TECHila2: The 2010 NIST-SRE System

Juan A. Nolazco-Flores, PhD; and Leibny Paola Garcia-Perera, MSc;
Roberto Aceves, BSc; Daniel Escobar, BSc; Benjamín Elizalde, BSc;
Tecnológico de Monterrey, México

TECNOLÓGICO DE MONTERREY®

Introduction

Objectives

- To test TECHila2, Speaker Verification System from Tec de Monterrey, under NIST SRE 2010 core database.
- To test our computer infrastructure and configuration on such computing demanding task.
- To show the evolution of the state of the art algorithm implementation in SV.

Approach

Our system follows the hypothesis testing theory using a Gaussian Mixture Models (GMM) framework in two stages: enrollment (training) and verification (test) (Figure 1). The maximum a priori (MAP) algorithm adapts a UBM (universal background model or anti-model) to compute target models. These models are evaluated applying several trials to produce an error rate.

Infrastructure

TECHila2

Matlab

SGE (http://gridengine.sunsource.net)

GNU-Linux

Beowulf cluster 20 CPUs	I686 @ 3GHz	1Gbps	7TB storage

System Description

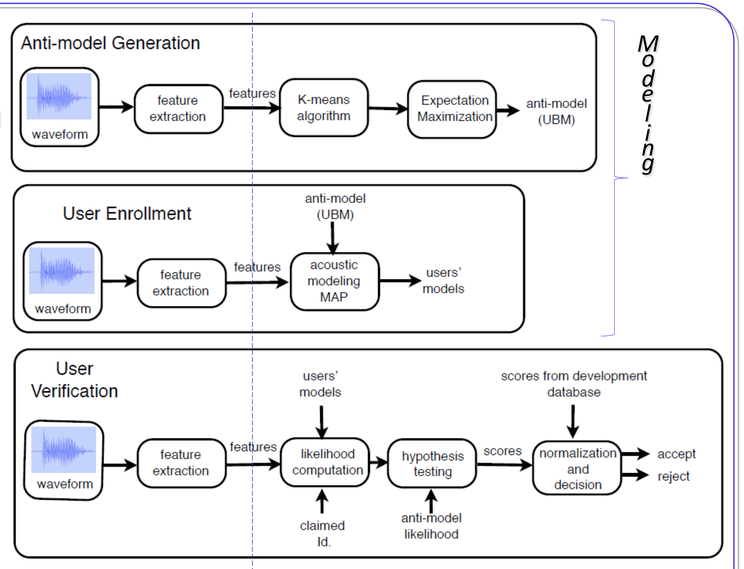


Figure 1

Feature Extraction

- A 10 ms frames rate and a 25ms analysis windows are used.

Feature Vectors

16 MFCC	•16 MFCC
Δ Log(E)	•Delta (Log (E))
Δ 16 MFCC	•Deltas from 16 MFCC
Δ^2 16 MFCC	•Double Deltas from 16 MFCC

Frame Removal

- MFCC's frames are labeled as low, medium or high, based on the frame energy. Using these tags a three-mixture GMM is built.
- Those frames belonging to the low energy Gaussian and the bottom 80% of the medium Gaussian are discarded. This threshold is defined on an heuristic basis.
- It is assumed that frames from silences (with low energy or noise) don't have significant speaker information.
- Dynamic coefficients are computed afterwards from the remaining frames.

Feature Warping Normalization

- To compensate channel distortion, feature warping is used to Gaussianise the MFCCs.

Speaker Modeling

- A gender-dependent target-independent anti-model, also known as UBM, is generated based on a GMM with 512-mixtures .
- To reach a faster convergence the UBM is initialized using a parallel K-means algorithm and ended up with 512-centroids.
- UBM is trained from a pool of utterances of NIST-SRE 2004 database.
- EM (Expectation Maximization) is used to get the GMM parameters. It's iterated until converging (~5 iterations).
- Target-models are obtained using MAP (Maximum APosteriori) speaker adaptation.

Evaluation and Decision

Evaluation Score

- It is based on hypothesis testing theory:
  - H0: to accept the speaker as legitimate.
  - H1: to reject him/her.

Then the score computation is as follows:

$$score = \log\left(\frac{\text{likelihood(H0)}}{\text{likelihood(H1)}}\right)$$

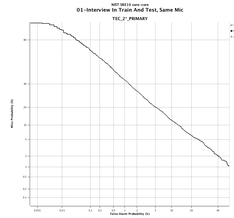
Decision

- A target-dependent threshold is pursued.
- The distribution of impostor scores is normalized to have zero mean and unit variance.
- Estimate of the distribution of the target-trials is built using the training data.

Results

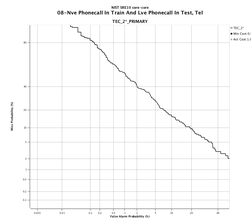
- Some of the obtained DET graphs are shown:

Same mic
Train: Interview
Test: Interview



Mic
Train: Nve phonecall
Test: Hve phonecall

Tel
Train: Nve phonecall
Test: Lve phonecall



Mic
Train: Nve phonecall
Test: Lve phonecall

References

[1] D. Petrovska-Delacretaz, A. El-Hannani, and G. Chollet. "*Text-Independent Speaker Verification: State of the Art and Challenges*", LNCS Springer, May 2007.

[2] J. Pelcanos and S. Sridharan. "*Feature warping for robust speaker verification*". 2001: A Speaker Odyssey Workshop. June 2001.

[3] S. Chen and R. Gopinath. "*Gaussianization*", NIPS 2000.

[4] J. Gauvain and C. Lee, "*MAP Estimation of Continuous Density HMM: Theory and Applications*", DARPA Sp. & Nat. Lang. Workshop, Feb. 1992.

[5] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz and D. A. Reynolds. "*A Tutorial on Text-Independent Speaker Verification*", EURASIP Journal on Applied Signal Processing 2004.