Computación y Sistemas
ISSN: 1405-5546
computacion-y-sistemas@cic.ipn.mx
Instituto Politécnico Nacional
México

García Perera, Leibny Paola; Aceves López, Roberto; Nolazco Flores, Juan
Speaker Verification in Different Database Scenarios
Computación y Sistemas, vol. 15, núm. 1, septiembre, 2011, pp. 17-26
Instituto Politécnico Nacional
Distrito Federal, México

Speaker Verification in Different Database Scenarios

Verificación de hablante en diferentes escenarios de base de datos

Leibny Paola García Perera, Roberto Aceves López, and Juan Nolazco Flores
Departamento de Ciencias Computacionales, Tecnológico de Monterrey,
Monterrey, Nuevo León, México
{paola.garcia, aceves, jnolazco}@itesm.mx

Article received on July 30, 2010; accepted on January 15, 2011

Abstract. This document shows the results of our Speaker Verification System under two scenarios: the Face and Speaker Verification Evaluation organized by MOBIO (MObile BIOmetric consortium) and the results for the Speaker Recognition Evaluation 2010 organized by NIST. The core of our system is based on a Gaussian Mixture Model (GMM) and maximum likelihood (ML) framework. First, it extracts the important speech features by computing the Mel Frequency Cepstral Coefficients (MFCC). Then, the MFCCs train gender-dependent GMMs that are later adapted to obtain target models. To obtain reliable performance statistics those target-models evaluate a set of trials and final scores are calculated. Finally, those scores are tagged as target or impostor. We tried several system configurations and found that each database requires a specific tuning to improve the performance. For the MOBIO database we obtained an average equal error rate (EER) of 16.43 %. For the NIST 2010 database we accomplished an average EER of 16.61%. NIST2010 database considers various conditions. From those conditions, the interview training and testing conditions showed the best EER of 10.94 %, followed by the phone call training phone call testing conditions of 13.35%.
Keywords. Speaker verification and authentication.

Resumen. Este documento muestra los resultados de nuestro sistema de verificación de hablante bajo dos escenarios: la Evaluación *Face and Speaker Verification Evaluation* organizada por MOBIO (MObile BIOmetric consortium) y la Evaluación de Reconociemiento de personas 2010 organizada por NIST. La parte central de nuestro esquema se basa en un modelado de Mezclas de Gaussianas (GMM) y máxima verosimilitud. Primero, se extraen los parámetros importantes de la voz calculando los coeficientes ceptrales en escala mel, Mel Frequency Cepstral Coefficients (MFCC). Después, dichos MFFCs entrenan las mezclas de Gaussianas dependientes del género que posteriormente serán adaptadas y se obtendrán los modelos de los usuarios objetivo. Para obtener estadísticas confiables esos modelos objetivo son evaluados por un conjunto de señales no conocidas y se obtienen puntuaciones finales. Por último, esas puntuaciones son etiquetadas como usuario objetivo o impostor. Hemos analizado diferentes configuraciones y encontramos que cada base de datos requiere una sintonización adecuada para mejorar su desempeño. Para la base de datos MOBIO, obtuvimos un porcentaje de error promedio de 16.43 %. Para la base de datos NIST2010, logramos un promedio de error de 16.61%. La base de datos NIST2010 considera varias condiciones. De esas condiciones, la condición de entrevista para entrenamiento y prueba mostró el mejor error con 10.94 %, seguida por la condición de llamada telefónica en entrenamiento y llamada telefónica en prueba con 13.35%.
Palabras clave. Verificación de hablante y autenticación.

1 Introduction

During the past three decades we have experienced a rapid evolution of technology. An example of this evolution is in the way in which transactions are performed. Internet, mobile devices and telephone are nowadays the most common means for these transactions. The spreading of the technology has transformed the security issue into a necessity.

Speech has shown to be a promising option to provide security by implementing authentication systems. The most important reasons to employ the voice signature are: a) speech is the most common way of human communication, b) to speak to a device is also considered a non-

invasive technology (the interaction with the user just needs an utterance).

The secure authentication of users has become an important security issue that can be solved through speaker recognition. This field is divided into two main tasks: speaker verification and speaker identification. Speaker verification (SV) verifies if the speaker utterance belongs to a target speaker or not. Speaker identification (SI) verifies the identity of a speaker among a set of prospect users.

In this research we focus on SV systems. The speech signal inputs the system, which produces an acceptance or rejection result. In the ideal case, if the speech belongs to the target model, the result is tagged as accepted. If the speech belongs to an impostor, the result is tagged as rejected. The system presented here shows the evolution of our SV system, named TECHila2. It is based on the Gaussian Mixture Model (GMM) and the Maximum likelihood (ML) framework.

Section 2 describes the architecture of the SV System. Section 3 presents the characteristics of the databases used to accomplish the experiments. Section 4 compares the results obtained in *NIST Evaluation* [SRE, 2010] and *MOBIO Competition* [Marcel *et. al.* 2010]. Section 5 opens the discussion and suggests special trends for future research.

2 SV Architecture

Speaker Verification is to correctly accept or reject the identity of a user given a speech segment and a target speaker model. To verify can be considered a classification problem based on hypothesis testing. We want to verify if the speech signal $\mathbf{X}$ belongs to a target user S. Two errors can be easily considered:

- Type I: known as false rejection, meaning that signal $\mathbf{X}$ is incorrectly rejected being a target speaker.
- Type II: known as false acceptance, meaning that the signal $\mathbf{X}$ is incorrectly accepted being an impostor.

To analyze this errors we first need to review hypothesis testing [Wald ,1947; Duda and Hart, 1973]. In the case of SV, the score obtained for every trial follows a hypothesis test framework, where the null hypothesis H_S accepts the speaker as legitimate and the alternative hypothesis $H_{\bar{S}}$ rejects him/her. Under this framework, the score is given as the likelihood ratio of two models: target-model and anti-model. The decision between the two hypotheses is defined as follows:

$$\theta(\mathbf{x}) = \frac{p(\mathbf{x}|\lambda_S)}{p(\mathbf{x}|\lambda_{\bar{S}})} \begin{cases} > \tau \text{ accept } \lambda_S \\ < \tau \text{ accept } \lambda_{\bar{S}} \end{cases} \quad (1)$$

where $p(\mathbf{x}|\lambda_S)$ and $p(\mathbf{x}|\lambda_{\bar{S}})$ are the probability density functions of S and $\bar{S}$ respectively (also known as likelihoods).We accept the hypothesis λ_S if θ, the score, is greater than τ. We reject (tag as impostor $\bar{S}$) the hypothesis λ_S if the score, θ is less than τ. For the purpose of this research we employ a gender-dependent set of impostors, usually named UBM (universal background model or anti-model), for all S target users.

If Equation 1 is transformed into the log domain:

$$\theta(\mathbf{x}) = \log\left(p(\mathbf{x}|\lambda_S)\right) - \log\left(p(\mathbf{x}|\lambda_{\bar{S}})\right) \quad (2)$$

where $\log\left(p(\mathbf{x}|\lambda_S)\right)$and $\log\left(p(\mathbf{x}|\lambda_{\bar{S}})\right)$ are the log-likelihoods corresponding to the target model S and the anti model $\bar{S}$.

Finally, we just need to correctly estimate the log-likelihoods by a modeling approach. To achieve this objective the SV process encompasses three main stages: Feature extraction, Speaker modeling and Evaluation. In the *feature extraction*, the system computes relevant information vectors from each user. In the *speaker modeling*, the speaker data is used to build a model. Finally in the *evaluation*, the models are tested using several trials and a reliable statistic of the system performance is

obtained. In the next sections we will describe in detail the different parts of the system.

2.1 Feature Extraction

The feature extraction includes the following stages: MFCCs computation, feature normalisations and frame removal.

2.1.1 MFCCs Computation

The computation of the MFCCs is composed of several stages [Davis and Mermelstein, 1980; ETSI 2000]. The first stage is the pre-emphasis followed by the short-time Fourier analysis on an overlapping hamming window. We can extract either the power or the magnitude of the Fourier coefficients. Afterwards, a filterbank transforms this signal into a smooth spectrum representation (close to the envelope of the speech signal). The filterbank output is then transformed to the log-domain. Finally, we apply the DCT to decorrelate and produce the cepstral coefficients [Bogert *et. al.* 1963]. The filterbank can be linearly spaced, and the resulting coefficients are named Linear Frequency Cepstral Coefficients. However, the most common computed are the MFCCs. They follow the mel scale that resembles the way a person hears. To emphasize the dynamic features of the speech in time, the time derivative (Δ) and the time-acceleration (Δ^2) are usually computed. It is common to compute 12 MFCC, one Energy coefficient and its corresponding (Δ)and (Δ^2). However, recent studies have shown that using more than 12 MFCCs can give better results [Martin and Greenberg, 2009; Burget *et. al.* 2009].

2.1.2 Feature Normalizations

Normalizations at this stage are implemented to reduce the effects of the noise and the channel distortion. For instance, the cepstral mean subtraction (CMS) [Furui, 1979] is a blind deconvolution that comprises the subtraction of the utterance mean of the cepstral coefficients from each feature. In the same way, the variance normalization (CVN) [Viikki and Laurila, 1978] is also applied. Hence, the new features will fit a zero mean and variance one distribution. Another well-known feature normalization is RASTA (Relative Spectra) [Hermansky *et. al.* 1992]. While CMS focus on the stationary convolution of the noise due to the channel, RASTA reduces the effect of the varying channel; it removes low and high modulation frequencies. The three of them are commonly used in the SV architecture.

2.1.3 Feature Warping

Another important normalization at the feature stage is the feature warping. It belongs to the Gaussianisation methods [Pelecanos and Sridharan, 2001; Chen and Gopinath 2001]. The underlying concept in this normalization scheme is that every spectral attribute (cepstral coefficient in our case) is normally distributed across time, but the transmission channel distorts such distribution. 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 set has a normal distribution. Feature 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.

2.1.4 Feature Frame Removal

Frame removal is based on the idea that low energy frames do not provide information about the identity of a person. The frames' log-energy of each utterance are modeled by a three-component GMM. ω_1 corresponds to the highest weight of the rightmost Gaussian, ω_2 to the middle Gaussian, and ω_3 to the leftmost Gaussian. According to this model every frame log-energy is labeled as high if it belongs to the rightmost Gaussian; medium if it belongs to the middle Gaussian; and low if it belongs to the leftmost Gaussian [Petrovska-Delacrétaz *et. al.,* 2007]. The following equation is used to determine which frames can be extracted.

$$N = \omega_1 + (g * \alpha * \omega_2), \tag{3}$$

where g is a value between 0 and 1, and α is an heuristic weighting parameter. N is the percentage (between 0 and 100) of the frames with highest energy that will be extracted. If an accurate voice activity detector (VAD) or the speech transcriptions to perform speech recognition are not available the frame removal is a suitable solution.

2.2 Speaker Modeling

Most of the current modeling strategies to compute the target model and the anti-model [Reynolds, 1992; Reynolds *et. al.* 2000; Reynolds 1995] are based on GMM (Gaussian Mixture Model) approach. For a d-dimensional feature vector, the multivariate Gaussian Mixture $p(\mathbf{x}|\lambda)$ p.d.f. is defined as:

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} \omega_i N(\mathbf{x}, \mu_i, \Sigma_i) \quad (4)$$

where M are the number of components of the model, $\omega_i$are the weights of each component $\sum_{i=1}^{M} \omega_i = 1$, and $N(\mathbf{x}, \mu_i, \Sigma_i)$ is the probability density function with μ_i mean, and $\Sigma_i$diagonal covariance matrix.

The Expectation maximization (EM) algorithm [Dempster *et. al.* 1977], the acoustic modeling under maximum likelihood criterion, is the leading algorithm for training the GMM. It is used to iteratively calculate the maximum likelihood estimates of the GMM parameters. By using the EM algorithm, the following model λ is computed such that:

$$p(\mathbf{x}|\lambda^{(n+1)}) \geq p(\mathbf{x}|\lambda^{n}) \quad (5)$$

To reach convergence in a shorter time, some systems employ a clustering algorithm as a previous step before the EM algorithm. For instance, a good solution is to first apply k-means algorithm to the data and identify possible clusters and centroids (mean vectors of a Gaussian distribution), instead of initialising the algorithm with random vectors.

By applying the EM algorithm to a target-independent set of data a GMM model is computed. This model is also known as anti-model or Universal Background Model (UBM). It embraces the characteristics of all the data vectors of the users not belonging to the target set. In most cases it is gender-dependent.

Due to the small amount of data available to produce target speaker models, the anti-model is adapted by using the maximum a posteriori (MAP) algorithm [Gauvain and Lee, 1994]. It follows the decision rule,

$$\zeta = \arg\max_{\lambda,\vartheta} p(\mathbf{x}|\lambda;\vartheta)\, p(\lambda;\vartheta) \quad (6)$$

where $\zeta = (\lambda, \vartheta)$ and ϑ is the hyper-parameter of the distribution of λ. Equation 6 is also solved under the ML criterion using the EM algorithm.

Note that for MAP, three issues should be considered,

- the definition of the prior densities that are the basis of the target models,
- the estimation of the prior densities,
- solution of the MAP by EM algorithm.

For the purpose of SV the prior densities are based on the GMM UBM model. Then, the UBM distributions are adapted and transformed into the target probability distributions. The new parameters are estimated by using the EM algorithm.

2.3 Evaluation

As shown, the score obtained for each trial follows a hypothesis test framework. The score evaluates the log-likelihood ratio of two models: target-model and anti-model (see Equation 2).

After calculating a final score, an accept/reject decision is obtained. The task of giving a decision is still a challenge. It can be computed in two ways: gender dependent and gender independent (both of them use a development database to tune the desired threshold). The gender dependent approach employs a priori information of each target to set a target-dependent threshold. In the gender independent approach

the scores are first normalized and a single threshold is calculated. If the system only had a single target (costumer), it would be simple to determine the acceptance/rejection based on the score. However, most systems have multiple targets, therefore it is convenient to scale (or normalize) the scores so that they are comparable across multiple targets. ZNORM [Mariéthoz and Bengio, 2005], TNORM [Navratil and Ramaswamy, 2003], and ZTNORM, among others, are used for this purpose [Petrovska-Delacrétaz *et. al.,* 2007].

Note that the threshold is a trade off (see Figure 1). Let's suppose that the scores of the target speakers follow a Gaussian distribution. The same is valid for the impostor speakers. The full picture points out two overlapping Gaussian distributions (the right one represents the target speakers and the left one the impostors). One approach is to set the threshold as high as possible, but that will produce few false acceptances and many false rejections (very secure). On the other had, if the threshold has a low value, there will be many false acceptances and few false rejection. Neither of them is desirable. A good tradeoff that can minimize the threshold loss is appropriate. According to the last necessities, this threshold should also adapt to the different channels.

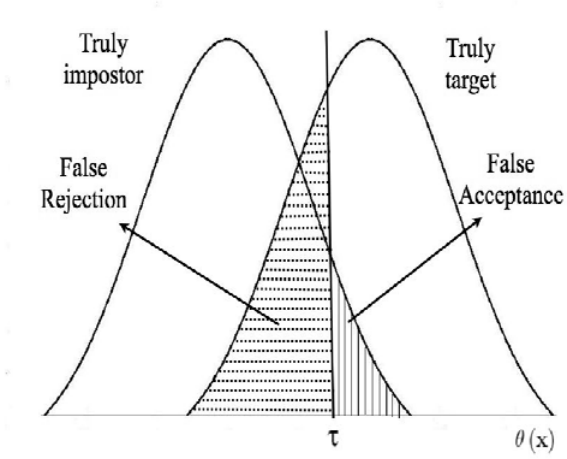

Fig. 1. Threshold

3 Databases

Two databases are discussed: MOBIO [Marcel *et.al., 2010*] and NIST2010 [SRE, 2010].

3.1 MOBIO

MOBIO [McCool and Marcel, 2010] is a large cellphone bi-modal (audio/visual) database. It includes both the speech and face data divided in six sessions. From this database, we extracted just the speech signal from the mp4.

The database was recorded as follows:

A) Set responses: the users were asked questions such as: What is your name? (around 7 seconds)
B) Read Speech from a paper: the users read 3 fixed sentences (maximum 30 seconds)
C) Free speech: the users provide five to ten second answers to random questions.

All this recordings were performed in several sessions. The evaluation is divided in three phases: development, enrollment and testing. The development includes the generation of a UBM and is composed of set of responses (A). The enrollment uses the data in the set (B) of response questions. The testing is composed by the free speech answers, set (C).

3.2 NIST2010

We focused on the core-core evaluation by NIST 2010 [SRE, 2010], which is a defined task.

The UBM was generated using NIST2004 database. The train and the test are composed of either one two-channel telephone conversation of approximately five minutes total duration, with the (target or proposed trial) channel designated previously or a microphone conversation segment of three to fifteen minutes involving both the interviewee (target speaker) and an interviewer. There were over 570,000 trials (female and male).

The data included various levels of vocal effort (low, normal and high vocal effort) from speakers of previous evaluations. The data also presented various types of microphones (seven) and both conversational and interview sessions. The studied common conditions are described in Table 1.

The experiments were conducted using cross-conditions of the above elements. The competition requirements allowed a cost of false rejection of 1 and the target probability of 0.001. The cost function caused the minimum cost operating point to be between 0.01% and 0.1% false alarm percentage.

Table 1. Table presenting the common conditions

Train condition	Test condition	Channel
Interview	interview	same microphone
Interview	interview	different microphone
Interview	normal vocal effort phone call	telephone
Interview	normal vocal effort phone call	microphone
normal vocal effort phone call	normal vocal effort phone call	different telephone
normal vocal effort phone call	high vocal effort phone call	different telephone
normal vocal effort phone call	high vocal effort phone call	microphone
normal vocal effort phone call	low vocal effort phone call	different telephone
normal vocal effort phone call	low vocal effort phone call	microphone

4 Experiments and Results

4.1 MOBIO and NIST 2010 Common Processing

At the beginning of our study, our system was entirely built to follow NIST2008 competition requirements. From this basic scheme more complex variants will be constructed or tuned depending on the application. The basic architecture employed in terms of signal processing is described in the following paragraphs.

The speech signal is treated as an 8 kHz signal. Subsequently, a 25 ms analysis overlapping Hamming window, 10 ms frame rate, and pre-emphasis coefficient of 0.97 was applied. Afterwards, the features are extracted (as explained in Section 2.1) followed by a GMM training approach. The experiments were performed independently for each gender. A gender-dependent and target-independent 512-mixture GMM anti-model [Lee, 1997] was trained from a pool of the corresponding development database. The EM (expectation maximization) algorithm was used to obtain the maximum likelihood estimates of the GMM parameters. TECHila2s's implementation of the EM algorithm for GMM uses the MPI (Message Passing Interface) environment to take full advantage of parallel computing infrastructure.

The GMM is first initialized using the K-means algorithm to obtain a set of 512 centroids. By using the k-means algorithm, the convergence of the EM is known to be faster. However, it is always important to check that the local bounds are not very restrictive, so that EM can make a satisfactory estimation. The EM is then repeated after the model had converged (about 3-5 iterations). Target-dependent models were then obtained with a traditional MAP (maximum a posteriori) speaker adaptation [Gauvain and Lee, 1994]. The score obtained for every trial follows the hypothesis test framework. The score is given by the log-likelihood ratio of two models: target-model and antimodel.

TECHila2 was, then, modified to fulfill MOBIO competition and NIST evaluation requirements. The next sections show how these changes were addressed.

4.2 MOBIO

For MOBIO, the signal was down sampled from 48kHz to 8kHz (with this adjustment it can be treated as telephone speech signal). Two approaches were followed:

- **System 1**: Feature vector of 33 attributes: 16 static Cepstral, 1log Energy, and 16 delta Cepstral coefficients, A single file from each target user (the average time of these utterances is 7 seconds) was used for training phase.
- **System 2**: Feature vector of 49 attributes: 16 static Cepstral, 1log Energy, 16 delta Ceptral coefficients, and 16 double delta Ceptral coefficients. The complete set of target files were used to compute the models.

A gender-dependent and target-independent 512-mixture GMM anti-model [Lee, 1997] was trained employing a pool of the MOBIO speech database (4893 audio files for male, 1764 for female). Target-dependent models were then obtained employing MAP.

The results obtained using our approaches on MOBIO database are summarized in

Table 2. Table presenting the final results (EER) on the Test set for the MOBIO database

	Male	Female	Average
System 1	20.55%	25.23%	22:89%
System 2	15.45%	17.41%	16:43%

The main purpose of system 1 is to test MOBIO database in the worst scenario (baseline). The first constraint is to reduce the relevant speaker information by down sampling the signal to suit a common telephone processing. The second one is to compute just the first delta coefficients to have a rapid training. The third one is to use just one file to test our MAP implementation. Sometimes the speakers are not cooperative and we should address an adaptation with the minimum data. For system 2, we considered the down sampling, but included the double delta coefficients and the complete pool of available training files. As shown, the system performance clearly depends on the amount of data used in the adaptation. Moreover, there is a significant improvement when the double deltas are appended to the feature vector.

Considering that the best results obtained were for system 2.We used this scheme to improve our architecture for NIST2010 evaluation.

4.3 NIST2010

For the case of NIST2010 database, the experiments were conducted using 49 attributes: 16 static Cepstral, 1log Energy, 16 delta Ceptral coefficients, and 16 double delta Ceptral coefficients. A gender-dependent and target-independent 512-mixture GMM anti-model model was trained from the core-core of NIST-SRE 2004 database.

For every iteration of the EM algorithm, TECHila2 randomly polls 25% of the training tokens (belonging to the core-core NIST2004 database), corresponding approximately to 3 hours of speech. The results obtained for NIST2010 are summarized in Table 3.

Table 3 . Table presenting the final results (EER) on the Test set for NIST 2010

	Female	Male	Average
1 Interview interview same mic	13.5%	8.39%	10.94%
2 Interview interview different mic	23.47%	17.29%	20.38%
3 Interview nvephonecall tel	18.42%	16.24%	17.54%
4 Interview nvephonecall mic	17.27%	13.07%	15.17%
5 nvephonecall nvephonecall different tel	17.18%	15.86%	16.52%
6 nvephonecall hvephonecall different tel	22.95%	20.22%	21.58%
7 nvephonecall hvephonecall mic	23.43%	19.49%	21.24%
8 nvephonecall lvephonecall different tel	13.40%	13.30%	13.35%
9 nvephonecall lvephonecall mic	12.71%	12.65%	12.78%
Average	18.03%	15.16%	16.61%

The results in Table 3 are consistent with the ones computed in MOBIO competition. We observe that the EERs are better for male than for female for all conditions. Note that the interview training and testing condition showed the best result of 10.94%, as well as the phone call training phone call testing condition of 13.35%. However, when there is a mismatch between conditions, the EER increases. Another interesting observation is for instance, row 1 and 2 where the EER is doubled just by changing the microphone. This occurs due two facts: the anti-model is trained just with a mixture of telephone data and the lack of an algorithm such as joint factor analysis to model the channel variability.

4.4 Infrastructure

All experiments were conducted under an autonomous Beowulf cluster with 20 CPUs i686 3GHz, 1Gbps LAN, 7TB storage. And we used the following software: SGE, Matlab, Python, Perl,

GNU-Linux. Our iterative algorithms, such as EM, MAP, k-means, emulate MPI environment to take full advantage of parallel computing infrastructure.

5 Discussion and Conclusions

The results for both databases follow the same trend. We can observe that increasing the number of MFCCs and including the delta and double delta improve the results. Moreover, most of the best results were obtained when there is no mismatch between conditions. We will consider further normalisation techniques (such as Znorm) to obtain better results as part of our future work. In addition, we should include an algorithm such as joint factor analysis that can handle the mismatch between training and testing channel conditions.

Each of the databases has problems to solve. On the one hand, we have found that the main issue of MOBIO database is how to treat the few number of pure speech frames that can be extracted from the phrases. This issue caused problems for the feature warping and frame removal algorithms, because they needed at least 3 second speech utterances. We should address this problem by reducing the number of GMM components. On the other hand, the main challenge of NIST2010 database was the number of trials to test the target models, causing an increase in the computing time.

Due to computation requirements the configuration of our cluster was a challenge too. Although the implementation was carefully done to avoid waste of computation (easily done in Matlab), we realized that our system needs a faster implementation and a comparable result.

References

1. **Bogert, B. P., Healy, M. J. R. & Tukey, J. W. (1963).** The quefrency alanysis of time series for echoes: Cepstrum, pseudo-autocovariance, cross-cepstrum and saphe cracking. *Symposium on Time Series Analysis*, New York, USA, 209–243.
2. **Burget, L., Fapso, M. Hubeika, V., Glembek, O., Karafiát, M., Kockmann, M., Matějka, P., Schwarz, P., & Černocký, J. (2009).** But system for nist 2008 speaker recognition evaluation. *Interspeech 2009.* Brighton, Great Britain, 2335–2338.
3. **Chen, S. S., & Gopinath, R. A. (2001).** Gaussianization. In Todd K. Leen, Thomas G. Dietterich,Volker Tresp (Eds.). *Advances in neural information processing systems 13*, (423-429), Massachusetts, USA, The MIT Press.
4. **Davis, S. & Mermelstein, P. (1980).** Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357-366.
5. **Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977)**. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1), 1–38.
6. **Duda, R. O.** & **Hart, P. E. (1973).** *Pattern classification and scene analysis*. New York: Wiley.
7. **Furui, S. (1981).** Cepstral analysis techniques for automatic speaker verification. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29 (2), 254–272.
8. **Gauvain, J. L. & Lee, C. H. (1994).** Maximum a posteriori estimation for multivariate Gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 2 (2), 291–298.
9. **Hermansky, H.**, **Morgan, N.**, **Bayya, A.**, **& Kohn, P. (1992).** RASTA-PLP speech analysis technique. *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP-92*, San Francisco, USA, 1, 121–124.
10. **Lee, C.-H. (1997).** A unified statistical hypothesis testing approach to speaker verification and verbal information verification. *Proceedings COST, Workshop on Speech Technology in the Public Telephone Network: Where are we today?,* Rhodes, Greece, 63–72.
11. **Marcel, S., McCool, C., Matejka, P., Ahonen, T., Cernocky, J. (2010).** *Mobile biometry (mobio) face and speaker verification evaluation.* Retrieved from http://publications.idiap.ch/index.php/publications/show/1848
12. **Mariéthoz, J. & Bengio, S. (2005).** A unified framework for score normalization techniques applied to text-independent speaker verification. *IEEE Signal Processing Letters*, 12 (7), 532-535.
13. **Martin, A.F. & Greenberg, C.S. (2009).** NIST 2008 Speaker Recognition Evaluation: Performance Across Telephone and Room Microphone Channels. *Interspeech 2009*, Brighton, United Kingdom, 2579-2582.

14. **McCool, C. & Marcel, S. (2010).** *Mobio database for the ICPR 2010 face and speech competition*. Retrieved from http://publications.idiap.ch/index.php/publications/show/1757
15. **Navratil, J. & Ramaswamy, G.N. (2003).** The awe and mystery of t-norm. *8th European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2009-2012.
16. **Pelecanos, J. & Sridharan, S. (2001).** Feature warping for robust speaker verification. *A Speaker Odyssey-The Speaker Recognition Workshop*, Crete, Greece, 213-218.
17. **Petrovska-Delacrétaz, D.**, **Hannani, A. E.**, **& Chollet, G. (2007).** Text-independent speaker verification: state of the art and challenges. *Progress in nonlinear speech processing*. *Lecture Notes in Computer Science*, 4391, 135–169.
18. **Reynolds, D.A. (1992).** *A Gaussian mixture modeling approach to text-independent speaker identification*. Ph.D. dissertation, Georgia Institute of Technology, Atlanta, Georgia, USA.
19. **Reynolds, D.A. (1995)**, Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17 (1-2), 91–108.
20. **Reynolds, D.A.**, **Quatieri, T. F. & Dunn, R. B. (2000).** Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10 (1-3), 19–41.
21. Speaker Recognition Evaluation. Retrieved from http://www.itl.nist.gov/iad/mig/tests/sre/
22. *Speech processing, Transmission and Quality aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms*. ETSI ES 201 108 V1.1.2 (2000-04), 2000.
23. **Viikki, O. & Laurila, K. (1998).** Cepstral domain segmental feature vector normalization for noise robust speech recognition. *Speech Communication- Special issue on robust speech recognition*, 25 (1-3), 133–147.
24. **Wald, A. (1947).** *Sequential analysis*. New York: John Wiley and Sons.

Leibny Paola García Perera

She is a research assistant of the Computer Science Department at the ITESM, campus Monterrey. She received her BSc. in Electronic Communications in 2000 and M.Sc. in Communications Engineering 2002. She is currently enrolled in a PhD. from the Universidad de Zaragoza, Spain. She has been a research scholar at Georgia Tech 2010. Her main research includes Speech Processing, Speech robustness, and discriminative Modeling methods in Automatic Recognition and Speaker Verification.

Roberto Alfredo Aceves López

He is currently a student of the Master in Electronic Systems and works as research Assistant in the Computer Science Department at the ITESM, campus Monterrey. He received his BSc. in Electronic Systems from the ITESM, campus Monterrey in 2006. His main research areas include Automatic Speech Recognition based on Hidden Markov Models and Speaker Verification.

Juan Arturo Nolazco Flores

He is currently professor and head of the Computer Science Department at the ITESM, campus Monterrey. He received his BSc. in Electronic Systems and M.Sc. in Control Engineering from the ITESM, campus Monterrey in 1986 and 1987, respectively; his MPhil. and his Ph.D. from the University of Cambridge at the UK in and 1991 and 1995. He has also been research scholar at Marburg, Germany, 1998, Carnegie Mellon University, USA, in 2000, Mannheim, Germany, 2002, Zaragoza, España, 2004. He is a Member of the Mexican National Research System since 2006. He obtained the Teaching and Research Excellence Award, ITESM, campus Monterrey, 2005 and 2009.His main research areas include Speech Processing, the application of Hidden Markov Models and Pattern Matching methods in Automatic Recognition and Speaker Verification.