Acoustic Models Optimization by Applying AIC/BIC Methods to Mexican Spanish Speech Commands for Planar System Control

[1]Pedro Mayorga O., [2]Ana M. Hernández-Campos., [3]Florencio Meza P., [4]J. Martin Olguin-Espinoza

[1,3]División de Estudios de Posgrado e Investigación. Instituto Tecnológico de Mexicali. Av. Tecnológico s/n. Col. Elías Calles. Mexicali. B.C., México. Tel 01 (686)580 49 80. FAX 01(686) 568-7803

[2]Facultad de Ingeniería. Universidad Autónoma de Baja California. Blvd. Benito Juárez s/n. Col. Insurgentes Este. Mexicali. B.C.

[1]pedromayorga@hotmail.com, [2]ahmxli@yahoo.com, [3]meza@skyworksinc

Paper received on 28/07/08, accepted on 06/09/08.

Abstract. This paper presents a modeling of Mexican Spanish Speech commands with GMM techniques and AIC/BIC methods in an Automatic Speech Recognition (ASR) system for controlling robots. An ASR platform is implemented to recognize Mexican Spanish spoken commands. An embedded microcontroller in the robot (planar position system), will be driven by the computer according to the recognized spoken commands. The Computer acquires speech commands which are processed by our platform using robust Mel Frequency Cepstral Coefficients (MFCC). For these experiments, Gaussian Mixture Modeling (GMM) techniques were used to model commands; these are successful methods in speech recognition context but rarely used in robotics domain. We focus in the optimization of the number of Gaussian components used in order to have an efficient system. We apply two methods: the first one manually reducing mixtures and the second one by applying AIC/BIC.

1 Introduction

There are successful state-of-the-art techniques for speech processing and for speech recognition [1], [2], among them we find GMM, ANN, and HMM models. In speech recognition field, front-ends are used to process voice for speech or speaker recognition goals. In this context, MFCC, PLP (Perceptual Linear Predictive) and others are popular front-ends [1-3].

In Latin America not enough research exists to standardize platforms for speech recognition in Spanish language, for this reason is difficult to find Spanish corpus in order to do experiments. Several universities from Spain such as UPC, UPM have made a lot of work about speech and speaker recognition and they continue doing research for Iberia Spanish. Unfortunately, the evolution of the Spanish language in Latin America (i. e. the Castilian) has taken diverse ways; the people from Chile or from Mexico speak with distinct accent from people who were born and live in Spain. The phoneme concept is related with how the speech sounds are pronounced. The sound associated to a same phoneme has varieties (allophones) depending on

©M. G. Medina Barrera, J. F. Ramirez Cruz, J. H Sossa Azuela. (Eds.).
Advances in Intelligent and Information Technologies.
Research in Computing Science 38, 2008, pp.17 -27

who's speaking, from the region usage, the regionalisms accents. The Mexican Spanish term is referred to the diversity of the Spanish spoken by people in Mexico; the Mexican Spanish is divided in six regions according to the dialectal analysis: North, Centre, West, Coast, Yucatan and Chiapas. Characters or letters are the graphic representation of the spoken sound, and there exist a classification named voiced an unvoiced phonemes; these are a consequence of the vocal cords vibration [4-8]. The Spanish phonemes are simpler than those in other languages like English or French, because their sound is the same as their graphic representation. The vocalic phonemes in Mexican Spanish are: a, e, i, o and u. In others languages these phonemes have variants and they are pronounced in different forms. The vowels sound depends of their location in the word, and also on their neighbours [4-8].

Many biomedical researches works deals mostly with heart signals and arm aid devices rather than systems controlled by voice. Speech Manipulation of robotics arms or wheelchairs is a research topic with many things still to do [9]. Robots can be more useful and easier to manage introducing a speech module. In common language, the commands must sound natural and understandable. It is also important the computational cost because in autonomous systems there are constraints in processing and power consumption, so it is mandatory having a good deal between accuracy and computational requirements. To reach a good level of recognition performance, it is necessary to work with well trained acoustic models, but at same time it is necessary to optimize models for autonomous systems.

In our experiments, we deal with a system, which behaves like a planar x-y system with movements such as: right, left, front, back, etc. Therefore, we are focused in finding the best words that might be useful to use in a system of this kind. For example, the command "*para*" have a clear semantic meaning in the robotic motion context for the Spanish language. At the same time, the ideas were training models in a computationally cost effective way, and optimize them in order to reach our goals.

2 General Architecture of the System

The planar robot controlled by speech commands (PRCSC) maps from human utterances to a set of control signals used to drive a planar system. It also allows a small set of discrete spoken words usable as commands, to set a X-Y position. At the moment we have a corpus with the follow command words: *Arriba, Abajo, Derecha, Izquierda, Adelante, Reversa, Arrancar, Comenzar, Empezar, Inicio, Base* and *Alto*.

In this work one-direction approach is enough because normally a person commands the machine through a voice interface. The system acoustic model should be optimized in order to build an embedded system. Figure 1 shows the sequence to follow in order to capture the voice signal and execute motions corresponding to speech commands. In this X-Y robot with a linear positioning system controller, the basic elements and control techniques for a robotic system are applied.

The last module shown in Figure 1 is related with a planar x-y system . the planar robot architecture has two main purposes. First, we consider the positioning system as a part of the Position control for a planar robot through voice commands

where the robot movements are executed by introducing voice commands. On the other hand, the design has important characteristics as the implementation of a proportional-integral-derivative algorithm to improve the transient response. A trapezoidal trajectory generation routine is included too, the main purpose is keeping velocity and acceleration always known and controlled. In the next paragraphs of this section, the dc-brushed servomotor for use as the robotic stage of a voice recognition system is discussed.

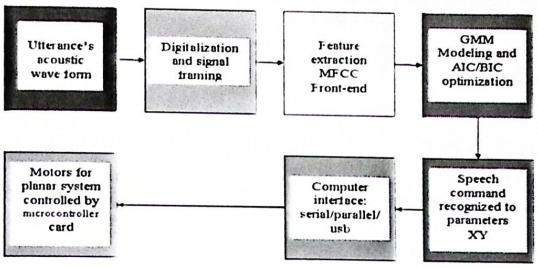

Fig1. Modular representation of PRCSC

A block diagram of the servomotor system is provided in Figure 2. The system is comprised of the following elements: PIC18F2331 microcontroller, RS-232 serial interface, power amplifier (H bridge), brush-DC motor and rotary encoder.

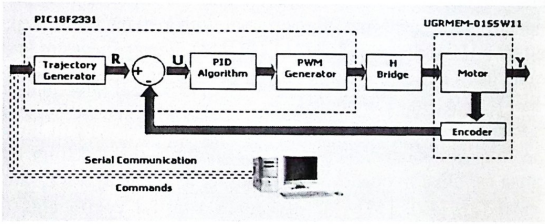

Fig 2. Block diagram of the servomotor

The microcontroller unit (MCU) is responsible for communications with the host system. An RS-232 interface is the main form of communication with the PC. The USART available on the PIC18F2331 is used for this purpose. The user can control the motor movement by sending predefined commands to the controller using the host computer and the graphical user interface.

A PID algorithm is used as a servo compensator and position trajectories are derived from linear velocity ramp segments with a total calculation time for the PID algorithm of 50μs. The servo calculation (compensation algorithm and motion pro-

file) is included in the interrupt routine, which is generated every 800μs by the PWM time base.

The PWM frequency is 20 kHz and the module provides 8 bits of resolution. The torque applied to the motor is determined by the PWM duty cycle. This system uses 50%-null PWM as the D/A conversion technique: 0%, 50% and 100% generate a maximum negative, null or maximum positive output, respectively.
The PWM signal is applied to a Yaskawa RM series motor through a 5A power MOSFET based H bridge, HIP4082 device.
The communication system also lets change configuration parameters from the host computer: these parameters include Position 1, Position 2, Maximum Velocity, Maximum Acceleration, Proportional, Integral, and Derivative Gain.

3. Speech technologies for human-machine interface

Here, we present an overview of the first four modules from Figure 1. This part concerns the capture of the voice signal up to the pattern recognition module, and eventually to transmit codes from the computer to the planar system to be interpreted as motions tasks.

3.1. Speech Parameterization

The voice signal was digitalized taken 8000 data by second. A preemphasis filter of 0.97 is applied to enhance the high and low frequencies of the spectrum, which are generally reduced by the microphone and the digitalization process [3]. The speech signal is segmented into 20 ms frames with a 10 ms frame step; each frame is multiplied by a Hamming window [3], [10], [11]. Power spectrum is computed with a 512 points fast Fourier transform (FFT), because the spectrum is symmetric only the first half is kept (first 256 points).

So as to reduce the spectral vectors quantity, the envelope of the previously obtained spectrum is obtained. Therefore, we multiply the spectrum by a Mel filterbank (human ear alike frequency scale) in order to get an average value by each frequency band [3], [12], [13]. Mel-frequency cepstral coefficients (MFCC) were extracted applying the discrete cosine transform to the log-energy outputs of mel-scaling filter-bank [14].

Specifically, in our case we carried out experiments with 39 acoustic parameters, these parameters were composed by 13 MFCC features, their first and second derivates.

3.2. Pattern Recognition using GMM Modeling

This section describes the Gaussian mixture model (GMM); this model was popularized by Reynolds works [13], [15] in the speaker identification context. Here, we use GMM first as a classification method and command recognition eventually: the GMM models are applied over clusters composed by their MFCC acous-

tic vectors. A Gaussian mixture density is a weighted sum of M component densities, as depicted by the following equations [16]:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} m_i b_i(\vec{x}) \tag{1}$$

With

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\Sigma_i^{-1}(\vec{x} - \vec{\mu}_i)\right\} \tag{2}$$

Where $\vec{x}$ is a D-dimensional random vector ($\vec{x}$ is a MFCC front-end. and D=13 in our case). The i-th Gaussian density is noted as $b_i(\vec{x})$, with mean vector $\vec{\mu}_i$, covariance matrix $\sum$i and mixture weights mi; where i=1, .., M are the component densities, we carried out experiments with values from 16 to 4 for M [13], [15].

A GMM is completely represented with three parameters: mean vectors, covariance matrices, and mixture weights. [22]. In this text, each command is represented and designated as a model λ [13]-[16].

$$\lambda = \{m_i, \vec{\mu}_i, \Sigma_i\} \quad i=1,...,M \tag{3}$$

Model is trained with the Expectation-Maximization (EM) algorithm in order to estimate GMM's parameters. The classification is done by estimating the probability of each class given the observation, and the class that gives the highest probability is chosen as the result [14], [17]. In a group of I commands represented by GMM's λ_1, λ_2,.., λ_I, the recognition rule is:

$$C = \arg_{1 \le k \le I} \max p(X \mid \lambda_k) \tag{4}$$

Model is trained with the Expectation-Maximization (EM) algorithm in order to estimate GMM's parameters. The classification is done by estimating the probability of each class given the observation, and the class that gives the highest probability is chosen as the result [14], [17].

The *Mexican Spanish voiced phonemes* are essentially: a, e, i, o and u. From the point of view of speech recognition, this is important because the energy and crossing zero rate of the voiced phonemes are distinguished from other phonemes. This principle make less difficult to detect the start and the end of the words necessary to obtain better GMM models. Although GMM is a successful method in speech recognition domain, there is no consensus on how to compute the optimal number of mixture components in some specific problem [18-19]. In this section, two approaches will explain how to overcome the difficulty of optimization of the mixture number components in GMM models.

3.2.1. Akaike information criterion (AIC). This model selection technique penalizes the model based on its complexity, as it shows in the equation:

$$AIC(\lambda) = -2 \log p(X|\lambda) + 2k \tag{5}$$

In this equation, log p(X|λ) is the log of the probability of X given λ, and k is the number of parameters in the model λ [18-19]. In this method, as the number of components in the GMM model increases log p(X|λ) and k increases. The model selected will be that which have the lowest AIC score.

3.2.2. Akaike information criterion corrected (AICc). The AIC method is not well adapted in cases where the number of input vectors is small relative to k, here AIC tends to have a negative result, to cope with this problem a variant of AIC method was developed, this technique was named AICc.

$$AICc(\lambda) = -2\log p(X|\lambda) + 2k\left(\frac{n}{n-k-1}\right) \quad (6)$$

In this equation, n is the number of input vectors. However, a modification of AICc technique results is the Bayes Information Criterion [16].

3.2.3. Bayes Information Criterion (BIC). In this equation n is the number of input vectors, and k the number of parameters, as a result, the model that minimizes BIC will be selected.

$$BIC(\lambda) = -2\log p(X|\lambda) + k\log(n) \quad (7)$$

The recognized command (the command related to one action) is translated into digital signals (x-y parameters) that will be transmitted to the microcontroller.

4. Experiments and Results

The experiment was done using the GMM technique, and we make especial emphasis in optimize the number of components Gaussian densities for the models. We trained models with values for M ranging from 16 to 2 mixtures, and we measured the accuracy recognition for each experiment. The models were trained using 40 wav files and evaluated with 30 wav files by each command.

4.1 Accuracy Recognition using Models Composed with 16 to 2 Gaussian Mixtures

In typical speech recognition platforms, 16 and more mixtures are used [1, 3, 9, 20, 22], but in the case of speech command applications with a few words, is possible to work with less mixtures. In our experiments we found that it is possible to train GMM models with 4 Gaussian mixtures (97.3% recognition), and the lost of accuracy recognition is not significant (99% for 16 or more mixtures in other platforms) [1, 3, 9, 20, 22].

The command corpus and the results were divided in two groups in order to remark certain details. The criterion used to divide the results was according to the number of phonemes in the command word. The first group of commands is com-

posed by short words: *abajo, arriba, inicio, base, casa and alto*. The second group is composed with: *derecha, izquierda, adelante, reversa, arrancar, comenzar and empezar*.

In Figure 3 the horizontal axis shows the number of Gaussian used to train the models of the command, whereas the vertical axis represents the accuracy percentage of command recognition; the result was averaged over all commands. The best results were obtained with 16, 13 and 6 mixtures; even though the best trade-off in terms of the number of Gaussians by accuracy was 5. So, we optimized the GMM models using 5 densities and got 96.67 % of recognition accuracy. This is important because in robotics and embedded systems, most of the time, are autonomous systems with limited resources.

Table 1. Recognition percentage for each command according to the number of Gaussians

	Number of gaussians							
Comman	16	14	12	10	8	6	4	2
Abaj	100	100	100	100	100	100	100	100
Arriba	100	100	100	100	100	100	100	96
Derecha	100	97	97	97	97	96	100	96
Izquierda	100	100	100	100	100	100	100	97
Adelante	100	100	100	100	100	100	100	97
Reversa	97	97	97	96	96	96	96	84
Arrancar	100	100	100	100	100	96	97	84
Comenzar	100	100	100	100	100	96	100	86
Empezar	100	100	100	100	100	100	96	100
Inicio	100	96	100	100	100	100	100	100
Base	96	96	96	96	96	96	96	83
Alto	93	96	96	96	96	96	83	77
Average	99.00	98.00	98.50	98.33	98.33	97.50	97.33	91.50

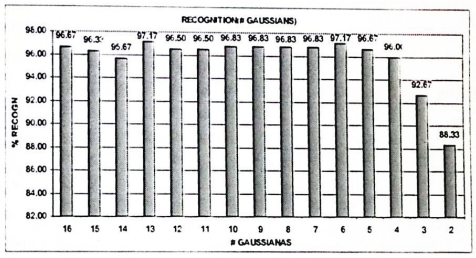

Fig.3. Recognition accuracy by all commands (first group) averaged over the number of Gaussians

In Figure 4 the best results were obtained with 16-12 mixtures; the best trade-off in terms of the number of Gaussians by accuracy was 4. So, we optimized the GMM models using 4 densities and got 98.14% of recognition accuracy. In this case the command words were longer than in Figure 3 (first group). The idea is to have less data; this means it would be better to use shorter words in order to process less information and less computing time, but in our case the best results were obtained with longer words group.

It is well known that voiced signals, are better for the construction of speech models aimed to recognition. Actually, vowels (a, e, i, o, u) signals are well distinguishably from background noise and other undesirables signals. According with the two ideas exposed above, we found better results with words that start and end with vowel phonemes. On the contrary, unvoiced signals are easily confused with noise; therefore, it is not a simple task to correctly detect the start and end of the command word. Consequently, it is logic to consider that words that start and end with vowel phonemes are easier to detect and therefore good for creating better models. Moreover, GMM models are clustering techniques based and do not depend on the sequence of the phonemes involved in the word; thus, the parameters (mean and covariance) are impacted by words with various voiced and vowel phonemes. For this reason, the longest words in our corpus leaded to better models and better results.

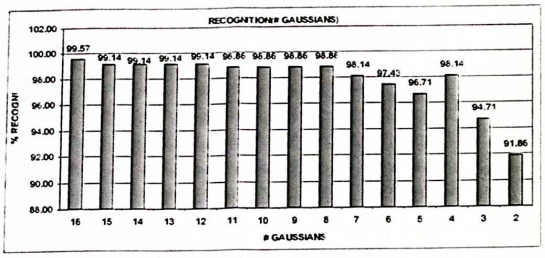

Fig.4. Recognition accuracy by all commands (second group) averaged over the number of Gaussians

4.3. Obtaining the optimum number of GMM by using AICc

The process of obtaining the optimum number of Gaussian mixtures via the conventional test phase already described and implemented is too time-consuming. Using the previously GMMs generated for each command with M mixture components, then using equations (4)-(6) for each mixture components. Given a specific command, the GMM with the lowest information criterion is selected. Since only Gaussians with diagonal covariance matrices are considered, the number of parameters per mixture component was 79 (1 prior + 39 means + 39 variances) [19].
Both AIC and BIC scores decreases as the number of mixtures used in the modeling increases and their values are very similar too. These techniques did not generate information that allows modeling optimization while reducing the number of mixtures. However, the AICc behavior is different, because when the number of samples and the complexity of the model are of the same order of magnitude, the penalty term for AICc will be much larger than AIC [19]. Minimum AICc corresponds to 7 GMM for the spoken command *reversa*, which is a lot better than modeling with 16 GMM as usually and arbitrary have been done. Figure 5 illustrates both, AICc scores and recognition accuracy obtained in the evaluation phase for the *reversa* command. From the left bar graphs is seen that the lowest AICc score belongs to 7 GMM

model and corresponding accuracy recognition is 97%, when the 7 GMM model is used, as shown at the right plot in the same figure.

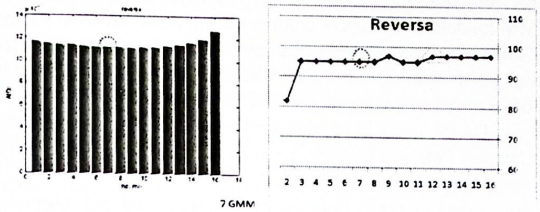

Fig.5. *Reversa* lowest AICc score and corresponding recognition accuracy.

Applying the same analysis to the complete corpus, the two left columns of Table 2 shows mixtures number selection according to its minimum AICc score. Additionally, in order to compare, the two rightmost columns show the mixture number of the GMM model elected according to best accuracy recognition for each command.

Table 2. Mixture Number selected according to minimum AICc score or best accuracy recognition in evaluation phase

	AICC RESULTS		TEST PHASE RESULTS	
COMMAND	GMM With Minimum AICc	Corresponding Recognition Accuracy	GMM With Best Accuracy Recognition	Corresponding Accuracy Recognition
Abajo	8	100	2	100
Adelante	12	100	6	100
Alto	7*	90	15	98
Arrancar	9	100	7	100
Arriba	7*	100	4	100
Atras	6*	87	16	84
Base	5*	98	4	98
Casa	6	95	13	95
Comenzar	9	100	7	100
Derecha	8*	97	10	97
Detener	7	85	14	94
Empezar	9	100	8	100
Frente	7*	100	14	95
Inicio	7	100	2	100
Izquierdo	8	100	3	100
Parar	6*	64	16	84
Reversa	7	95	12	97
MEAN	8	95	9	96

We can observe from Table 2 that AICc criterion is in accordance with the results obtained from the testing phase. Therefore, in order to acquire GMM models we can apply AICc criterion to optimize GMM modeling eliminating the tedious testing phase used to obtain the optimum number of Gaussian mixtures. As expected, results from AICc technique and experimental phase would not be exactly the same; we must to remember that AICc Criterion is a theory approach, and does not consider some real issues as noise and quality of the recording equipment.

5. Conclusions

This work presented the architecture of our experimental planar system which employs GMM models to recognize commands utterance. In order to know how acceptable the AICc criterion performs; we carried out experiments to obtain the number of optimal densities in GMM models and at the same time we computed the optimum number with the AICc approach. After we compared results, we found a very good similarity. On the other hand, this kind of methods makes possible the reduction of computation time and memory resources; consequently, a very important issue for autonomous or robotics applications.

Finally, for command words from our corpus, we found that it is possible to train GMM models with as less as 4 Gaussians mixtures with accuracy recognition up to 97.3% of average. At the moment, the experiments were implemented in our platform in a PC to process information; in the future it will be implemented as an embedded system using DSP processors seeking the advantage of an autonomous system.

References

1. P. Mayorga-Ortiz et al (2003). Audio Packet Loss over IP and Speech Recognition, ASRU IEEE 2003 (Automatic Speech Recognition & Understanding). St. Thomas, Virgin Islands, USA, Nov. 1- Dec. 4, pp. 607-612.
2. R. V.Cox et al (2000). Speech and Language Processing for Next- Millennium Communications Services, Proc. of the IEEE, Vol 88, No. 8, August 2000.
3. R. Bimbot et al (2004). A Tutorial on Text_Independent Speaker Verification. Eurasip Journal on Applied Signal Processing. Vol. 4, April 2004. pp. 430-451.
4. J. Moreno De Alba et. Al (1994). Atlas Lingüístico de México. Tomo I. Fonética. El Colegio de México. México, 1994, 222 p.p.
5. P. M. Butragueño (2002). Más sobre la evaluación global de los procesos fonológicos: la geografía fónica de México" (en Variación lingüística y teoría fonológica. México: El Colegio de México, 2002, pp. 63-104.
6. A. Espinoza (2006). Variación del segmento /-s/ en El Ciruelo, Oaxaca (para el Coloquio "Fonología instrumental: patrones fónicos y variación lingüística"), El Colegio de México, 23 a 27 de octubre de 2006.
7. J.Serrano (2000). Contacto dialectal (¿y cambio lingüístico?) en español: el caso de la /ʧ/ sonorense (en Estructuras en contexto. Estudios de variación y cambio.) Ed. P. Martín. México: El Colegio de México, 2000.
8. P. Henriquez (1991). Observaciones sobre el español de América, Revista de Filiología Española, VII (1991), p.p. 357-390.
9. J. A. Bilmes et al (2006). The Vocal Joystick, ICASSP 2006, Toulouse France, May 14-19, 2006.
10. L. Rabiner and J Biing-Hwang (1993). Fundamentals of Speech Recognition, Prentice Hall, ISBN 0 13 015157, 1993.
11. J. G. Poakis and D. K Manolakis (2006). Digital Signal Processing, Prentice Hall. 4th edition, ISBN 978-0131873742.
12. D. Pearce (2000). An Overview of ETSI Standards Activities for Distributed Speech Recognition Front-Ends. AVIOS 2000: The Speech Applications Conference, San Jose, CA, USA, May 22-24.
13. D. A Reynolds (1992). A Gaussian Mixture Modeling Approach to Text-Independent speaker Identification. Ph. D. Thesis. Georgia Institute of Technology, August 1992.

14. Tuomi et al (2002). "Computational auditory scene recognition." IEEE International Conference on Audio, Speech and Signal Processing, Orlando, Florida.
15. D. A. Reynolds and R.C. Rose (1995). Robust Text-independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Trans. Speech, and Audio Processing, vol. 3, no. 1, pp. 72-83.
16. A. R. Webb (2002). Statistical Pattern Recognition, John Wiley & Sons, Second Edition.
17. A. J. Bilmes (1997). Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical Report, University of Berkeley, ICSI-TR-97-021.
18. P. Mayorga and L. Besacier (2006). Voice over IP and Vocal Recognition, ICEEE 2006 (IEEE) Conference, Veracruz, Mexico, Septiembre 6-8 2006, ISBN:1-4244-0403-7.
19. Y. Kai (2006). Generating Gaussian Mixture Models by Model Selection For Speech Recognition, F06 10-701 Final Project Report, School of Computer Science, Carnegie-Mellon University, autumn 2006.
20. P. Mayorga O., L. Besacier, and A. M. Hernandez C.(2006). Packet Loss and Compression Effects over Vocal Recognition, CERMA2006 (IEEE), Conference, Cuernavaca, Morelos, Mexico, 26-29 Septiembre 2006, ISSN 200691349.
21. P. McKenzie and M. Alder (1994). Selecting the Optimal Number of Componenets for a Gaussian Mixture Model, Information Theory, Proceedings., 1994 IEEE International Symposium on.
22. J.J. Verbeek, N. Vlassis, B. Kröse (2003). Efficient Greedy Learning of Gaussian Mixture Models, Neural Computation, 5(2):469–485.