Loquendo
VOCAL TECHNOLOGY AND SERVICES

NIST SRE 2006 Workshop

Loquendo - Politecnico di Torino site presentation

Claudio Vair, Daniele Colibro, Alba Fiorio
Loquendo

Emanuele Dalmasso, Pietro Laface
Dipartimento di Automatica e Informatica Politecnico di Torino

June 26, 2006

About Loquendo

- **Loquendo is a Telecom Italia company headquartered in Turin, Italy**
- **Loquendo's offering is a complete range of speech technology components, including:**
  - Loquendo TTS synthetic speech engine
  - Loquendo ASR speaker-independent speech recognition engine
  - Loquendo Free Speech Identification engine
  - VoiceXML Interpreter and a range of platform solutions
- **First participation in NIST Speaker Recognition Evaluation**

Outline

- System description
- Feature domain intersession compensation
- Development data
- Analysis of the results

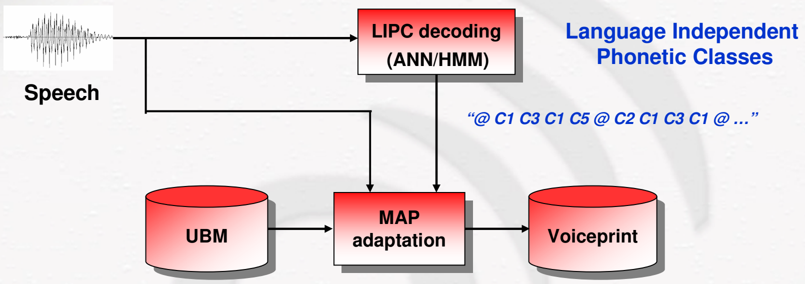
Outline

- System description
- Feature domain intersession compensation
- Development data
- Analysis of the results

System description

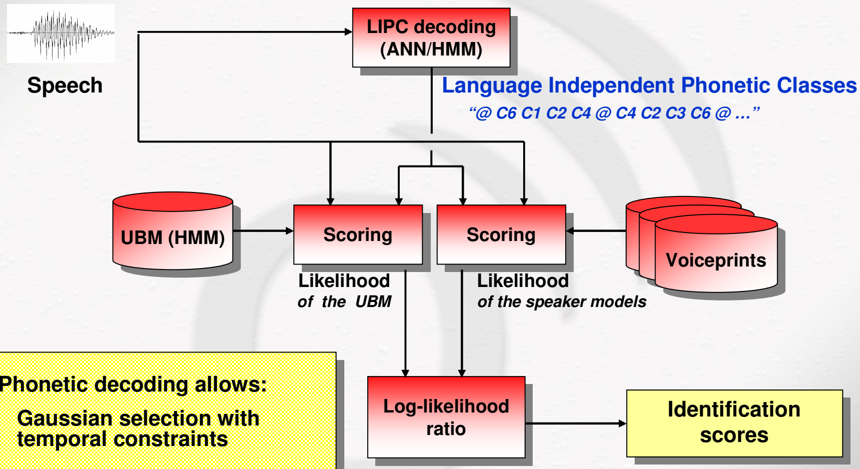
- **Two independent GMM systems were used for the evaluation**
  - Phonetic GMM (PGMM) [Loquendo]
  - GMM [Politecnico di Torino]
- **The primary system's scores were obtained by the linear fusion of the two GMM systems**

Phonetic GMM - Training



- **Phonetic decoding of the utterance producing language independent broad phonetic class segments**
- **ANN trained pooling 20 hours of speech of 10 different languages (SpeechDat2 corpora)**
- **Gender independent UBM trained on the same ANN training data, ~2000 Gaussians**

Phonetic GMM – Testing



- **Phonetic decoding allows:**
  Gaussian selection with temporal constraints

GMM system

- The GMM system is similar to the PGMM without the phonetic decoding step
- The UBM is gender independent with 512 Gaussians
- It was trained with 20 hours of speech from the NIST 2000, the OGI National Cellular, and HTIMIT corpora
- Fast Gaussians selection is achieved by means of a "road-map" based approach

Acoustic features

- **MFCC parameters with appended delta**
  - GMM: 13 cepstrals + delta, excluding C0
  - PGMM: 19 cepstrals + delta, excluding C0
- **Both systems perform feature warping to a Gaussian distribution**
  - each parameter stream warped
  - 3 sec sliding window excluding silence frames
- **The GMM system performs also feature mapping**
  - 10 models, gender and channel dependent (carbon, electrect, GSM, analog and digital)

Performed tests

- **The SRE06 primary system has been tested on all the evaluation conditions**
- **The unsupervised adaptation scores have been submitted on the core test condition**
- **The SRE05 mothball system has been tested on the 1conv4w-1conv4w condition**

Unsupervised Adaptation

- **The adaptation has been carried out in a sub-optimal "batch" mode:**
  - Testing using the un-adapted primary system
  - Selection of the adaptation test utterances, on the basis of the ZT-normed, un-adapted scores (threshold 4.0)
  - Training of the adapted model and Z-normalization: 4.76 models / target on average
  - Testing using the adapted models

Outline

- System description
- **Feature domain intersession compensation**
- Development data
- Analysis of the results

Intersession variability compensation

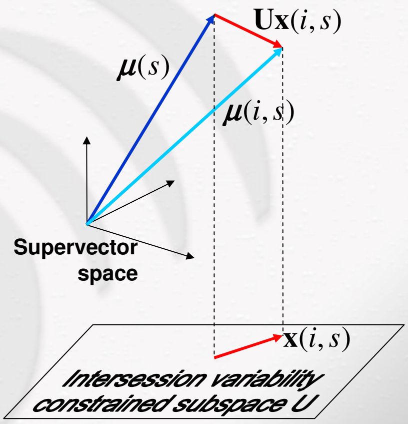
- **SDV 04 and QUT 05 results demonstrated that the intersession compensation greatly improves accuracy**
- **P. Kenny developed a complete theory on factor analysis for Speaker Recognition, applicable to intersession variability compensation**
- **Intersession variability is indeed one of the most important factor affecting the performance of SR systems**
  - Environment, recording condition, phonetic content, speaker attitude, … are examples of intersession variability
  - The variability can occur between training and testing conditions, introducing a source of mismatch
- **The proposed approaches basically make 2 assumptions:**
  - The acoustic parameters, in the models domain, are corrupted by session dependent contributes, which affect speaker recognition performance
  - The session corruption can be constrained to a low dimensional space: this allows discarding session contributes and obtaining best results

Intersession variability compensation in constrained subspace

$$\boxed{\boldsymbol{\mu}(i,s) = \boldsymbol{\mu}(s) + \mathbf{U}\mathbf{x}(i,s)}$$

- $\boldsymbol{\mu}(i,s)$ is the session dependent supervector (*) of speaker *s* for utterance i
- $\boldsymbol{\mu}(s)$ is the session independent supervector
- $\mathbf{x}(i,s)$ is the speaker dependent intersession factor vector in the constrained subspace defined by $\mathbf{U}$

(*) *The supervector of a GMM is obtained appending the mean value of all the Gaussians in a single stream*



Model domain intersession compensation

- **Model domain compensation:**

$$\mu(i,s) = \mu(s) + \mathbf{U}\mathbf{x}(i,s) \qquad (1)$$

  - In training $\mu(s)$ and $\mathbf{x}(i,s)$ are jointly estimated
  - In testing $\mu(s)$ is fixed (from training), $\mathbf{x}(i,s)$ is estimated and $\mu(i,s)$ is obtained using (1)
- **Limitations:**
  - The model domain approach is not suited for other classifiers (e.g. SVM)
  - Each model should be compensated

Loquendo VOCAL TECHNOLOGY AND SERVICES

Feature domain intersession compensation

- **For feature compensation, we estimate the intersession factor vector $\mathbf{x}(i)$ on the UBM, neglecting the speaker dependency:**

$$\mu(i) = \mu + \mathbf{U}\mathbf{x}(i)$$

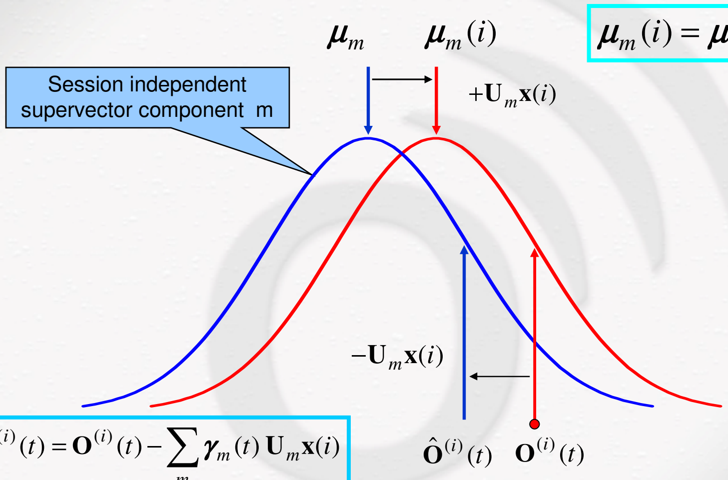
- **The compensation, defined by the intersession factor vector $\mathbf{x}(i)$, is projected in the feature domain, weighted by the *m-th* Gaussian occupation probability $\gamma_m(t)$**

$$\boxed{\hat{\mathbf{O}}^{(i)}(t) = \mathbf{O}^{(i)}(t) - \sum_m \gamma_m(t)\,\mathbf{U}_m\mathbf{x}(i)}$$

Loquendo VOCAL TECHNOLOGY AND SERVICES

Model domain versus Feature domain compensation

$$\mu_m(i) = \mu_m + \mathbf{U}_m\mathbf{x}(i)$$



$$\hat{\mathbf{o}}^{(i)}(t) = \mathbf{o}^{(i)}(t) - \sum_m \gamma_m(t)\,\mathbf{U}_m\mathbf{x}(i)$$

Outline

- System description
- Feature domain intersession compensation
- **Development data**
- Analysis of the results

Development data

- **Intersession subspace matrix training:**
  - Telephone tests: SRE04/05
  - Microphone tests: SRE05
- **Z-Norm and T-Norm: same setup used last year**
  - 160 Male + 160 Female speakers from SRE04
  - Z-norm performed on the same conditions of the test
  - T-norm performed on the same conditions of the training
- **Development and threshold tuning: SRE05**

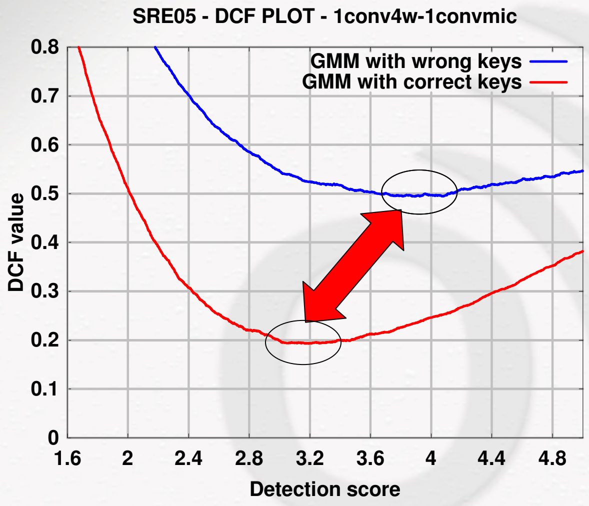
Intersession subspace matrix

- **The intersession subspace matrix training was done using different recordings (sessions) for each speaker**
- **Two subspaces trained on:**
  - SRE04 $\Rightarrow$ development purpose:
    - Female: 186 speakers, 6.5 sessions / spk
    - Male: 122 speakers, 8.8 sessions /spk
  - SRE04 + SRE05 $\Rightarrow$ SRE06 test purpose:
    - Female: 408 speakers, 10.4 sessions / spk
    - Male: 269 speakers, 11.8 sessions / spk
- **Gender dependent subspace matrixes**

Intersession subspace matrix: Xchan condition

- **SRE05 used to train the Xchan subspace**
  - Only 86 (M+F) speakers available
  - 7.6 mic. sessions / speaker
  - 1.4 tel. sessions /speaker
- **Single Xchan subspace matrix for both training and test**

XChan DCF plots



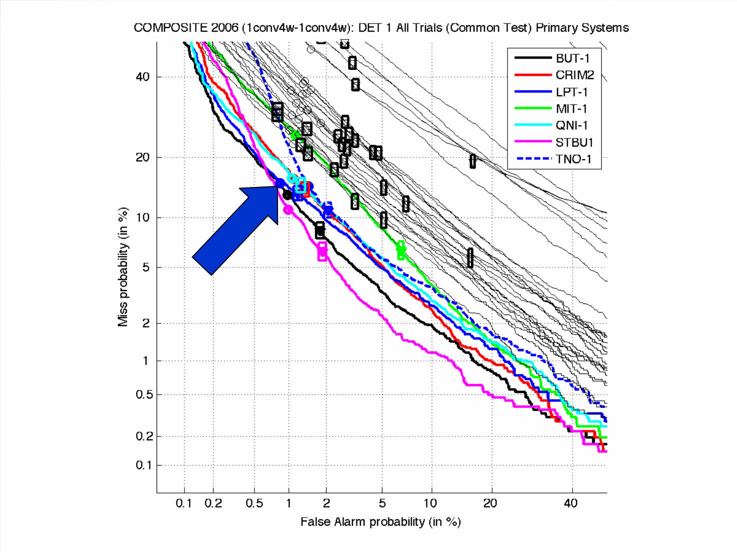
The wrong keys induced an over estimation of the min DCF threshold

Bad actual DCF on SRE 2006

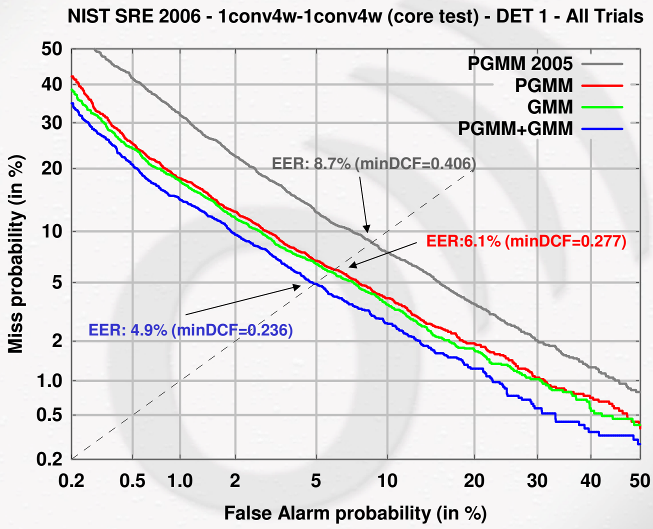
Outline

- System description
- Feature domain intersession compensation
- Development data
- **Analysis of the results**

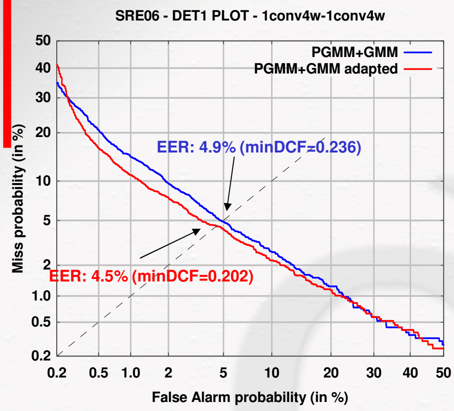
Core Test – All trials – DET1

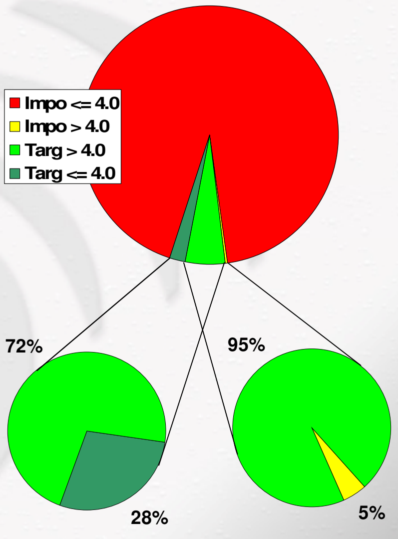


Core Test – All trials – DET1



Unsupervised adaptation





- 72% of true speaker trials were correctly selected for adaptation
- Among the trials scoring > 4.0, only 5% were impostors
- 14.4% of DCF reduction due to adaptation

Conclusions

- **Significant improvements were obtained with the new intersession compensation technique in the feature domain**
  - **⇒ 31.8% of DCF reduction**
- **The orthogonality of the fused system is a key factor for obtaining further improvement**
  - **⇒ 14.8% of DCF reduction**
- **The acoustic-only primary system demonstrate its robustness in almost all conditions and languages**
  - **⇒ best system on 12/15 all trials tests**

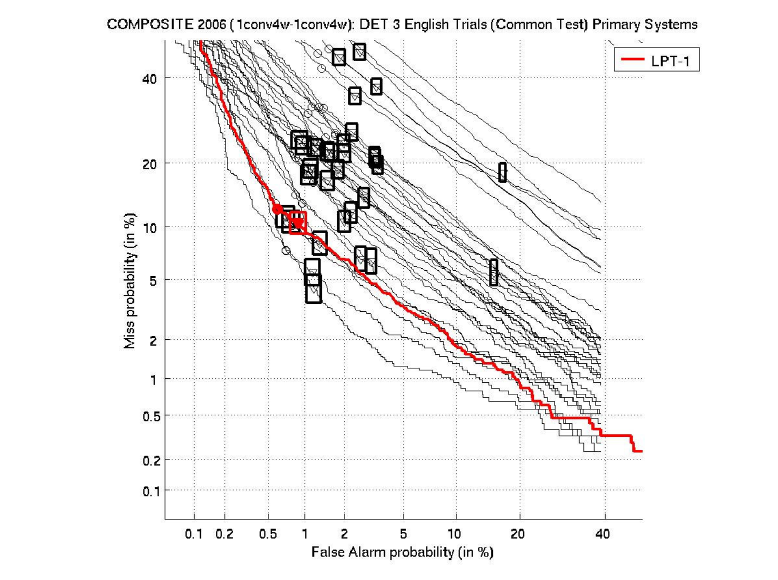
References (i)

- C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, P. Laface, *"Channel Factors Compensation in Model and Feature Domain for Speaker Recognition"*, Odyssey 2006 Workshop on Speaker and Language Recognition.
- P. Kenny, V. Gupta, G. Boulianne, P. Ouellet and P. Dumouchel, *"Feature Normalization Using Smoothed Mixture Transformations"*, in *Proc. ICSLP*, Pittsburgh, Pennsylvania, Sept. 2006 (submitted).
- P. Kenny, P. Dumouchel, *"Disentangling Speaker and Channel Effects in Speaker Verification"*, Proc. ICASSP 2004, pp. I-37-40, 2004**.**
- N. Brümmer, *"NIST SRE 2004 Evaluation Workshop"*, Toledo, Spain, 2004.
- R. Vogt, B. Baker and S. Sridharan, *"Modelling Session Variability in Text-independent Speaker Verification", Proc. INTERSPEECH-2005*, pp. 3117-3120, 2005.
- D. A. Reynolds, T. F. Quatieri, R. B. Dunn, *"Speaker Verification Using Adapted Gaussian Mixture Models"*, Digital Signal Processing, Vol. 10, pp. 19-41, 2000.
- J. Pelecanos, S. Sridharan, *"Feature warping for robust speaker verification"*, Proc. 2001: A Speaker Odyssey, pp. 213-218, 2001.

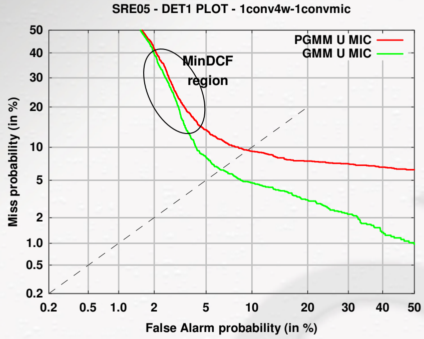
References (ii)

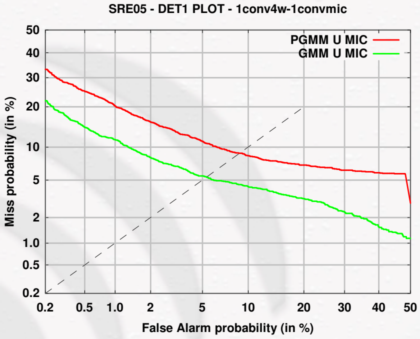
- S. Lucey, T. Chen, *"Improved Speaker Verification through Probabilistic Subspace Adaptation"*, Proc. EUROSPEECH-2003, pp. 2021-2024, 2003.
- D. A. Reynolds, *"Channel Robust Speaker Verification via Feature Mapping"*, Proc. ICASSP 2003, vol. 2, pp. 53–56.
- Povey D. &Woodland P.C., *"Frame Discrimination training of HMMs for Large Vocabulary Speech Recognition"*, Proc. ICASSP'99, pp. 333-336, Phoenix.
- R. Kuhn J.C. Junqua, P. Nguyen, and N. Niedzielski, *"Rapid Speaker Adaptation in Eigenvoice Space"*, IEEE Trans. on Speech and Audio Processing, Vol.8, No.6, Nov. 2000, pp. 695-707.
- E. Dalmasso, P. Laface, D. Colibro, C. Vair, *"Unsupervised Segmentation and Verification of Multi-Speaker Conversational Speech"*, Proc. INTERSPEECH-2005, pp. 1001-1004.
- R. Auckenthaler, M. Carey and H. Lloyd-Thomas, *"Score Normalization for Text-Independent Speaker Verification Systems"*, Digital Signal Processing, 10 (2000), pp. 42-54.

Core Test – English trials – DET3



GMM vs PGMM on Xchan condition

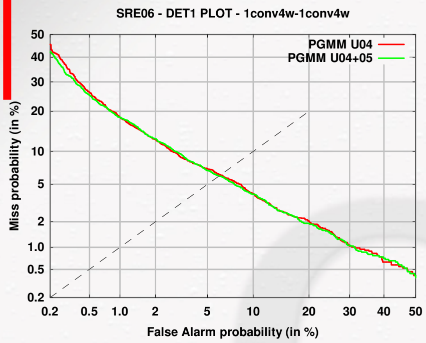


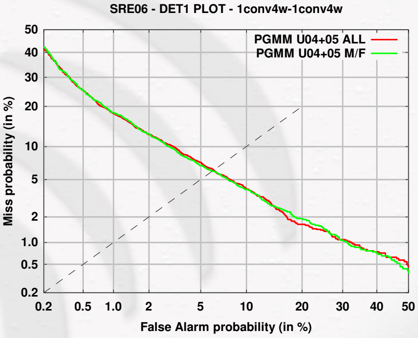


- **PGMM and GMM seem to be equivalent in the min DCF region (left plot)**

  …BUT… the keys were WRONG !!
- **With corrected keys the GMM outperform PGMM (right plot). Possible reasons:**
  - Too few Mic. data to train the big PGMM subspace matrix
  - GMM Feature Mapping
  - Unreliable phonetic decoding
- **Further investigation required…**

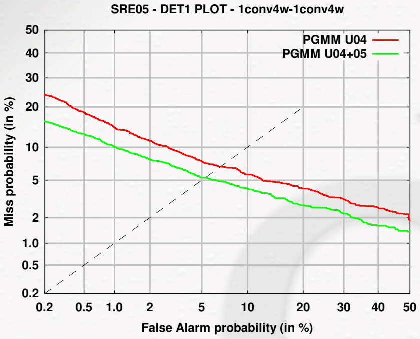
Comparison of subspace matrixes

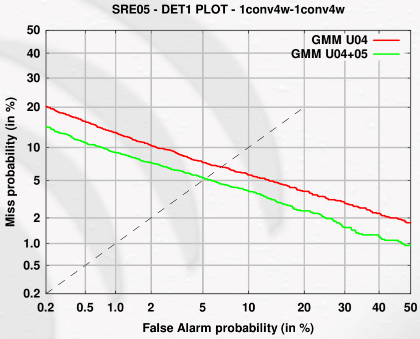




- **No difference using SRE04 or SRE04+SRE05 to train the intersession subspace matrixes (left plot)**
- **No difference using gender dependent or gender independent intersession subspace matrixes (right plot)**

Intersession subspace training: The effect of data overlap

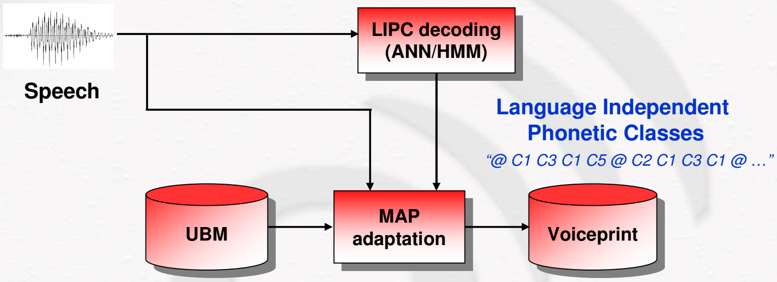




	EER	Min DCF
SRE04	6.8%	0.231
SRE04+SRE05	5.3%	0.173

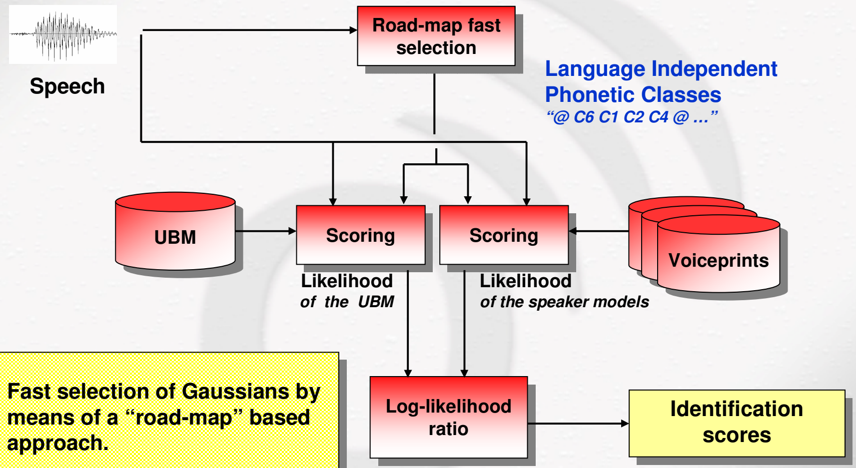
	EER	Min DCF
SRE04	6.8%	0.208
SRE04+SRE05	5.2%	0.156

GMM – Training



- The UBM is a gender independent GMM with 512 Gaussians
- Trained with 20 hours of speech from the NIST 2000, the OGI National Cellular, and HTIMIT corpora

GMM – Testing



- Fast selection of Gaussians by means of a "road-map" based approach.

Feature domain intersession compensation (2)

- The compensation, defined by the intersession factor vector $\mathbf{x}(i)$, is projected in the feature domain, weighted by the *m-th* Gaussian occupation probability $\gamma_m(t)$

$$\hat{\mathbf{O}}^{(i)}(t) = \mathbf{O}^{(i)}(t) - \sum_m \gamma_m(t)\, \mathbf{U}_m \mathbf{x}(i)$$

LPT1 Standing – Actual DCF
English trials

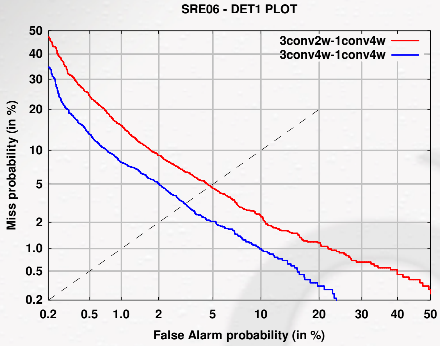
		Test Segment Conditions			
		10 sec. 2 chan.	1 conversation 2 chan.	1 conversation summed chan.	1 conversation aux mic
Training conditions	10 seconds 2 channels (4 wires)	16			
	1 conversation 2 channels (4 wires)	2	7	1	9
	3 conversations 2 channels (4 wires)	1	2	1	1
	8 conversations 2 channels (4 wires)	1	7	1	5
	3 conversations summed chan. (2 wires)		1	1	

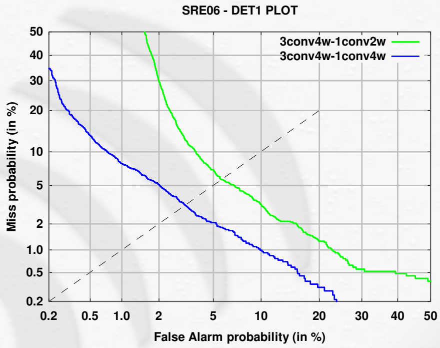
LPT1 Standing – Actual DCF
All trials

		Test Segment Conditions			
		10 sec. 2 chan.	1 conversation 2 chan.	1 conversation summed chan.	1 conversation aux mic
Training conditions	10 seconds 2 channels (4 wires)	16			
	1 conversation 2 channels (4 wires)	1	1	1	7
	3 conversations 2 channels (4 wires)	1	1	1	1
	8 conversations 2 channels (4 wires)	1	1	1	7
	3 conversations summed chan. (2 wires)		1	1	

- The acoustic only approach of our system demonstrate its robustness for all conditions and all languages

2 Wires Conditions





- We used unsupervised speech segmentation to detect speaker cluster in all 2 wires train / test conditions
- For 2w tests, each putative speaker cluster is scored against the speakers models in the index file and the best score is selected

One subspace matrix vs two on Xchan condition