and prosodic features

→ Spectral features (I,III,IV)

- 60 features (20ms frame): 19 PLP static with RASTA + E + Δ + ΔΔ
- Mean and variance feature normalization

→ Prosodic Features (II)

- Energy and pitch contours of syllable-like region modeled with Legendre polynomials of order 5.
- Log-pitch and log-energy of voiced regions with Snack toolkit.
- 13 features: 6 pitch + 6 energy + 1 length of the syllable-like region

→ Phonetic alignment generation (V)

- Use of NIST automatic transcriptions.
- MLP acoustic models trained with PLP (13 static + Δ), log-RASTA (13 static + Δ), MSG (28 static), ETSI AFE (13 + Δ + ΔΔ)

→ ZT normalization (I,II,IV)

- From SRE2004 and SRE2005 data: 400 Z-segments and 400 T-segments
- Gender-dependent: 200 male and 200 female

3

The L^2F-UPC SR sub-systems

→ JFA

- Cookbook developed at Brno University of Technology
- 300 eigenvoices using 372 male and 519 female multi-session speakers
- 80 eigenchannels trained on telephone data (184 male and 245 female)
- *D* estimated from NIST SRE 2006 data (298 male and 402 female)
- Linear Scoring and ZT normalization (200 male and 200 female)

→ GSV-SVM

- Gaussian supervector approach based on stacked GMM-means
- Linear SVM kernel for speaker models training (libSVM toolkit)
- Background formed by 874 male and 1204 from SRE 1 side corpora.

→ GSV-GMM

- Speaker SVM models are *pushed back* to *positive* and a *negative* GMMs
- Z Normalization with only 100 male and 100 female

→ TN-SVM-NAP

- **Novel** approach that makes use of ASR speaker adaptation features
- Based on connectionist ANN/HMM ASR ⇒ Linear Input Network that maps SD input vectors to SI characteristics
- NAP dimension 32 (670 male and 921 female multi-session speakers)

4

Fusion and Calibration

→ SRE2008 *short2-short3* data set was used for calibration and fusion

→ Linear logistic regression (FoCal) to perform a two stage calibration:

- Independent sub-system calibration
- Joint calibration of the five sub-systems

→ Three different calibration and fusion configurations

→ New cost parameters for calibration and decision threshold setting

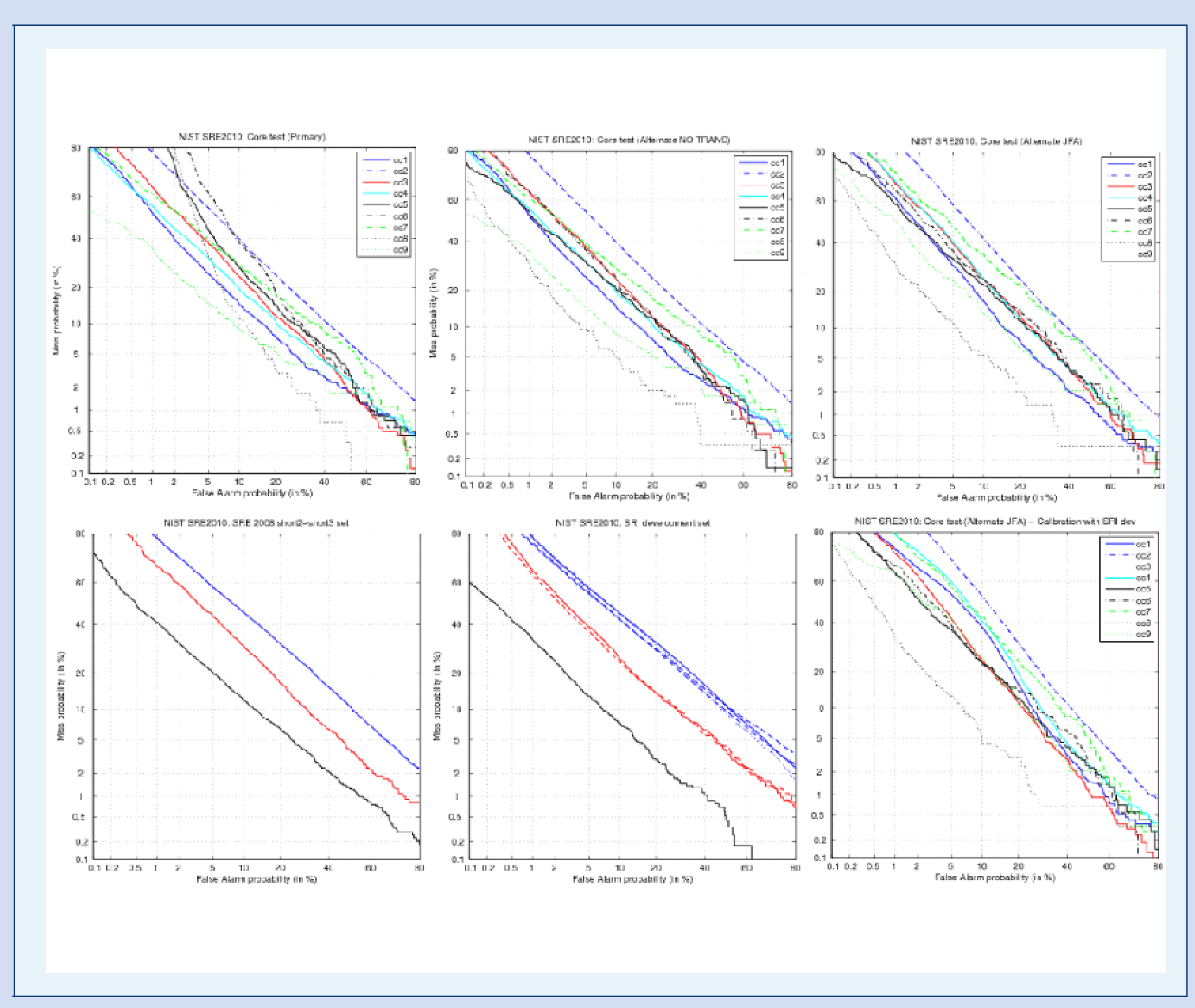
Configuration	*SRE'08*	*SRE'10*	
		model	**segment**
MIC - TEL	*interview-phonecall/telephone*	*interview*	*phonecall-telephone*
TEL - TEL	*phonecall-phonecall/telephone*	*phonecall-telephone*	*phonecall-telephone*
MIC - MIC	*interview-interview*	Rest of trials	

→ An error was detected in the **primary** *tel-tel* configuration

5

Experiments

Development on SRE'08, results on SRE'10 and post-evaluation



- tel-tel SRE'08 development constrains results on non-seen conditions
- Post-evaluation results do not show a calibration issue

6

Conclusions

- First participation focused on algorithms development/assessment
- Cross-channel problems not sufficiently addressed (focused on *tel-tel*)
- New cost function and vocal effort problem challenges were ignored
- Some methods not applied due to time restrictions (i.e. NAP in **(III)**)
- Need more post-evaluation analysis

7