Científica
Instituto Politécnico Nacional
revista@maya.esimez.ipn.mx
ISSN (Versión impresa): 1665-0654
MÉXICO

2006
Eric Simancas Acevedo / Mariko Nakano Miyatake / Hector Perez Meana
EVALUATION OF GMM BASED SPEAKER RECOGNITION SYSTEMS USING DYNAMIC FEATURES
Científica, año/vol. 10, número 003
Instituto Politécnico Nacional
Distrito Federal, México
pp. 151-156

Científica Vol. 10 Núm. 3 pp.151-156
© 2006 ESIME-IPN. *ISSN 1665-0654.* Impreso en México

Evaluation of GMM Based Speaker Recognition Systems Using Dynamic Features

Eric Simancas-Acevedo[1]
Mariko Nakano-Miyatake[2]
Hector Perez-Meana[2]

[1]University Polytechnic of Pachuca,
Ex-Hacienda de Santa Barbara,
km. 20 Carretera Pachuca-Cd. Sahún,
43830 Zempoala, Hidalgo.
MEXICO.

[2]SEPI, ESIME Culhuacan,
National Polytechnic Institute of Mexico,
Av. Santa Ana 1000, Col. San Francisco Culhuacan
04430 Mexico City.
MEXICO.

email: hmpm@prodigy.net.mx
ericsimancas@upp.edu.mx

Recibido el 26 de mayo de 2005; aceptado el 05 de diciembre de 2005.

1. Abstract

The dynamic features of the LPC-Cepstral coefficients (delta and double-delta cepstral) can be used to improve the performance of a Speaker Recognition System (SRS), because the delta and double-delta represent the derivative of the LPC-Cepstral coefficients with respect to the time (speed and acceleration respectively), allowing that the speaker features become less sensitive to channel and environment distortion. Taking this fact in account, this paper presents an analysis of SRS performance using feature vectors obtained from delta and double-delta LPC-Cepstral coefficients that complements a previously published paper. The evaluation results show that the dynamic features improve the performance of speaker recognition system compared with the baseline SRS which use only the LPC-cepstral coefficients.

Key words: Speaker recognition, LPC-Cepstral, Delta LPC-Cepstral, Double-delta LPC-Cepstral, GMM

2. Resumen (Evaluación del funcionamiento de sistemas de reconocimiento de hablante basados en GMM usando rasgos dinámicos)

Las características dinámicas de los coeficientes LPC-Cepstral (delta y doble delta cepstral) pueden ser usados para mejorar el funcionamiento de un sistema de reconocimiento de hablante, debido a que los coeficientes delta y doble delta representan la derivada con respecto al tiempo de los coeficientes LPC-Cepstral (velocidad y aceleración), lo que permite reducir la sensibilidad de los coeficientes LPC-Cepstral a variaciones del canal. Tomando esto en consideración, en este artículo se presenta un análisis del funcionamiento de un sistema de reconocimiento de hablante usando los coeficientes delta y doble delta de los parámetros LPC-Cepstral, que complementa un artículo publicado previamente por los autores. Los resultados obtenidos por simulación muestran que usando estos vectores característicos, el funcionamiento del sistema mejora en comparación con el istema convencional.

Palabras clave: Reconocimiento de hablante, LPC-Cepstral, Delta LPC-Cepstral, Doble-delta LPC-Cepstral, GMM

3. Introduction

One of the main problems of most Automatic Speaker Recognition Systems (ASRS) is the large degradation of system performance when it is required to operate with speech signals different from those used during the training intervals, i.e. open test. This degradation is due to, among several other facts to, differences on the acoustic conditions such as: noisy telephone lines, environment variations, different microphones, etc. that the system confronts during the training and testing stages.

Several studies show that the SRS archive good recognition performance if the conditions during training are similar to those during the testing [7], [9], [11]. However, one of the most attractive

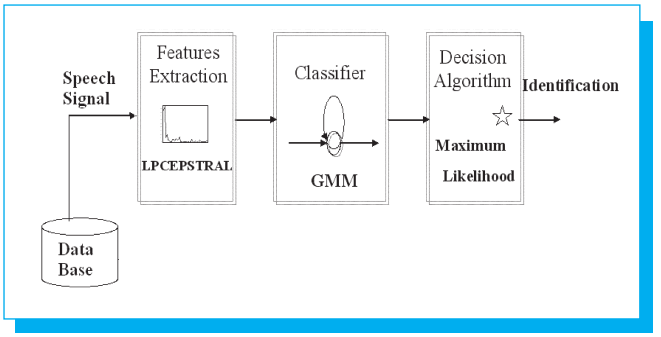

Fig. 1. Proposed speaker recognition system.

applications of the SRS is the long distance speaker identification by using some kind of telecommunication system such as the telephone one in which the communication channel may changes during different time accesses to the system, such as different calls. A difficulty in this case is the fact that the data for training is only a small representation of all the acoustic conditions that the system can met during its operation. Thus, several researchers have proposed the use of more than one speaker feature to obtain a better speaker representation [1], [2] while others intend to suppress those feature variations [11]. Typically, compared with the use of channel normalization techniques such as Cepstral Mean Normalization (CMN) and RASTA filtering, the CMN has shown to be more effective to compensate the variations of the communication channel and to reduce the distortion effects produced for the noise environment [1], [3], [4], because the features derived from the speech spectrum have proven to perform fairly good in most in SRS.

All the techniques used to enhance the speech spectrum try to reduce the distortions introduced in the speech signal by the environment and communication channel, however when a large spectral degradation is present, the speech spectrum can not be enough enhanced. To reduce this problem, dynamic features of the LPC-cepstral coefficients such as the delta and double-delta cepstral coefficients can be used because the delta features are less affected by the channel and environment effects. These dynamic features have been widely used in speech recognition application where have played important roles in the syllable and phoneme perception; and then they can be expected to effectively select the word candidates in a large vocabulary recognition [5], [6]. Taking these facts in account we can develop a robust SRS by enhancing the speaker feature vector using delta and double-delta to analyze the improvement in the accuracy of the speaker recognition system that can be obtained using these features as compare with the conventional system using the LPC-Cepstral coefficients.

4. Development

4.1 Speaker Recognition System

A general speaker recognition system, shown in Fig 1, consists mainly, of three stages: the feature extraction stage, where appropriate information is estimated in a suitable form and size, from the speech signal to obtain a good representation of the speaker features, and the classifier stage, where the speaker models are adapted using the feature vectors, and the decision stage, where the recognition decision is taken.

The SRS system under analysis firstly extracts the features vector from the speaker voice. To this end, firstly estimates the first 16 LPC-Cepstral coefficients using only the voiced parts of speech signals. Then using the estimated LPC-Cepstral coefficients, the dynamic features are estimated to enhance the speaker features vector. Next the estimated dynamical features vector is feed to a Gaussian Mixture Model, GMM, which is used to obtain a representative model for each speaker. In a previous research [7], we developed a speaker recognition system using LPC-Cepstral features vector extracted from the whole speech signal that provides a good performance in closed test. However this SRS can be improved taking in account that the voiced part of the speech signal contains the main information relative to the speaker identity [8], [9], [11]. For this reason in this SRS the features vector will be derived from the LPC-cepstral extracted only from the voiced part speech signal.

4.1.1 Voiced Part Detection

The voiced part detection has a very important roll in the speaker recognition systems because it contains the main characteristics of speech signal, and then suitable feature vectors can be extracted from them. To this end, firstly the pitch period is estimated, using the autocorrelation method [9], because it is well known that in a phrase, the vowels belong to voiced part and the pitch always appears where vowels exist [2]. Thus, suitable voiced part detection can be carried out using the following steps: 1) The speech signal is segmented in overlapping frames, each one of 30ms length with 20 ms overlapping. 2) Each frame is then multiplied by a Hamming window of 240 samples assuming a sampling rate of a sampling rate of 8KHz. 3) The windowed frame is processed using the center clipper method [10] to reduce the noise influence intrinsic within of the signal. 4) Next the autocorrelation of each processed frame is computed. 5) Finally the pitch period, given as the distance between the first and second pick of the autocorrelation function, is estimated using a dynamic threshold. 6) Finally, using the estimated pitch period, the voiced segments are determined [2], [9].

4.1.2. Estimation of LPC-Cepstral Parameters

The features extracted from the speech signal spectrum have shown to provide better performances in the speaker recognition system, specially the LPC-Cepstral, because these have proved to increase the robustness of the speaker recognition systems reducing the problem of speech signal distortion introduced by the communication channel [7]. The computation of the LPC-Cepstral is relatively simple since they can be obtained using a simple recursion after the Linear Prediction Coefficients (LPC) was estimated as follows [7]:

$$c_n = -a_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a_i c_{n-i}, \qquad n>0 \tag{1}$$

where cn is the nth LPC-Cepstral coefficient, ai are the linear prediction coefficients which are obtained by Levinson Durbin algorithm. In this application, 16 LPC-Cepstral coefficients were extracted in each frame.

4.1.3. Dynamic Features

The dynamic spectral features, or simply dynamic feature, contain information that complements the instantaneous spectrum information provided by the static features, such as the LPC-Cepstral coefficients [6]. Widely used dynamic features are the Delta Cepstra and Double-delta Cepstra which represent the derivatives of the time trajectories of the LPC Cepstral coefficients, *i. e*. speed and acceleration respectively. The delta Cepstral feature vector is then a linear regression of the LPC-Cepstral coefficients which can be estimated as follows

$$\mathbf{d}_t = \frac{\sum_{w=1}^{W} w(\mathbf{c}_{t+w} - \mathbf{c}_{t-w})}{2\sum_{w=1}^{W} w^2} \tag{2}$$

where $\mathbf{d}_t$ is the delta feature at time *t,* $\mathbf{C}_t$ is the LPC-Cepstral feature and ***W*** is the window size. Because the eq.(2) relies on the past and future parameter values, some modification is needed at the beginning and at the end of the feature vector as shown eqs. (3) and (4)

$$\mathbf{d}_t = \mathbf{c}_{t+1} - \mathbf{c}_t \quad t < w \tag{3}$$

$$\mathbf{d}_t = \mathbf{c}_t - \mathbf{c}_{t-1} \quad t \geq T - w \tag{4}$$

The double-delta Cepstra,, or acceleration, feature vector can be obtained by applying the same eqs.(2) and (3) substituting now to delta Cepstra feature $\mathbf{d}_t$.

4.1.4. Gaussian Mixture Model (GMM)

In this paper GMM, with 8 Mixtures Gaussian densities, is used to estimate the speaker models using the feature vectors described in section 4.3, and trained using the Expectation Maximization (EM) algorithm. In the GMM model, the features distributions of the speech signal are modeled for each speaker as follows.

$$p(\vec{\mathbf{x}}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{\mathbf{x}}), \tag{5}$$

where

$$\sum_{i=1}^{M} p_i = 1 \tag{6}$$

where $\mathbf{X}$ is a random vector of D-dimension, $p(\mathbf{X}/\boldsymbol{\lambda})$ is the speaker model; p_i is the ith mixture weights; $b_i(\mathbf{X})$ is the ith pdf component that is formed by the ith mean $\boldsymbol{\mu}_i$ and ith covariance matrix, where i = 1,2,3,…,M, and each density component is a D-variants Gaussian distribution given eq. (7).

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \tag{7}$$

The mean vector, $\boldsymbol{\mu}_i$, covariance matrix, $\boldsymbol{\Sigma}_i$, and mixture weights p_i of all density components, determines the complete Gaussian Mixture Density which represents to each speaker. To obtain an optimum model representing each speaker we need to obtain a good estimation of the GMM parameters. To this end, the Maximum-Likelihood Estimation (ML) approach, which is a very efficient method, can be used. Here after the feature vectors to be used during the training period are obtained, the ML method maximizes the Likelihood of the GMM; where for a given of T vectors used for training, $\mathbf{X}=(x_1,x_2,...,x_T)$, the likelihood of GMM can be written as [7], [9].

$$p(\mathbf{X}|\lambda) = \prod_{t=1}^{T} p(\mathbf{x}|\lambda) \tag{8}$$

However it is a no-linear function of the parameters of the speaker model, λ. Thus eq. (8) can not be maximized directly, then the estimation of the ML parameters must be carried out

using an iterative algorithm called Baum-Welch algorithm. The Baum-Welch algorithm, which is the same algorithm used by HMM to estimate its parameters has the same basic principle of the Expectation-Maximization (EM) algorithm, whose main idea is as follows: Beginning with an initial model, λ, a new model $\underline{\lambda}$ is estimated such that

$$p(\mathbf{X}/\bar{\lambda}) \geq p(\mathbf{X}/\lambda)$$

Next, we set the model parameters λ equal to those estimated in the actual stage, λ, so that, the new model becomes the initial model for the next iteration, and so on. Then during the estimation of the GMM parameters, to obtain an optimum model for each speaker, the parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$ and p_i should be estimated iteratively until convergence is achieved. The initial condition $p(i/\mathbf{X},\boldsymbol{\lambda})$ can obtained using the Viterbi algorithm [5]. Subsequently, the parameters p_i, the Mean vector $\boldsymbol{\mu}$ and Variance $\boldsymbol{\Sigma}_i$, which is assumed to be a diagonal matrix, required in the subsequent iterations are given by

$$\bar{p}_i = \frac{1}{T} = \sum_{t=1}^{T} p(i|\vec{\mathbf{x}}_t, \lambda) \tag{9}$$

$$\vec{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\vec{\mathbf{x}}_t, \lambda)\vec{\mathbf{x}}_t}{\sum_{t=1}^{T} p(i|\vec{\mathbf{x}}_t, \lambda)} \tag{10}$$

$$\bar{\Sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\vec{\mathbf{x}}_t, \lambda)\vec{\mathbf{x}}_t}{\sum_{t=1}^{T} p(i|\vec{\mathbf{x}}_t, \lambda)} - \bar{\mu}_i^2 \tag{11}$$

where the likelihood a *posteriori* of the i-th class is given by

$$p(i|\vec{x}_t, \lambda) = \frac{p_i b_i(\vec{x}_t)}{\sum_{k=1}^{M} p_k b_k(\vec{x}_t)} \tag{12}$$

This process is repeated until convergence is achieved. Here the Mixes order and the model parameters previous to the Maximization of the likelihood of GMM, can be different depending on the application.

4.1.5 Decision Stage

In this stage, after the GMM models for each speaker are estimated, the target is to find the model with the maximum likelihood a posteriori for an observation sequence.

$$\hat{S} = \arg\max_{1 \leq k \leq S} P(\mathbf{X}|\lambda_k) \tag{13}$$

where $P(\mathbf{X}/\boldsymbol{\lambda}_\kappa)$ denotes the Gaussian Mixture density given by eq. (5).

4.2 Enhancing the Features Vector

The performance of speaker recognition system can be seriously degraded when the SRS uses a channel communication, such as the telephone lines, because the transfer characteristics of the communication channel, the environment or the microphone characteristics. The LPC-cepstral coefficients have shown to be robust when operates with low speech signal quality, providing a good performance when the same data set are used for training and evaluation, i.e. closed test. However their performance is considerately degraded when the SRS is tested with different data set, i.e. open test. This is because the data used for closed and open test of each speaker may have different acoustic conditions. Thus, the channel normalization techniques have been proposed to reduce the features distortion and keep up the recognition performance. For channel normalization, the Cepstral Mean Normalization (CMN) and RASTA filtering have been proposed which can archive a considerable amount of environmental robustness at almost negligible cost [4]. A comparison of these channel normalization techniques, presented in [1], show that the CMN is better than the RASTA filtering because the RASTA filtering introduces phase distortion in the time domain, while the recognition results for corrected RASTA technique are identical to those of CMN. Another suitable approach is the use of dynamic features, because the higher dynamic features are invariant to any constant bias within the temporal window used for their derivation. In addition they are invariant to slowly time varying linear distortion of the speech signal introduced by the communication channel as well as by the noise environment. Thus the speaker feature vector can be enhanced by using the first and second order regression coefficients, estimated as described in section 4.3.

4.3. Evaluation Results

The system evaluation was performed using two different data set stored from several telephone calls, one for closed test and the other one for open test. To this end, the baseline system using 16 LPC-cepstral coefficients extracted only from the voiced parts is used. The second evaluation is carried out using the CMN technique to enhance quality of feature vector, which has been affected by the communication channel and the environment noise. The third evaluation use a feature vector obtained by concatenating the dynamic features (delta and double-delta cepstral) and the LPC-cepstral features vector. Several conclusions are explained after each evaluation.

Table 1. Result comparisons of system performance with the combination of dynamic features.

Features Vector	LPC-C	LPC-CMN
(Closed set) 6 581 phrases	93.31%	80.72%
(Open set) 3 282 phrases	76.97%	70.88%

The Database used for training and evaluation was provided by the KDD (Telecommunications company of Japan that provides long distance telephone services) that consists of 80 different speakers, with a pronunciations of 10~25 phrases of 2.5-3s in Japanese language for each speaker by telephone, those phrases are repeated for each speaker 6 times with a total number of phrases equal to 10805. For training of the GMM and open test evaluation, 4 repetitions of the same phrases were used, giving in total 7147 phrases. For system evaluation in close test, the same data for training was used and for system evaluation in open test, 2 different repetitions were used which were stored in different times, giving a total of 3658 phrases.

Each repetition was stored by telephone in intervals of 1 month with sampling a frequency of 8 KHz. For the baseline system, in this study, we will use as feature vector 16 LPC-cepstral coefficients extracted from the voiced part of the KDD database because it is very well known that using feature vectors extracted only from the voiced part; the SRS performance improves. In addition there are some other advantages against the use of whole speech signal; for instance it saves storage requirements due to the reduction of the length of feature vector; in some cases until a 50%. Another advantage is the saving time for training. For instance we used 20 iterations in the preliminary experimentation while using only the voiced part it was not necessary more than 10 iterations to achieve even better results.

For the classification stage, a GMM with 8 Gaussian mixtures using a diagonal covariance matrix is used. Experimental results show that the SRS provides a recognition rate of 93.31% for close test and 76.97% for open test. This system performance is compared with other GMM based SRS using several features vectors, such as the Cepstral Mean Normalization (CMN) and the Dynamic features (Delta and Double-delta cepstra).

Table 2. Experimental results of using LPC-Cepstral from voiced part (LPC-C), LPC-Cepstral from voiced part applying delta (LPC-D) and applying double delta (LPC-2D).

Features Vector	LPC-C	LPC-D	LPC-2D
(Closed set) 6 581 phrases	93.31%	94.32%	94.01%
(Open set) 3 282 phrases	76.97%	78.06%	78.40%

4.3.1. Evaluation Results Using Ceptral Mean Normalization

The technique of CMN, use to compensate for channel distortion, is equivalent to a high-pass filtering of the LPC-cepstral coefficients, that is the CMN subtracts the mean the value of the LPC-cepstral coefficients from the original LPC-cepstral coefficients follows

$$CMN_i(n) = C_i(n) - \frac{1}{N}\left(\sum_{n=1}^{N} C_i(n)\right) \qquad (14)$$

where i is the coefficient number and N is the total frames number in which the signal was divided to carried out the feature vectors extraction. The CMN forces the average values of the cepstral coefficients to be zero in both the training and testing set, so it is not possible to apply it only in the testing data set. Thus we need to apply the CMN in both training and testing data to archive good compensation of the unknown linear filtering effects [4].

Table 1 shows the performance of the Speaker Recognition System using as feature vector a 16 elements LPC-Cepstral coefficients vector extracted from the voiced part, compared against the system performance applying the Cepstral Mean Normalization (CMN). Using the Cepstral Mean Normalization to enhance the feature vector, the system performance reduces considerately in close and open test, as shown in the Table1. We observed that the recognition performance for those speakers with bad performance by using only the LPC-cepstral improves when the CMN is used; while for those speakers with in which the SRS achieves a good recognition performance using the LPC-cepstral, the SRS recognition performance considerably degraded after the CMN was used. Some other researchers have also presented similar decreasing performance when CMN is used [1]. This is because the CMN improves

the performance in robust conditions, but decreases when clean speech signal is used or when the communication channel is time invariant. This is because the CMN assumes that the mean value of cepstral coefficients of clean speech has zero mean; which is not entirely correct. In addition, the CMN eliminates the convolution effects but it does not eliminates the additive noise and does not take in account the nonlinearities and non stationary condition of the communications channel [1].

4.3.2. Combining LPC-Cepstrum and Dynamic Features

While begin more robust to the channel effects, the delta features do not perform as well as the LPC-Cepstral in matches condition. Thus the delta features are appended to the LPC-cepstral to enhance the feature vector that will be used to represent the speaker in both the training and testing stages as shown in eq. (15)

$$\vec{y}_t = \left[\vec{c}_t \quad \vec{d}_t\right] \tag{15}$$

Table2 show the SRS performance using a LPC-cepstral feature vector enhanced by the Dynamic features (Delta and Double-delta cepstra). Since the delta feature vector are less affected by the channel variations and noise environment they provide an enhanced feature vector for speaker representation. In Table2 we can see the improvement of the system performance using the dynamic features.

5. Conclusions

A SRS evaluation using dynamical features to improve the performance of Speaker Recognition System affected by different acoustic conditions is presented. These dynamic features used during training and testing stages, allows a reduction of the channel distortion and intra-speaker variations when they store the same phrases in different interval times. In the SRS under analysis the features vectors are extracted only from voiced parts of speech signals allowing a reduction of approximately 50% in the training time. Evaluation results show that using the dynamical features extracted from the voiced parts of speech signal, the SRS performance can be improved also when it operates in open test conditions.

Acknowledgements

The authors thanks to The National Science and Technology Council of Mexico, CONACYT for the support provided during the realization of this research; and to the Japanese Government for the support provided through the JUSST Exchange Student Program. An important part of this work was performed at the Electronic Engineering Department of the University of Electro-Communication, Tokyo, Japan.

6. References

[1] Hema A., Murthy, Francoise Beaufays, Larry P. Heck, «Robust Text-Independent Speaker Identification over Telephone Channels», *IEEE Transactios on Speech and Audio Processing*. Vol.7. No.5, Septiember 1999.

[2] Michael D. Plumper, Thomas F. Quatieri,\ and Douglas A. Reynolds, «Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification», *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No. 5, September 1999.

[3] Johan de Vetch, Louis Boves, «Comparison of Channel Normalization Techniques For Automatic Speech Recognition Over the Thelephone», *Department of Language and Speech*, The Netherlands.

[4] Fu-Hua Liu, Richard M. Stern, Xuedong Huang, Alejandro Acero, «Efficient Cepstral Normalization For Robust Speech recognition», *Department of Electrical and Computor Engineering*, CMU, U.S.A.

[5] Hynek Hermansky, «Exploring Temporal Domain for Robustness in Speech Recognition», *proceedings the 15th International Congress on Acoustic*, Trondheim, Norway, June 1995.

[6] Sadaoki Furui, «On the use of Hierarchical Spectral Dynamics in Speech Recognition», *NTT Human Interface Laboratories*, Musashino-shi, Tokyo, 180 Japan, IEEE, 1990.

[7] Eric Simancas Acevedo, Akira Kurematsu, Mariko Nakano Miyatake and H. Perez Meana. «Speaker Recognition Using Gaussian Mixtures Model». *Lecture Notes in Computer Science, Bio-Inspired Applications of Connectionism*, pp. 287-294, Springer Verlag, Berlin, 2001.

[8] Martin Plsek, Martin Vondra, *Pitch detection in noisy speech recordings*, Brnos University of Technology, Faculty of Electrical Engineering and Communications, Czech Republic.

[9] E. Simancas, M. Nakano Miyatake and H. Perez.Meana, «Speaker Verification Using Pitch and Melspec Information», *Journal of Telecommunications and Radio Engineering*, vol. 56, pp. 46-57, Jan. 2001.

[10] L. R. Rabiner, M. Cheng, A. Rosemberg, C. McGoegal, «A Comparative Performance Study of Several Pitch Detection Algorithms», *IEEE Trans. on Acoustics, Speech, and Signal Processing*, October 1976.

[11] Akira Kurematsu, Mariko Nakano-Miyatake, Héctor Pérez-Meana, Eric Simancas-Acevedo, «Performance Analysis of Gaussian Mixture Model Speaker Recognition Systems with Different Speaker Features», *Electronic Journal of Technical Acoustics*,2005. http://webcenter.ru/~eeaa/ejta/eng/ejtaeng.shtml