TECHila2 2010 NIST-SRE System Description

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

ITESM, campus Monterrey, Mexico

1. Introduction

Among all biometrics, speech is a competitive alternative for secure authentication due to its natural production, non-invasive technology, and its capacity of performing remote transactions [5]. Speaker Verification (SV) is the task of accepting or rejecting a user according to a given segment of speech and a claimed identity. It presents two stages: enrollment and verification. In the enrollment, each registered user provides positive tokens to build a speaker model. During the verification each trial is evaluated as an impostor or a target using several trials.

The progress of SV has shown its evolution in every each NIST competition. It has demonstrated several challenges that need to be addressed.The Speech Group from Tec de Monterrey, developed a system for NIST 2010 evaluation, called TEChila2. It represents the evolution of algorithm implementation from our earlier systems that had used alternative data sets (YOHO, SV-TIMIT and NIST2008). Despite that the evaluation consist in varied tests, we have only concentrated in the core test. Furthermore, due to computation requirements the configuration of our cluster was a challenge too.

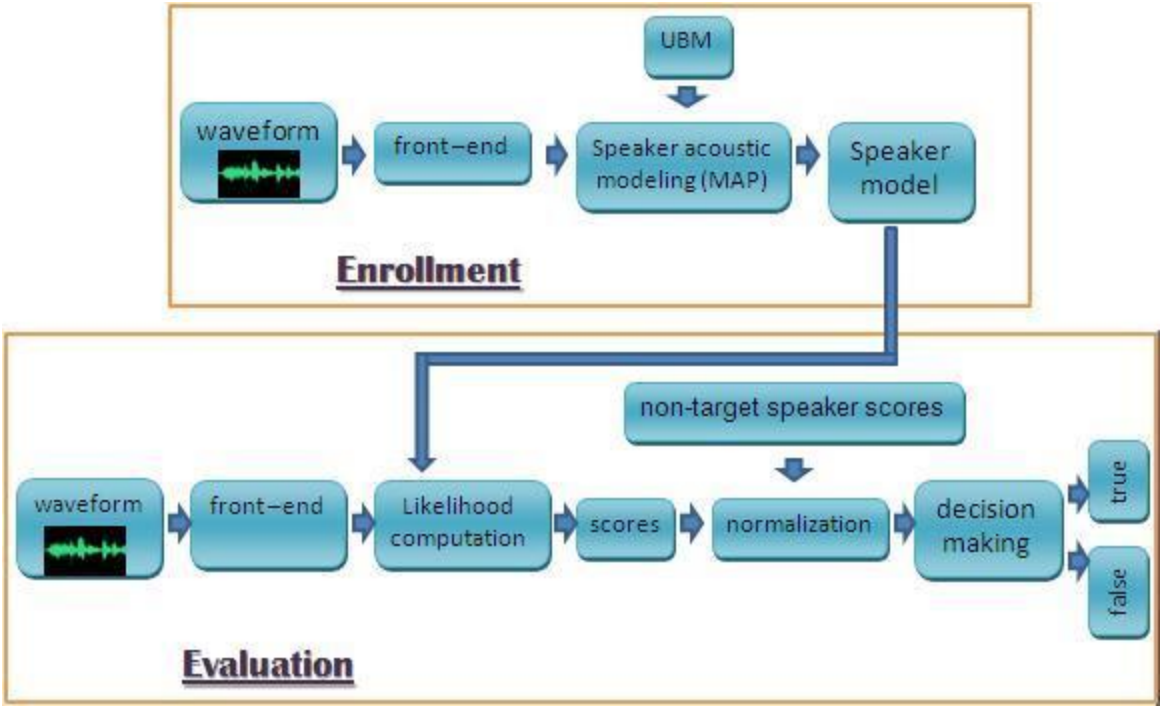
2. System Description

2.1 Infrastructure and Tools

Autonomous Beowulf cluster with 20 CPUs i686 3GHz, 1Gbps LAN, 7TB storage.
SGE, Matlab, Python, Perl, GNU-Linux.

2.2 TECHila2 System

The following block diagram describes TEChila2.



Parametrization

Feature-Extraction

A short-time 256-pt Fourier analysis is performed on a 25ms analysis window and 10ms frame rate. The magnitude spectrum was transformed to a truncated vector of Mel-Frequency Cepstral Coefficients (MFCC) as follows:

Cep. Coeff. 1..13:	16
delta(*logE*):	1
delta(*Cep*) :	16
delta(delta(*Cep*)):	16

where *logE* is the logarithm of the frame's waveform energy, and *Cep* represents the cepstral coefficients with order 1 through 16. We have observed that including more than 13 coefficients can improve the performance of the gain, however the computation time increases as well. A trade off between both of them should be taken into consideration for future implementation.

Feature-Frame-Removal

We decided to use frame removal instead of VAD, or a speech recognizer. The main idea is to eliminate in just one effort the silence intra-words and the silence between words. For a given conversation side, every frame log-energy was tagged as *high*, *medium* and *low*, as suggested in [1] using a 3 mixture - GMM. Low and 80% of the medium log-energy frames

were simply discarded. The delta and double delta were obtained after these silent frames have been removed. This 80% threshold was a heuristic.

Feature Warping-Normalization
The Gaussianization methods [2,3] have been very effective in SV. The task of feature warping is to undo the distortion caused by the channel by warping each attribute's scale so that the resulting attribute has a normal distribution. The underlying idea in this normalization scheme is that every spectral attribute (cepstral coefficient in our case) is normally distributed across time, and that the transmission channel distorts such distribution.
Traditionally, this warping is accomplished by first assembling an empirical CDF (cumulative distribution function) from the ranked features within 1.5 seconds after and before the current frame (3 seconds total), and then perform the CDF-inverse at the current frame.

Modeling

A gender-dependent and target-independent 512-mixture GMM anti-model model was trained from the core core of NIST-SRE 2004 database. EM (expectation maximization) algorithm was used to obtain the maximum likelihood estimates of the GMM parameters. TECHila2's implementation of EM algorithm for GMM emulates MPI environment to take full advantage of parallel computing infrastructure. For every iteration of EM, TECHila randomly polls 25% of the training tokens, corresponding approximately to 3 hours of speech.
The GMM is first initialized using the K-means algorithm to obtain a set of 512 centroids. By using the k-means the performance of the EM was simplified, however it is always important to check that the local bounds are not very restrictive, so that EM can provide a satisfactory estimation.
The EM is then repeated after the model had converged (~3 to 5 iterations).

The target-models are obtained with a traditional MAP (maximum a posteriori) speaker adaptation [4].

Scores
Under this framework, the score is given by the log likelihood ratio of two models: target-model and anti-model. In the current implementation, the anti-model is target-independent. We used ZNORM [1]. A set of synthetic impostor trials was induced offline and these values were used to re-scale the evaluation trials' scores so that impostor trials have a score centered at zero with unit variance.

2.3 Decision Making
Since the distribution of impostor scores was normalized to have zero mean and unit variance, the only part needed in order to choose a value for the threshold is an estimate of the distribution of the target-trials. Since no data is available for every target, other than the training data, we attempted to build such a distribution from this data. This turned out to be not a good idea, as the scores obtained were optimistically high, therefore, leading to thresholds that are too high as well.

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.

About Our Lab and Future Research Agenda

The Tecnologico de Monterrey have been doing research in speech recognition technology since 1994. Over the past decade, several projects have been developed, including cryptographic key generation based on speech signal, speaker verification (MOBIO competition) and speech recognition system for native Mexican languages, among others.

More recently in 2002, a baseline speaker verification system was developed. Since then, some improvements have evolved to finally obtain the TECHila2 system, designed to conform the guidelines for NIST SRE 2010. The main challenge of this competition was the number of trials to be processed.

Our future research agenda includes the improvement of decision making algorithm(accept/ reject) from the evaluation scores for the NIST-SRE.
Furthermore, although the implementation was carefully done to avoid waste of computation (easily done in Matlab). For NIST2010, we realized that our system needs to have faster implementation and a comparable result.