The IRISA/METISS systems for NIST SRE05

- **METISS Research Group:**
  - ⋆ Since 2001, research project of the INRIA (National Research Institute for Computer Science and Automatics);
  - ⋆ Audio Signal Processing (including Speech Processing):
    speaker characterization, information detection/tracking in audio streams,
    "advanced" processing (e.g., Blind Source Separation), speech recognition;
  - ⋆ 3 permanent researchers, 2 engineers, 1 post-doc, 5 PhD students.

- **More info:**

  `http://www.irisa.fr/metiss/`

- **Speaker:**

Sacha KRSTULOVIĆ `sacha@irisa.fr`

METISS and the NIST SRE05 campaign

Our research interests include:

- designing simpler/lighter score computation and normalization schemes;
- re-thinking the speaker verification process as a "template matching" procedure rather than a "stream scoring" procedure:

  $\Rightarrow$ reduction of the complexity of the scoring
  (e.g., for embedded systems, faster online scoring etc.);
- applying speaker verification techniques in the framework of the segmentation of audio documents:

  $\Rightarrow$ big audio databases demand lighter/faster recognition techniques.

The IRISA systems for NIST05

IRI_2

Model Space System

- Train a GMM on the test utt.
- Scoring in the model space (direct distance between Speaker/Utterance GMMs)

⇒ Simpler scoring
⇒ Simpler normalization

IRI_3

Piecewise Linear Score Approximation

- Feature space partition with a decision tree
- Region-dependent linear approximation of the LLR score

⇒ Faster/lighter scoring

IRI_5

Summed Channel System

- Frame clustering with Hierarchical Clustering and GMM distance

⇒ Efficient speaker segmentation

Baseline GMM/UBM

- LLR scoring in the feature space

IRI_1

The IRISA systems for NIST05

- IRI_1: *GMM-UBM with LLR+T-norm scoring*

  Tasks: 1conv4w-1conv4w, 1conv4w-10sec4w, 10sec4w-10sec4w.

- IRI_2: *GMM-UBM with model-space scoring*

  The LLR is replaced with an estimate of the Kullback-Leibler distances between the GMMs. New normalization paradigm: BiT-norm.

  Tasks: 1conv4w-1conv4w

- IRI_3: *Decision Trees and Linear Score Approximation*

  Fast/computationally light approximation of the GMM scoring.
  (Intended for use with SIM cards.)

  Tasks: 1conv4w-1conv4w, 1conv4w-10sec4w, 10sec4w-10sec4w.

- IRI_5: *Summed-Channel System*

  GMM-UBM with a speaker segmentation pre-processing.

  Tasks: 1conv4w-1conv2w.

IRI_1: Baseline GMM-UBM

1conv4w-1conv4w, 1conv4w-10sec4w,
10sec4w-10sec4w, 1conv4w-1conv2w

Sacha Krstulović	sacha@irisa.fr
Gilles Gonon	gonon@irisa.fr
Guillaume Gravier	ggravier@irisa.fr

IRI_1: Primary system (baseline)

"Plain vanilla" GMM-UBM:

- Preprocessing: silence removal, removal of low energy frames (2 mono-Gaussian EM classifiers);
- 16 LFCC + Delta + energy; CMS and variance normalization;
- UBM: 2048 Gaussians, diagonal covariances;
  training data: 100 males + 100 females from NIST 2004;
- speaker models: MAP (means only), 1 iteration, r=8;
  development set: 45 male, 159 female speakers from NIST04;
- scoring: LLR with T-norm (100 male, 99 female from NIST04).

IRI_2: Model space system

1conv4w-1conv4w

Mathieu Ben	mben@irisa.fr
Sacha Krstulović	sacha@irisa.fr
Frédéric Bimbot	bimbot@irisa.fr

IRI_2: Model space system (1/4)

General principle :

1. estimate a GMM on the test material (one pass MAP)
2. compute a detection score based on distances between models
3. apply normalizations in the model space

Model distance and score definition :

- for two adapted GMMs P an $\tilde{P}$ (adaptation of the means only):

$$D_E(P,\tilde{P})^2 = \sum_{k,d} w_k . \frac{(m_{k,d} - \tilde{m}_{k,d})^2}{\sigma_{k,d}^2} \quad \geq \quad KL2(P,\tilde{P})$$

  It can be shown [BEN04] that $D_E(P,\tilde{P})^2$ is strongly correlated with $KL2(P,\tilde{P})$: correlation coefficient = 0.99.

- Score : $S(X,Y) = D_E(P_Y, P_\Omega)^2 - D_E(P_Y, P_X)^2$

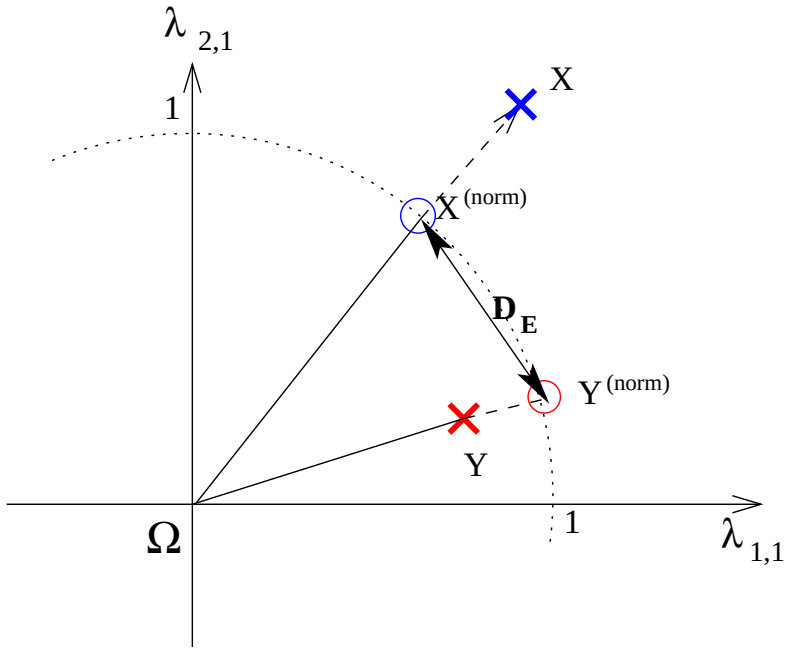
Model space system (2/4)

Euclidean model space :

- $D_E(P, \tilde{P})$ is homogeneous to a Euclidean distance :

$$D_E(P, \tilde{P}) = ||\underline{\lambda} - \underline{\tilde{\lambda}}||_2, \text{ with } \underline{\lambda} = \{\lambda_{k,d}\} = \{\sqrt{w_k}\frac{m_{k,d} - m^{\Omega}_{k,d}}{\sigma_{k,d}}\}$$

- Model normalization = projection on a unit hypersphere



Model space system (3/4)

Score normalization :

- Let $\mu_Y = \frac{1}{N} \sum_i S(T_i, Y)$ be the mean distance of the *test model* P_Y to the T-norm models P_{T_i} (σ_Y the standard deviation);
- let $\mu_X = \frac{1}{N} \sum_i S(T_i, X)$ be the mean distance of the *client model* P_X to the T-norm models P_{T_i} (σ_X the standard deviation);

$\Rightarrow$ classical T-Norm (asymmetric) : $$\frac{S(X,Y) - \mu_Y}{\sigma_Y}$$

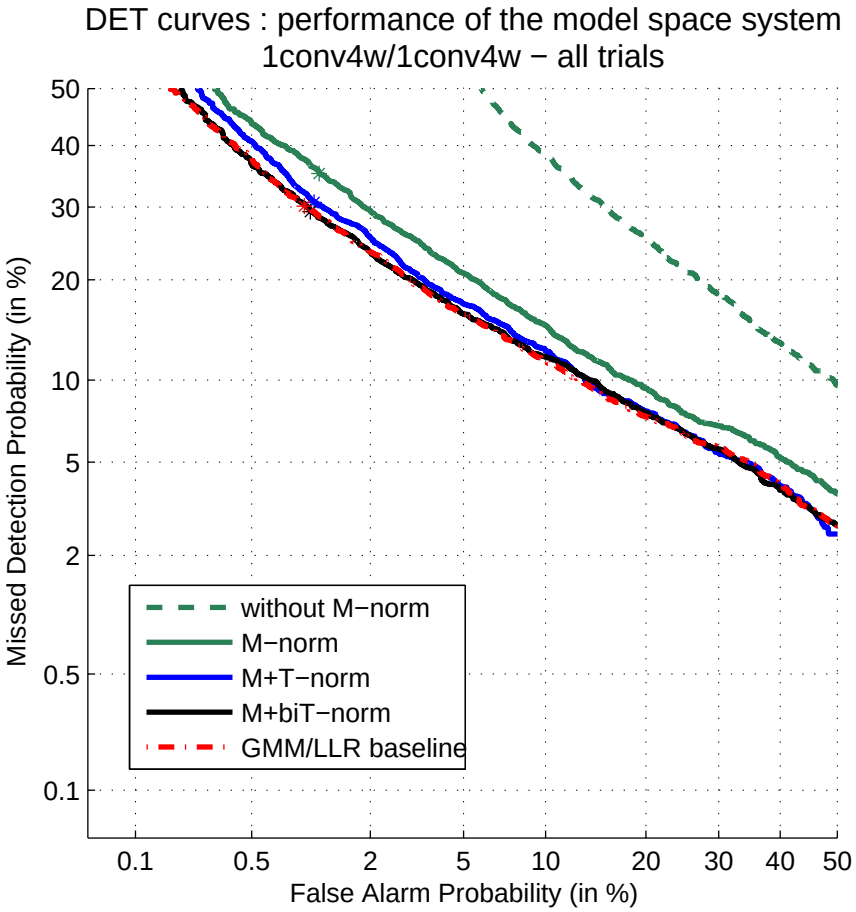
$\Rightarrow$ bi-directional T-Norm : $$\frac{S(X,Y) - \mu_X}{\sigma_X} + \frac{S(X,Y) - \mu_Y}{\sigma_Y}$$

$\Rightarrow$ No need to score the frames once the models are trained: scoring+normalization use simple Euclidean distances.

$\Rightarrow$ **Future work:** channel & other normalizations ? (PCA, ...)

Model space system (4/4)

Results



- submission to NIST was bugged
- de-bugged results :
  - model normalization (M-norm) is crucial
  - bi-directional T-norm improves performance over classical T-norm
  - same performance as the GMM/LLR baseline system when bi-Tnorm is used

IRI_3: Decision tree based system

1conv4w-1conv4w, 1conv4w-10sec4w

Gilles Gonon	gonon@irisa.fr
Rémi Gribonval	remi@irisa.fr
Frédéric Bimbot	bimbot@irisa.fr
Sacha Krstulović	sacha@irisa.fr

IRI_3: Decision tree based system (1/4)

- **Context of the work: the Inspired project**
  - Integrated Secure Platform for Interactive Personal Devices
  - Research, development, testing and certification on next generation secure smart devices
  - Feasibility of biometric authentication on such devices, including speaker recognition.

- **Description of the system**
  - Decision trees and Linear Regression are combined to provide a *piecewise linear approximation* of the scoring function of a GMM/UBM system.

$\Rightarrow$ builds on a trained GMM/UBM system:

  - $\star$ 12 LFCC+Δ+energy, GMMs w/ 128 components

  - The scoring complexity becomes suitable for embedded devices.

Decision tree based system (2/4)

- **Piecewise linear approximation of the LLR with trees:**

1. The tree divides the feature space in client/world regions.
   (Training: CART method.)

   ⋆ Trick: the feature space is augmented with a fixed set of *oblique discriminant features*, related to the underlying GMM system:

   □ find the Gaussians most shifted by the MAP adaptation

   ⇒ best locally discriminating directions: $\Delta\mu_i = \Sigma^{-1}(\mu_i^X - \mu_i^{\bar{X}})$

   ⇒ feature projection (scalar product): $< \Delta\mu_i, y_t >$

2. A linear scoring function is affected *a posteriori* to each region/leaf

   ⋆ Multiple regression on the development set's LLRs over each region (Ordinary Least Squares).

   Partly smooths the discontinuities of the piecewise approximation, as opposed to a hard score (+1/-1) or an average score over a region.

Decision tree based system (3/4)

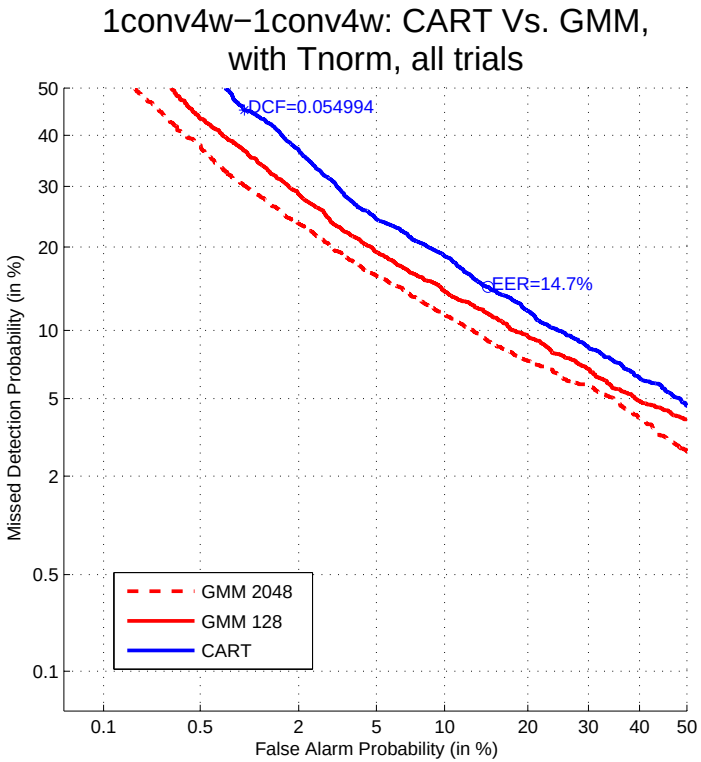
- **Complexity of the scoring for NIST 05 SRE:**

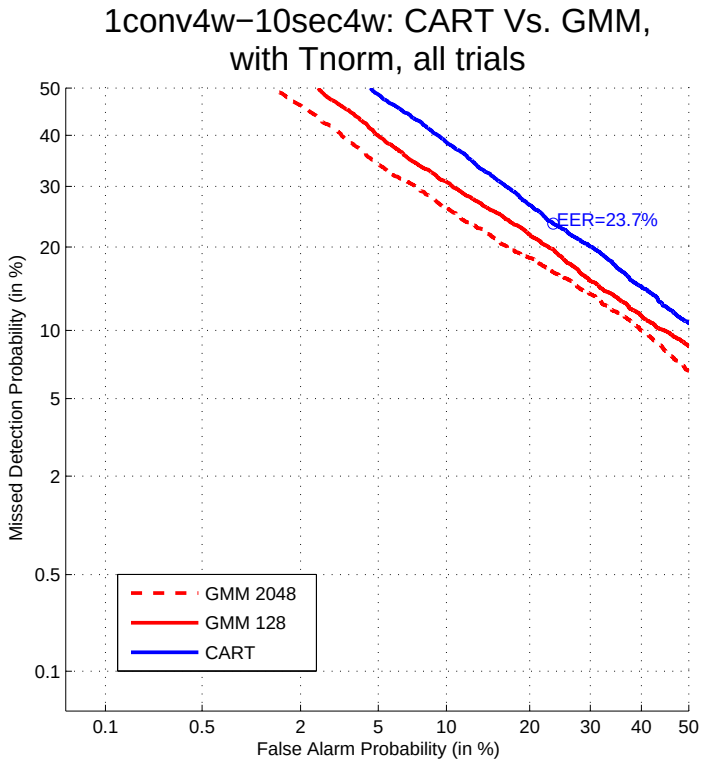
Frame score = region-dependent linear combination of frame features
Final score = mean score for all the frames

  - Average number of leaves (nbr of regions): 320 (min:183, max:355).
  - Tree depth (nbr of tests per frame): minimal=4, maximal=20.
  - Number of multiply-adds per frame: $100 < N < 500$.
  - Suitable for real-time/streaming scoring applications.
  - Memory size of the speaker templates: 64kB.

- **But:**
  - the tree training complexity is huge (greedy algorithm);

Decision tree based system: results (4/4)





IRI_5: Summed channel system

1conv4w-1conv2w

Daniel Moraru	dmoraru@irisa.fr
Guillaume Gravier	ggravier@irisa.fr
Sacha Krstulović	sacha@irisa.fr

IRI_5: Summed Channel System (1/3)

General principle of the summed channel system:
the speech segments of the test file are divided in two speaker-dependent frame subsets which are scored independently.

1. Segmentation of the test file in mono-speaker segments:
   silence detection + speaker change detection.
2. Clustering of the segments in 2 speaker-dependent classes:
   Agglomerative Hierarchical Clustering.
   $\Rightarrow$ Extraction of two subsets of speaker-dependent frames.
3. Speaker verification: core system applied to both frame subsets, final score taken as the max of both tests.

The parametrization is not the same for all the steps.

Summed Channel System (2/3)

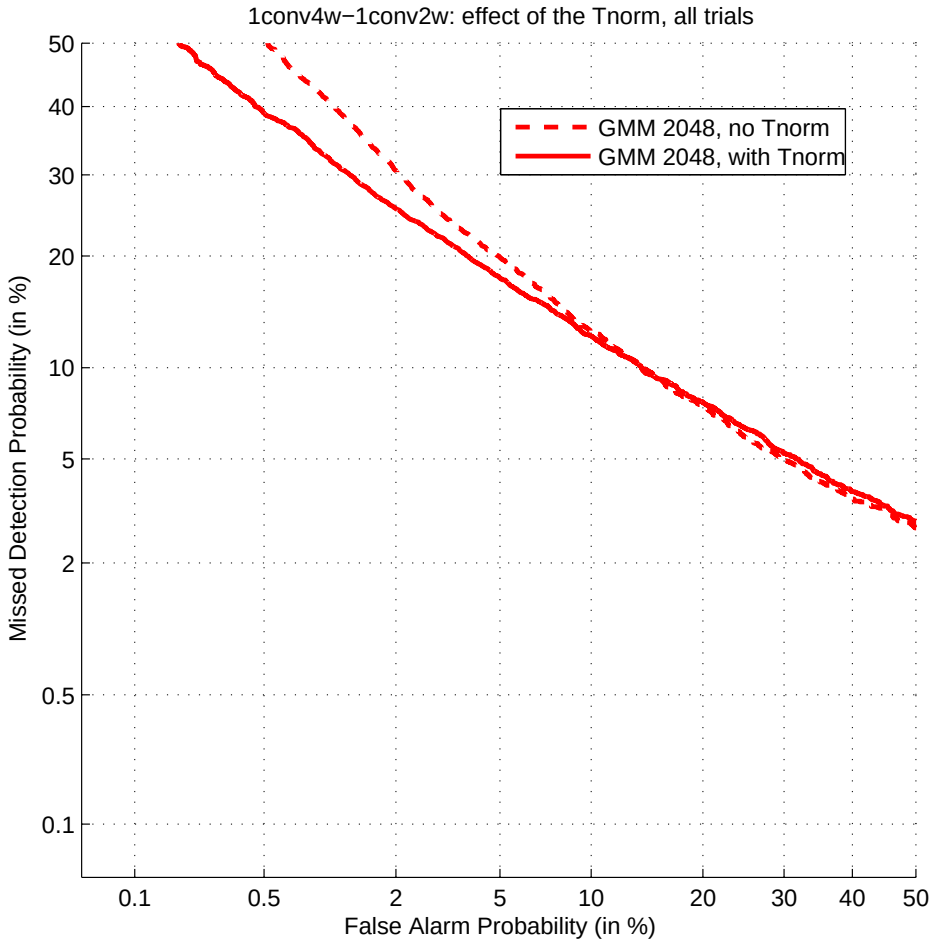
Step 1, segmentation:

- silence detection: energy bi-Gaussian;
- speaker change detection in the speech segments:
  Bayesian Information Criterion (BIC) w/ full cov. mono-Gaussian, 24 Mel Filter Banks coeffs. (∼3% misclassified frames in ESTER evaluations.)

Step 2, clustering:

- segment models: GMMs, MAP adapted from a UBM, 1 pass, r=30;
- UBM: 64 comp. diagonal GMM, gender independent, 16 MFCC+Energy;
- distance: model space based, approx. of the Kullback-Leibler Distance;
- re-estimate a model on the fused segments at each iteration;
  stop when 2 classes, ideally corresponding to the 2 speakers, are left.
  (∼17% misclassified frames in ESTER evaluations.)
- Very fast.

Summed Channel System (3/3)



Step 3, scoring:

- Independent scoring of the two frame subsets with the core system.
- Final score = max. of both scores.
- T-Norm improves the performance at the DCF point, but not the EER.

Summary and conclusion

Assets of the IRISA systems:

- *simpler normalization* of the scores in the *model space*;
  (Towards channel compensation in the model space ?)
- *faster/lighter scoring* for important amounts of data, both with the *model-space scoring* and the *piecewise linear score approximation*;
- application to *fast speaker segmentation*.

Thank you for your attention.