Electronic Journal «Technical Acoustics»
http://webcenter.ru/~eeaa/ejta/

2005, 14

Akira Kurematsu[1], Mariko Nakano-Miyatake[2], Hector Perez-Meana[2*], Eric Simancas Acevedo[3]

[1] *University of Electro-Communications, Chofu-shi, Tokyo 182-8585, Japan*
[2] *National Polytechnic Institute of Mexico, Av. Santa An, Col. San Francisco Culhuacan, 04430 Mexico D.F.*
[3] *Polytechnic University of Pachuca, Mexico*

Performance analysis of Gaussian Mixture Model speaker recognition systems with different speaker features

Received 28.03.2005, published 17.05.2005

This paper analyzes the effect of the speaker feature vector characteristics, in the performance of speaker recognition systems (SRS) based on the Gaussian Mixture Model (GMM). To this end, the performance of the SRS is analized using speaker features derived from: a) linear predictive cepstral coefficients (LPCepstral) extracted from the whole speech frame, b) LPCepstral derived from the voiced parts of the speech frame, c) LPCepstral extracted from voiced segments of speech frame together with the pitch information, d) LPCepstral extracted from voiced segments of each frame normalized using a Cepstral Mean Normalization (CMN). Evaluation results, using phrases of 2.5–3 second of telephone speech utterances in Japanese language, show that a fairly good performance of GMM-based SRS is achieved with most speaker features vectors with both, close test as well as with open-test, although the features vector providing the best recognition performance closely depends on each particular speaker.

INTRODUCTION

The development of efficient Speaker Recognition System (SRS) has been a topic of active research during the last decade, because they have a large number of potential applications in many fields that require accurate user identification or user identity verification such as: Shopping by telephone, bank transactions, access control to restricted places and information, voice mail and law enforcement, etc. According to the task that the SRS is required to perform, it can be divided in speaker identification system (SIS) or speaker verification systems (SVS), where the SIS has the task to determine the most likely speaker among a given speakers set, while the SVS has the task of deciding if the speaker is that she/he claims to be. Usually a SIS has M inputs and N outputs, where M depends on the feature vector size and N on the size of the speaker set, while the SVS usually has M inputs, as the SRS, and two possible outputs (accept or reject) or in some situations three possible outputs (accept, reject or indefinite).

According to the speech data set used, the SRS can also be divided in text-dependent speaker recognition system (TD-SRS) and text-independent speaker recognition system (TI-SRS), where the TD-SRS is trained and tested using the same kind of sentence or words, such that in the TD-SRS the sentence or words used for testing are known in advance, while

* Coresponding author, e-mail: hmpm@prodigy.net.mx

the TI-SRS is trained with a given data set and tested with a set of words or sentences that may be different from those used during training. Thus in the TI-SRS the sentences used during the verification stage are, in general, unknown in advance. This fact makes the development of efficient TI-SRS more complicated, especially in open tests. Because the performance TD-SRS and TI-SRS strongly depend on the data base used for training and testing, for TD-SRS short sentences are preferred, while for TI-SRS long sentences are necessary to correctly train the speaker model.

Several SRS have been proposed which use several efficient methods to properly estimate the main speaker characteristics voice. Among them we have statistical methods such as the Vector Quantization (VQ) [1, 2, 3, 4] and Dynamic Time Warping (DTW) [4], which use templates of small dimension [4]. These methods have shown to be accurate enough in TD-SRS applications when short-term utterances can be used. However their performance is not appropriate enough in TI-SRS applications where long-term utterances must be used [1, 5, 6].

A recently developed paradigm for speaker recognition tasks is the artificial neural network (ANN) whose target is to build an artificial system that tries to emulate the human brain behavior, by training one individual model for representing all speaker features [7]. Furthermore, the ANN can be used for many applications in the speaker recognition field, although the ANN are mainly used in speaker identification where they have shown be very efficient for features classification, providing a good performance in the identification task [7, 8]. In addition, the ANN uses a relatively few amount of parameters to carry out the recognition task. However, the ANN has the disadvantage that if one more speaker is added to the recognition system, in most cases, the ANN need to estimate all its parameters again, and for this reason its use has been limited.

The third kind of speaker recognition systems is based on stochastic methods, which have replaced to the statistical models in text independent applications where it is necessary to use long-terms utterances. Among the stochastic methods we have the SRS based on the Hidden Markov Models (HMM) [1, 4, 5, 6], which assign one model to each speaker features. Then, using these features during the training and test stages, the SRS compares them to find the model with the minimum distance, performing in this way the speaker recognition task [1, 7]. Another widely used stochastic method for speaker recognition tasks is the Gaussian Mixture Model (GMM), which is similar to the HMM with the difference that the GMM omits the temporal information implicit in the HMM [9, 10]. This means that the GMM has only one state and then it does not need the transition time from one state to other, as required in the HMM [9]. So, the GMM uses a unique Gaussian distribution matrix to represent each speaker. In addition, in most cases it is not necessary to use all covariance matrix components, because all Gaussian components are acting together to model the overall probability density function. Then, the full covariance matrix is not necessary even if the features are not statistically independent. This is because to take all components of the covariance matrix is equivalent to take only the main diagonal of the covariance matrix from each speaker model [9, 10]. Finally, because the GMM estimates one model for each speaker, if one speaker is added to the recognition system, it is necessary only to add the new speaker model, keeping unchanged the already trained ones. Thus, recently the GMM has been widely used in text-independent speaker recognition systems because, besides its desirable features described above, it has the capacity of representing broad acoustic classes with its individual Gaussian components, providing thus a very good performance in many speaker recognition applications [3, 4, 10, 11, 12].

The performance of all speaker recognition systems described above strongly depends on the speaker feature vectors, which are estimated from the speech signal by using time domain characteristics such as: linear prediction coefficients (LPC) and partial correlation coefficients

(PARCOR) or frequency domain characteristics such as: Mel-Spectrum, Mel-Cepstral and LPCepstral, etc. Among them, the feature vectors extracted from the LPCepstral provide a very good feature extraction performance, especially in GMM based TI-SRS [1, 5, 6, 8].

The GMM-based SRS performs fairly well in both text independent and text dependent applications when long utterances are available for training and recognition, as well as when the features vectors can be extracted from speech signals with relatively low distortion. Due to that the performance of most SRS proposed until now provide fairly good recognition rates when operate in closed tests, i.e. using the same data during both training and testing periods. However their performance degrades when the data used for testing are different from those used for training, i. e. open test. The main reason of this problem are the variations of the speaker feature vector, due to the acoustic conditions under which it is estimated such as noisy telephone lines, non stationary environment, the use of different microphones, etc. As an example we have the GMM-based SRS, reported in [3], which provides an identification accuracy of 96.8% using clean speech utterances, however its accuracy decrease to 80.8% when telephone speech utterances are used. In both cases the data base consists of utterances of 49 different speakers. To improve the SRS performance two alternatives will be analyzed: a) The use of speaker features vector with more than one speaker characteristics [2, 9, 11], or b) To reduce the speaker feature variations using channel normalization techniques such as Cepstral Mean Normalization (CMN) and RASTA filtering which provide a fairly good performance when it is necessary to compensate the communication channel variations and to reduce the distortion effects produced by a noisy environment [2, 13, 14]. However although all speaker feature vectors proposed until now performs fairly well in some specific situations, they may fail in other conditions or with some specific speakers, leading to a SRS performance degradation.

This paper presents an analysis of the GMM based TI-SRS with speaker feature vectors derived from: a) the LPCepstral coefficients extracted from the whole speech frame, b) the LPCeptral coefficients extracted only from the voiced parts of the speech signal, c) the features vector extracted from a combination of LPCepstral and pitch information, d) the features vector using the LPCepstral and enhancement feature techniques such as the Cepstral Mean Normalization (CMN) method, e) the features vector extracted from a combination of LPCepstral with CMN techniques and pitch information. Evaluation results were obtained using closed test as well as in open test with a data base of 12,000 telephone speech utterances, each one of 2.5 to 3 seconds, are given, and the advantages and disadvantages of above mentioned speaker features when used in GMM-based SRS are described in detail.

1. GMM-BASED TEXT INDEPENDENT SPEAKER RECOGNITION SYSTEM

Figure 1 shows a general structure of GMM-based text independent speaker recognition system (GMM-TI-SRS), shown in Fig. 1, which consists of two modules, the training module and the identification module, each one with three stage. During training, the speech signal is firstly acquired by the user interface. Subsequently the signal is feed to a preprocessing stage to speech enhancement, voice activity detection, etc. Next, the feature vector is estimated and finally the GMM is updated. During the identification or recognition operation the speech is feed into the preprocessing and features vector stages, which are the same used during training. Next the feature vector is feed into the GMM stage and its output used to perform the recognition task.

1.1. Preprocessing and segmentation stage

In this stage the speech signal is processed to reduce the distortion and the additive noise introduced by the communication channels, microphones, etc. by using standard speech enhancement algorithms [15]. It also limits the duration of the speech signal to be analized by detecting the initial and final points of the speech frame. This fact avoids the analysis of long silence intervals often present in the speech frames under analysis that can lead to SRS performance degradation. To this end several methods have been proposed, among them, the method proposed by Rabiner and Gold [15], provides a fairly good performance with a low computational complexity. In this method the speech power average over the time is estimated and analyzed point by point from left to right, as shown in Fig. 2, until both thresholds Th1 and Th2 are crossed. Then, when Th2 is crossed, the signal is analyzed from right to left until the first threshold Th1 is found and then, in that instant the initial point of the speech frame is determined. The final point is estimated in a similar form but in this case the analysis is carried out, firstly from right to left until Th2 is found and then from left to right to find Th1. Usually the threshold levels are set equal to 5% and 10% respectively of the maximum of speech signal power in the interval under analysis [16].

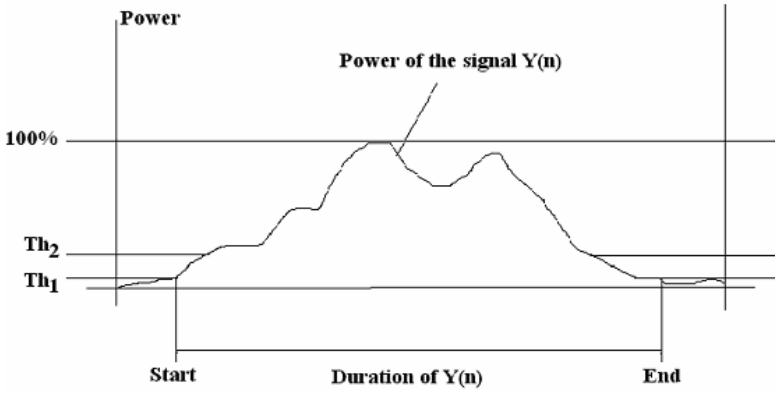

Figure 2. Initial and final point detection algorithm using power analysis

1.2. Feature vector extraction

A good performance of any pattern recognition system strongly depends on the extraction of a suitable feature vector that allow unambiguous representation of the pattern under analysis, with a number of parameters as small as possible. A simple way to estimate the speaker characteristics is the use of the linear prediction coefficients (LPC) of speech signal. The main reason about it is the fact that the structure of the vocal tract can be satisfactorily represented by using these parameters. However, it has been reported that better performance can be obtained if the LPC are combined with some frequency domain representation. One of these representations are the LPC features combined with the cepstral analysis, which allows to get a robust speaker characterization with low sensitivity to the distortion introduced in the signal transmitted through conventional communication channels [8].

The features vectors extracted from the whole speech signal provide a fairly good performance. However, when the LPCepstral coefficients are obtained from the LPC analysis, useful information of the speaker is still ignored or not taken in account, such as the pitch that is a specific feature of the individual speaker identity widely used to represent the glottal flow information.

The performance of SRS can be seriously degraded when the SRS uses speech signal transmitted though some communication channel, such as a telephone one, due to the

frequency response of the communication channel as well as the environment or the microphone characteristics. The LPCepstral coefficients have shown to be robust for reducing the problem of low speech quality. In most SRS even if they have a good performance using the same data training (closed test), their performance considerably degrades when the systems are used with different data set (open test), because the data for closed and open test for each speaker may have different acoustic conditions. Thus, channel normalization techniques may have to be used to reduce the speaker features distortion, keeping in such way a good recognition performance. Among the channel normalization technique we have the Cepstral Mean Normalization (CMN) and RASTA filtering, which can provide a considerable environmental robustness at a negligible computational cost [14]. Among them, the CMN provides better performance than the RASTA filtering because the latter introduces phase distortion and the recognition results obtained with a correct RASTA filtering are identical to those obtained using the CMN [13]. On the other hand, the use of more than one speaker feature is proposed as well as combination of them to get a more robust feature vector. To improve the SRS performance the LPCepstral coefficients obtained using the CMN can be combined with the pitch information because the pitch is a very important speaker feature. Thus to analyze the SRS performance with different speaker characteristics, the following speaker features will be estimated: (a) LPCepstral derive form the whole speech frame, (b) the LPCepstral coefficients extracted only from the voiced parts of the speech signal, (c) speaker features derived from a combination of LPCepstral and pitch information (d) enhanced feature vectors derived by using normalization techniques such as the Cepstral Mean Normalization (CMN), (e) speaker feature vectors derived from a combination of LPCepstral with CMN and pitch information.

1.2.1. Features vector derived form LPCepstral

To estimate the LPCepstral coefficients, firstly the speech signal is divided in segments of 20 ms length with 50% overlap using a Hamming window. Next, the LPC coefficients are estimated using the Levinson algorithm such that the mean square value of prediction error given by

$$E\left[e(n)\right] = E\left[S(n) - \sum_{i=1}^{P} a_i S(n-i) \right] \tag{1}$$

becomes a minimum, where $E[.]$ is the expectation operator, P is the predictor order and a_i is the i-th linear prediction coefficient (LPC). Next, once the LPC vector has been estimated, the LPCepstral coefficients can be obtained in a recursive way as follows [1]:

$$c_n = -a_n + \frac{1}{n}\sum_{i=1}^{n}(n-i)a_i c_{n-i}, \; n > 0, \tag{2}$$

where c_n is the n-th LPCepstral coefficient. Thus the SRS feature vector becomes

$$\mathbf{X}_t = \left[c_{1,t}, c_{2,t}, c_{3,t}, ..., c_{d,t}\right], \tag{3}$$

where t denotes the frame number.

1.2.2. Features vector derived from LPCepstral of voiced segments

The pitch and voiced part detection plays a very important roll in the speaker recognition systems because the pitch value and the voiced segments of speech signals contain the most important information about the speaker identity. Then the feature vector could be extracted only from the voiced segments of speech signal [16]. To this end, firstly the pitch period is detected using the autocorrelation method [5] as follows: Initially, the speech signal is segmented in frames of 20 ms with 10 ms overlap using a Hamming window; next the center clipper method [7] is applied to the windowed frame to reduce the effect of the additive noise intrinsic within of the speech signal. Subsequently, the autocorrelation of the center clipped segment is obtained. Finally the pitch value is estimate as the distance between two consecutive positions in which the normalized autocorrelation sequence is larger than a given threshold, as proposed in [15]. Using the pitch information, the speech segment is then classified as voiced or unvoiced, because the pitch only appears in the voiced segments. Thus, if the pitch does not exist the speech segment is considered as a unvoiced segment [5, 6, 8]; and the pitch exists then speech segment is classified as a voiced speech segment. The Fig. 3 shows clearly the result of this procedure that proves that the detection of the voiced part is correctly done.

If only the voiced segments of speech signal are taken in account for features extraction, the original speech signal is transformed into a new speech signal containing only voiced parts, neglecting in such way the unvoiced and noisy silence parts, as shown in Fig. 4. This may improve the feature extraction because the unvoiced and silence parts provide non-useful information [10].

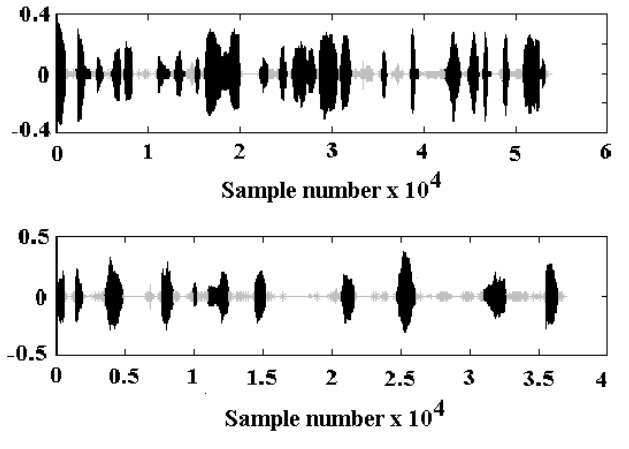

Figure 3.

Pitch period detection
in two different speech signals

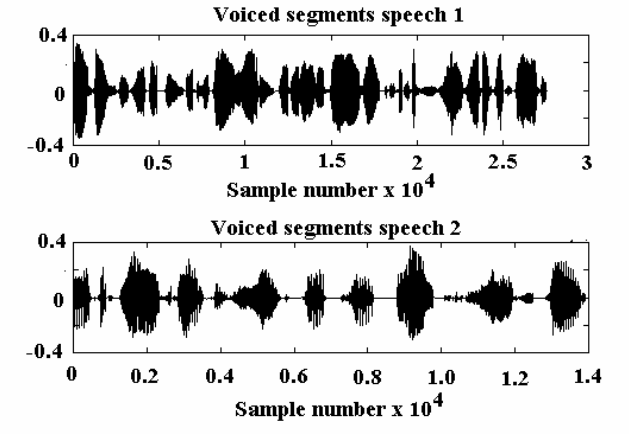

Figure 4.

Speech signal reduction
taking only voiced part

The new signal, with only the voiced parts, has less samples number than the original one, but contains the essential parts required to estimate the principal features that identify the speaker and, as shown in the Fig. 4, the number of data is reduced, in many cases, as far as 50%. Once the new signal is constructed only with the voiced parts, it is divided in segments of 20 ms length with 50% overlap using a Hamming window. Next, the LPC coefficients are estimated using the Levinson algorithm as mentioned above, and then the LPCepstral coefficients are estimated using the eq. (2). Here two different features vectors can be estimated. The first one consists only the LPCepstral of voiced segments

$$\mathbf{X}_t^v = \left[c_{1,t}^v, c_{2,t}^v, c_{3,t}^v, ..., c_{d,t}^v\right], \tag{4}$$

and the second one consists of LPCepstral coefficients and pitch information

$$\mathbf{X}_t^p = \left[c_{1,t}^v, c_{2,t}^v, c_{3,t}^v, ..., c_{d,t}^v, \log_{10} F_{0,t}\right], \tag{5}$$

where $c_{n,t}^v$ is the *n*-th LPCepstral coefficient and $F_{0,t}$ is the inverse of the pitch period at the block *t*. Here $\log_{10} F_{0,t}$ is used instead of the pitch period, because the probability distribution of the $\log_{10} F_{0,t}$ is close to the normal distribution.

1.2.3. Reinforcing and enhancing feature vectors

In long distance speaker recognition, the speech signal is transmitted through a channel communication and then is processed by the SRS. However, the speech signal suffers some distortion or variation due to the communication channel effects, noise environment, etc. Because these distortions are added to the principal components of the speech signal, it is necessary to remove the undesirable information before to proceed with the recognition process. To this end it would be convenient to enhance the estimated feature vector. Thus we can subtract the global average vector from all feature vector components. In this process, known as Cepstral Mean Normalization (CMN) [2, 3, 7, 17], it is assumed that the mean values of LPCepstral coefficients of clean speech is zero, so that if the mean value is subtracted from the feature vector components its mean value becomes zero. This avoids the distortion introduced by the additive noise when the signal passes through the communication channel. This technique is equivalent to a high-pass filtering of LPCepstral coefficients, because the CMN estimates the mean value of the LPCepstral vector coefficients and subtracts it from each component, as shown in eq. (6):

$$CMN_{n,t} = c_{n,t} - \frac{1}{T}\sum_{t=1}^{T} c_{n,t},\ 1 \le \mathrm{n} \le d\ , \tag{6}$$

where $c_{n,t}$ is *n*-th LPCepstral coefficient at block *t* and *T* is the total number of frames in which the speech signal was divided to extract the feature vectors. Fig. 5a and Fig. 5b show the effect produced in the feature vector when the Cepstral Mean Normalization (CMM) technique is applied to one of the LPCepstral coefficients extracted from the only the voiced part of the speech signal. In this situation the features vector becomes

$$\mathbf{CMN}_t = \left[CMN_{1,t}, CMN_{2,t}, CMN_{3,t}, ..., CMN_{d,t}\right]. \tag{7}$$

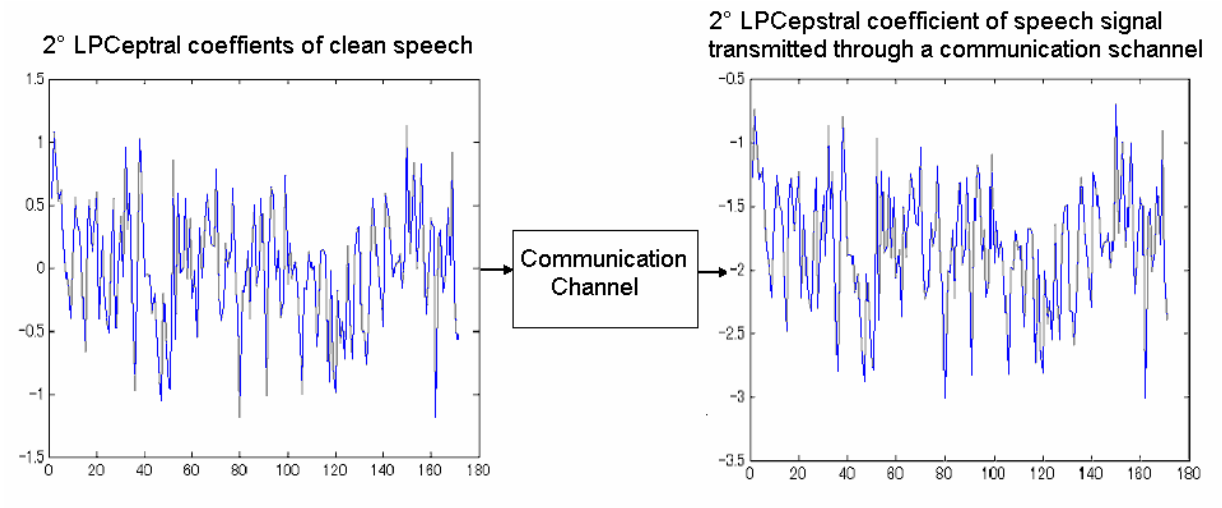

Figure 5. The second LPC-Cepstral coefficient extracted from the speech signal

1.2.4. Enhanced LPCepstral and pitch information

One way to improve the SRS performance is using more that one speaker characteristic in the feature vector. Thus, because the pitch carries specific information about the speaker identity, a combination at the feature level of the enhanced LPCepstral coefficients, CNM, with the pitch, which is widely used as a representation of the glottal flow information, can be used. Because the pitch does not appear in unvoiced parts of the speech signal, the combination of both features is done only if the signal frame is considered as a voiced part, in such case the feature vector is given as

$$\mathbf{CMN}_t^p = \left[CMN_{1,t}^v, CMN_{2,t}^v, CMN_{3,t}^v, ..., CMN_{d,t}^v, \log_{10} F_{0,t}\right], \tag{8}$$

where

$$CMN_{n,t}^v = c_{n,t}^v - \frac{1}{T}\sum_{t=1}^{T} c_{n,t}^v,\ 1 \le \mathrm{n} \le d\ , \tag{9}$$

where $c_{n,t}^v$ is the *n*-th LPCepstral coefficient of block *t*. This feature vector consists of 17 parameters, where 16 parameters are the LPCepstral coefficients represented by c_{dt} with $d = 16$ and 1 parameter of the pitch, represented by the $\log_{10} F_{0,t}$ because, as mentioned before, to the distribution of the $\log_{10} F_{0,t}$ is closer to the normal distribution.

1.3. Gaussian mixture model stage

In the classifier stage the Gaussian Mixture Model (GMM), shown in Fig. 6, is used to provide a speaker voice sound model. The GMM is similar to the widely used Hidden Markov Model (HMM) with the difference that the GMM only has a state with a Mixture Gaussian distribution that represents different acoustic classes and ignores the temporal information of the acoustic observation sequence. Here, only one model is built for each speaker that represent all his/her features using only the main diagonal of the covariance matrix, the mean vector and the mixture weights [3, 18]. This representation results in a few amounts of parameters which make it possible that the GMM model can be used in Text-Independent (TI) speaker recognition applications.

In the GMM model, the features distributions for each speaker are modeled by using the sum of the weighted speech signals Gaussian distributions densities of the speaker under analysis given as

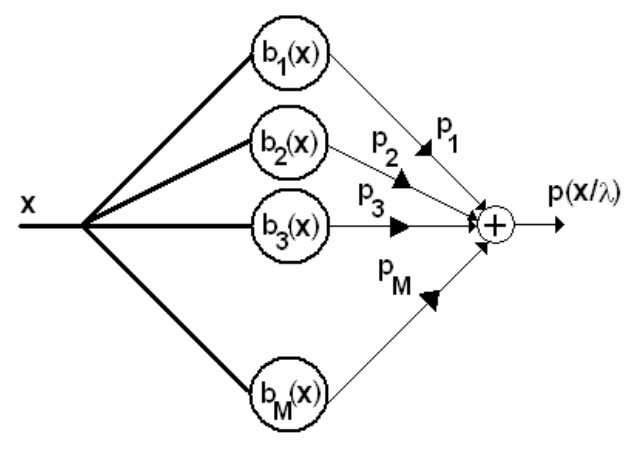

Figure 6.
Gaussian Mixture Model

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}), \tag{10}$$

where

$$\sum_{i=1}^{M} p_i = 1, \tag{11}$$

x is the D-dimensional features vector estimated in the feature stage described in section 2.2, $p(\mathbf{x}|\lambda)$ is the speaker model, p_i is the *i*-th mixture weight and $b_i(\mathbf{x})$ is a D-Variate-Gaussian distribution given by

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}, \tag{12}$$

which is characterized by its mean $\boldsymbol{\mu}_i$ and covariance matrix $\boldsymbol{\sigma}_i$, where T denotes the transpose operation. The mean vector, $\boldsymbol{\mu}_i$, covariance matrix, $\boldsymbol{\sigma}_i$, and mixture weights, p_i, of all the density components determine the complete Gaussian Mixture Density $\lambda(\boldsymbol{\mu}, \boldsymbol{\sigma}, p)$ used to represent the speaker model.

To obtain the optimum Gaussian Mixture Density, $\lambda(\boldsymbol{\mu}, \boldsymbol{\sigma}, p)$, for each speaker, it is necessary to estimate the optimum values of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and p. It can be done in an efficient manner using the Maximum-Likelihood (ML) estimation method in which, a set of D-dimensional feature vectors, derived as shown in section 2.2, form the vector $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,...\mathbf{x}_T)$ used for training the Gaussian Mixture Model (GMM) such that the Likelihood of the GMM for the estimated vector **X,** given by

$$p(\mathbf{X}|\lambda) = \prod_{t=1}^{T} p(\mathbf{x}_t|\lambda), \tag{13}$$

be maximized. However, Eq. (13) is a no-linear function of the speaker model parameters, λ, so that it can not be maximized directly.

To estimate the parameters of ML it must be used an iterative algorithm such as 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 $\bar{\lambda}$ is estimated such that $p(\mathbf{X}/\bar{\lambda}) \geq p(\mathbf{X}/\lambda)$. Next, we make the model parameters λ equal to those estimated in the actual stage, $\bar{\lambda}$, so that, the new model become 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. This optimization procedure can be summarized as follows.

1. Estimate the initial condition of $p(\mathbf{x}/\lambda)$ using the Viterbi algorithm [12, 15].
2. Compute a new L, D-dimensional features vectors using the methods described in section 1.2, to form the features vector $\mathbf{X}=(\mathbf{x}_1, \mathbf{x}_2,,..\mathbf{x}_N)$ to be used for training.
3. Compute the new mixtures weighting factors as:

$$\bar{p}_i = \frac{1}{T}\sum_{t=1}^{T} p(i|\bar{\mathbf{x}}_t, \lambda), \tag{14}$$

d) Compute the mean as

$$\bar{\boldsymbol{\mu}}_i = \frac{\sum_{t=1}^{T} p(i|\bar{\mathbf{x}}_t, \lambda)\bar{\mathbf{x}}_t}{\sum_{t=1}^{T} p(i|\bar{\mathbf{x}}_t, \lambda)}, \tag{15}$$

e) Compute the variance

$$\bar{\sigma}_i^2 = \frac{\sum_{t=1}^{T} p(i|\bar{\mathbf{x}}_t, \lambda)|\bar{\mathbf{x}}_t|^2}{\sum_{t=1}^{T} p(i|\bar{\mathbf{x}}_t, \lambda)} - |\bar{\boldsymbol{\mu}}_i|^2, \tag{16}$$

f) Compute the new conditional probabilities matrix given by a (M×N)-dimensional matrix **P** whose (*i*,*t*)-th component, $p(i|\bar{\mathbf{x}}_t, \lambda)$ is given by

$$p(i|\bar{\mathbf{x}}_t, \lambda) = \frac{p_i b_i(\bar{\mathbf{x}}_t)}{\sum_{k=1}^{M} p_k b_k(\bar{\mathbf{x}}_t)}. \tag{17}$$

The training proceed until the convergence is achieved using as initial conditions the conditional probabilities given by (10). The variables that need to be considered are the kind of Covariance matrix, the order of the Mixes and the model parameters previous to the maximization of the likelihood of GMM which can be different depending on the application.

1.4. Decision algorithm

During the recognition task, after the GMM parameters for each speaker have been estimated, the target is to find the model with the maximum likelihood a posteriori for a given observation sequence. Usually [6, 19]

$$\hat{S} = \arg\max_{1\leq k\leq S} P(\lambda_k|\mathbf{X}) = \arg\max_{1\leq k\leq S} \frac{p(\mathbf{X}|\lambda_k)\,p(\lambda_k)}{p(\mathbf{X})}, \tag{18}$$

where $\hat{S}$ is given by the Bayes rule as shown in eq. (18). Then assuming that all speaker are equally probable and noting that $p(\mathbf{X})$ is the same for all speakers models, the classifiers rule is reduced to

$$\hat{S} = \arg \max_{1 \le k \le S} \; p(\mathbf{X} | \lambda_k) \tag{19}$$

and then, from the above described algorithm, the estimated speaker identity becomes [3]

$$\hat{S} = \arg \max_{1 \le k \le S} \; \sum_{t=1}^{T} \log p(\bar{\mathbf{x}}_t \mid \lambda_k), \tag{20}$$

where $p(\bar{\mathbf{x}}_t | \lambda_k)$ represents the Gaussian Mixture density given by eq. (10).

2. EVALUATION RESULTS

The GMM-based SRS was trained and evaluated using a data-base of 80 speakers, provided by the KDD Corporation of Japan, which contains 10 different lists of 50 different contexts of 3–5 s duration and the digit numbers, both in Japanese language. The speech signal was recorded, during time intervals of one month, in real condition using the public telephone network with a sampling frequency of 8 KHz. For training and evaluation a total of 10,850 utterances are used. 7,147 of them are used for training and closed test evaluation, and 3656 for open test, where the last one were different from that used for closed test evaluation to be able to carried out a text independent evaluation. Firstly, the SRS was evaluated using speaker features vector derived from the whole speech frame corresponding to digits and phrases with a mixes order equal to 8 obtaining a global recognition rate equal to 88.01%.

The recognition results for each speaker, individually, are shown in Fig. 7. Next, the system was evaluated using separately phrases and digits. When the database is used separately the recognition rate improves. In his situation the recognition rate was 96.61% when only phrases are used and 90.14% when only digits are used.

Figure 7 shows the recognition performance obtained for each one of the 80 speaker individually. Here it can see that, even if there are 4 speakers with a recognition rate lower than 70%, most of them present a recognition rate higher than 80%. Finally, the system was trained and evaluated using all 80 speakers using phrases and digits with a mixes order of 16. In this situation the recognition rate was 92.87%. No further improvement was obtained when the number of mixes was increases. The experimental results were obtained for closed and open test to compare the performance of each development system. The system was evaluated for text independent speaker recognition by using data stored from the telephone. These results are closed to those reported by [3].

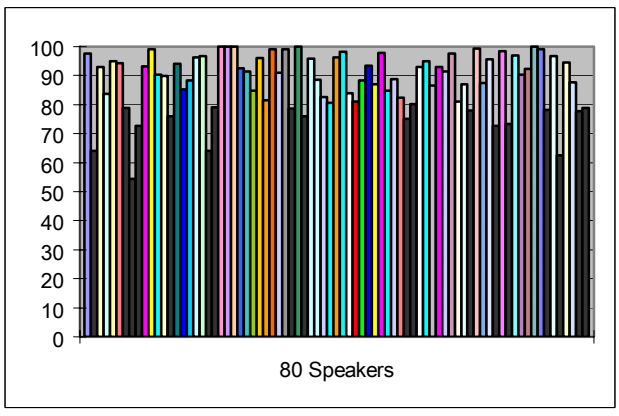

Figure 7. Percentage recognition for each speaker using phrases and digits

The SRS was evaluated using a feature vector with 16 LPCepstral coefficients extracted from only voiced part of the speech signals. The first evaluation of the baseline system is using these 16 LPCepstral coefficients extracted from only voiced part. The second evaluation is using the CMN technique to enhance the feature vector quality which has been affected by the channel communication and the environment noise. The third evaluation is using a combination of LPCepstral and pitch information and the forth evaluation is using the combination of LPCepstral coefficients and pitch information applying the CMN technique. All system evaluations, which are discussed in each respective section, are presented in the Table 1 and Table 2.

Table 1. Experimental results comparison of using whole and voiced part speech signal

Feature Vector	**LPCepstral From whole speech signal**	**LPCepstral From voiced part**
Evaluation **(Close test)** 7147 phrases	96.61%	97.13%
Evaluation **(Open test)** 3658 phrases	82.34%	83.57%

Table 2. Experimental results with different features vectors

Feature vector	LPCepstral from voiced part	LPCepstral from voiced parts using CMN	LPCepstral from voiced and pitch information	LPCepstral from voiced parts using pitch and CMN
Closed test 6581 phrases	93.31%	80.72%	99.18%	97.29%
Open test 3282 phrases	76.97%	70.88%	80.29%	77.57%

2.1. SRS using LPCepstral extracted from voiced part of the speech signal

For this evaluation of two systems were developed. The first system was trained with feature vector of 16 LPCepstral extracted from the voiced part, whose feature vector is given by eq. (4) and in the second one with a feature vector given by eq. (3). The results depicted in the Table 1, show the performance increment of 0.52% in close test and 1.23% in open test (97.13% and 83.57% respectively).

The first system was trained with feature vector extracted from voiced parts of only those phrases where pitch information was found. For system training, 6581 phrases were used in close testing and 3282 phrases in open test. Using features vector from only voiced part there are some advantages against the use of whole speech signal; it saves storage requirements due to the reduction of the length of feature vector parameters as well as 50%, saving training time as we used 20 iterations in the preliminary experimentation and with using voiced part was not necessary more than 10 iteration and it gives better results. In contrast, we are reducing the length of enrollment data providing less speaker information to GMM, and because of we discarded some phrases when voiced part detection stage did not found pitch information, the performance system does not have an improvement.

During evaluation, each SRS based on GMM is trained and tested with a data base of 9863 phrases in which the pitch detection algorithm found enough pitch information. For training and close test evaluation 6581 phrases were used and for open test 3282 different phrases were used.

2.2. SRS using LPCepstral enhanced by CMN

The performance of the Speaker Recognition System using as feature vector 16 LPCepstral coefficients extracted from the voiced part compared with 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 Table 1. We analyzed the results for those speakers with bad performance, especially in open test, by using the LPCepstral without CMN and after using CMN and we observe that those performances were improved considerately as show in the Fig. 8. On the other hand, for those speakers who had good performance in open test using the LPCepstral without CMN, after using the CMN, they performance decreased as show in the same Fig. 8. Some researches have presented similar decreasing performance when CMN is used [2, 9, 10]. This is because the CMN improves the performance in noise conditions, however it decreases when clean speech signal is used and when the channel communication does not have variety. This is because CMN assumes that the Cepstral mean of clean speech has zero mean, which is not entirely correct. In addition, CMN eliminates convolution effects but it does not eliminate additive noise and does not take into account nonlinear and non-stationary channel conditions [2].

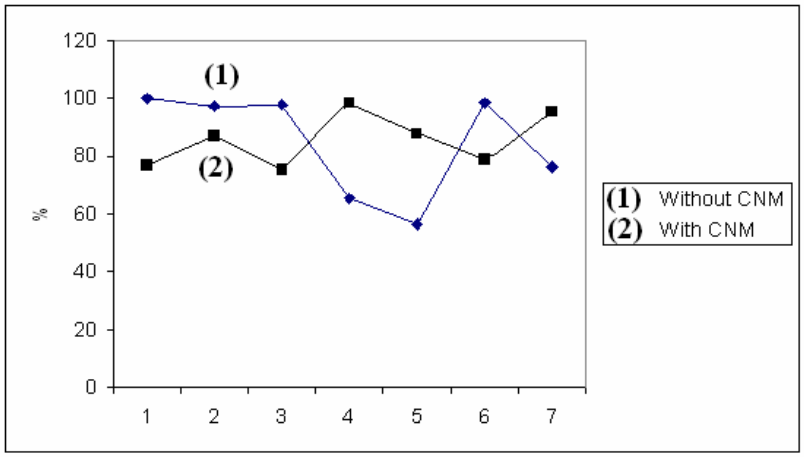

Figure 8. Effect of the use of CNM in the SRS recognition rate of the 7 speakers with the largest and lowest recognition rate without using the CNM

2.3. SRS using LPCepstral and pitch

The combination of the LPCepstral with the pitch shows a very good performance of the speaker recognition system based on GMM improving 5.87% in close test and a 3.32% in open test as we can see in Table 2. This is because the LPCepstral is a robust feature vector and combined with the pitch which is a specific feature of the individual speaker identity. However, the performance in open test still is far from close test performance. The pitch reinforces very well to the vector feature given more specific speaker information, however, due to the high intra-speaker variability of the pitch, the system still have mismatch for data not training.

If only the voiced segments of speech signal are taken in account for features extraction, the original speech signal is transformed into a new speech signal containing only voiced parts, neglecting in such way the unvoiced and noisy silence parts, as shown in Fig. 4. This may improve the feature extraction because the unvoiced and silence parts provide non-useful information [10].

The results shown in the Table 2 for the use of LPCepstral applying CMN and combined with the pitch have an improvement. However, this improvement is less than the achieved for using LPCepstral without CMN and pitch combination, thus, using CMN technique the system have the same problem of decreasing performance. In section 1.2.3, we explained why in certain condition the CMN not performed well in the system.

2.4. SRS proposed by other authors

Several algorithms have been proposed for speaker recognition, among them we have pattern matching based SRS and dynamic time warping (DTW) based SRS with feature vector extracted from which LPCepstrum which provides a recognition rate of 79% and 90% respectively [5]. SRS based on the Gaussian mixture models (GMM) have also been proposed with recognition rates between 89% and 92% [5, 19, 20], using feature vectors extracted from the Mel-Spectrum of the whole speech signal. Several approaches have been proposed using only the segments of speech signals in which voice activity is detected with the feature vectors extracted using the Mel-Spectrum and differential cepstral of speech signal; providing, for text dependent, a recognition rate 99% for clean data and a 95% using telephone [20]. While for text independent speaker recognition it is reported a recognition rate between 85% and 93% using conversational speech and a recognition rate between 65% and 80% using speech data obtained from a radio communication system [20]. In these two cases the training was carried out using speech segments with duration of 2 minutes for training and 30 seconds form testing.

3. CONCLUSIONS

The recognition performance of a GMM based SRS with different features vector was analyzed. Evaluation results show that the use of voiced parts to extract the feature vector gives better performance in comparison with the performance obtained using feature vector extracted from the whole speech signal where there is possible that feature vector extracted from unvoiced part degrades the GMM recognition performance. This fact agrees with the results reported in [20, 21, 22]. Evaluation results also show that the use of LPCepstral coefficients as principal feature vector enhances the peak frequency of the speech spectrum involved in the analysis, providing better performance in text-independent speaker recognition systems based on GMM as shown in the Table 1. However, the variations of the communications channel and the noise environment do more difficult an accurate estimation of the LPCepstral coefficients, doing it necessary the use of channel normalization techniques.

It is well known that CMN is a good channel normalization technique when the speech signal is degraded by the communication channel. However, as shown in section 2, in several situations the CMN is not a desirable solution to enhance the quality of the feature vectors due to the decreasing of the system performance when this channel normalization technique is used.

The pitch is a very important speaker feature to perform the speakers recognition, providing a considerable improvement of the SRS performance in several cases, as we can see from the evaluation results in which the system performance is considerably improved. Although this is only for the closed test data because for open training a larger amount of data is required to adapt the model with most of the intra-speaker variations [23, 24].

Considering the results and the analysis obtained using each feature extraction technique applied to the GMM based speaker recognition system, we can conclude that it is possible to improve the recognition rate if each speaker signal speech is analyzed separately, due to the different acoustic conditions that are present in the speech signal used to characterize each different speaker. Thus the use of any particular technique that should be applied depends on the acoustics conditions that the speech signal of each different speaker presents. We arrive to this conclusion because the evaluation results show that, the performance of the GMM based SRS using each feature extraction technique improves for some speakers but worse for others. Thus the idea of using the different feature extraction techniques separately for each speaker depending of theirs acoustics conditions presented in this speech signal can be useful in a speaker verification system where it is possible know in advance which model we should use for feature vector extraction of a given speaker. This is possible because the speaker, whose identity must be verified, should provide the information necessary to decided if his/her is or not the person who claim to be.

ACKNOWLEDGEMENTS

We thanks to the Japanese Government and to The University of Electro-Communication of Tokyo for the support provided through the JUSST program during the realization of this research. We are also grateful to the National Science and Technology Council of Mexico, CONACyT, for the support provided during the realization of this research.

REFERENCES

[1] E. Simancas-Acevedo, A. Kurematsu, M. Nakano-Miyatake, H. Perez-Meana. Speaker Recognition Using Gaussian Mixtures Model. Lecture Notes in Computer Science, Bio-Inspired Applications of Connectionism, Springer Verlag, Berlin, 2001, 287–294.

[2] H. A. Murthy, F. Beaufays, L. P. Heck, M. Weintraub. Robust Text-Independent Speaker Identification over Telephone Channels. IEEE Transactions on Speech and Audio Processing, vol. 7, N°5, September 1999.

[3] D. A. Reynolds. Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on Speech and Audio Processing, vol. 3, N°1, 72–83, January 1995.

[4] S. Van Vuren. Comparison of Text-Independent Speaker Recognition Methods on Telephone Speech with Acoustic Mismatch. Oregon Graduate Institute of Science & Technology Center for Spoken Language Understanding, 20000 N.W. Walker Road, Beaverton, Oregon 97006 USA.

[5] J. P. Campbell. Speaker Recognition: A Tutorial. Proceedings of the IEEE, vol. 85, N°9, 1437–1462, Sept. 1997.

[6] H. K. Kim, H. S. Lee. Use of Spectral Autocorrelation in Spectral Envelope Linear Prediction for Speech Recognition. IEEE Transactions on Speech and Audio Processing, vol. 7, N°5, September 1999.

[7] T. Ganchev, A. Tsopanoglou, N. Fakotakis, G. Kokkinakis. Probabilistic Neural Networks Combined with GMMs For Speaker Recognition over Telephone Channels. 14-th International Conference On Digital Signal Processing (DSP 2002), 2002, July 1-3, Santorini, Greece, Volume II, 1081–1084.

[8] D. A. Reynolds. Experimental Evaluation of Features for Robust Speaker Identification. IEEE Transactions on Speech and Audio Processing, vol. 2, N°4, October 1994.

[9] K. P. Markov, S. Nakagawa. Integrating Pitch and LPC-Residual Information with LPC-Cepstral for Text-independent Speaker Recognition. J. Acoustic Society of Japan (E), 20, 4, 281–291, 1999.

[10] J. Pool, J. A. du Preez. HF Speaker Recognition. Thesis notes, Digital Signal Processing Group, Department of Electrical and Electronic Engineering, University of Stellenbosch, March 1999.
[11] M. D. Plumper, T. F. Quatieri, D. A. Reynolds. Modeling of the Glottal Flow Derivative Waveform with Application to Speaker Identification. IEEE Transactions on Speech and Audio Processing, vol. 7, N°5, September 1999.
[12] K. Markov, S. Nakagawa. Frame Level Likehood Normalization For Text-Independent Speaker Identification Using Gaussian Mixture Models. The Fourth International Conference on Spoken Language Processing, ICSLP96, vol. 3, October 3–6, Wyndham Franklin Plaza Hotel, Philadelphia, PA, USA.
[13] J. de Vetch, L. Boves. Comparison of Channel Normalization Techniques For Automatic Speech Recognition Over the Telephone. Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmen, The Netherlands.
[14] F. Liu, Richard M. Stern, Xuedong Huang, Alejandro Acero. Efficient Cepstral Normalization For Robust Speech recognition. Department of Electrical and Computer Engineering, School of Computer Science, Carnegie Mellon University. Pittsburgh, PA 15213.
[15] 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, vol. ASSP-24, N°5, 399–418, October 1976.
[16] B. Rabiner, B. Gold. Digital Processing of Speech Signals. Prentice Hall, Englewood Cliffd, NJ, 1975.
[17] D. Hardt and K. Fellbaum. Spectral Subtraction and Rasta Filtering in Text Dependent HMM-based Speaker Verification. Proc. of ICASSP, vol. 2, 867-870, April 1997.
[18] E. Simancas, M. Nakano Miyatake, H. Perez-Meana. Speaker Verification Using Pitch and Melspec Information. Journal of Telecommunications and Radio Engineering, vol. 56, 46–57, Jan. 2000.
[19] F. Hou, B. Wong. Text Independent Speaker Recognition Using Probabilistic SVM with GMM Adjustment. Proc. of the International Conference of Speech, Acoustics and Signal Processing, 305–308, 2003.
[20] D. A. Reynolds. An Overview of Automatic Speaker Recognition Technology. Proc. of the International Conference of Speech, Acoustics and Signal Processing, vol. 4, 4072–4075, 2002.
[21] E. Simancas Acevedo, H. Perez-Meana, M. Nakano Miyatake, A. Kurematsu. Effect of Voiced Segments in Gaussian Mixture Model Text Independent Speaker Verification. Journal of Electromagnetics Waves and Electronic Systems, vol. 8, N°7, 34–42, August, 2003.
[22] R. Zheng, S. Zhang, B. S. Xu. Text Independent Speaker Identification Using GMM-UBM and Frame Level Likelihood Normalization. International Symposium on Chinese Spoken Language Processing, 289–292, Dec. 2004.
[23] M. Kepesi, J. Macku. Introducing the Single-Channel Speech Separation Problem. Department of Telecommunications, Brno University of Technology, Purkynova 118, 612 00 Brno.
[24] M. Plsek, M. Vondra. Pitch Detection in Noisy Speech Recordings. Brno University of Technology, Faculty of Electrical Engineering and Communications, Department of Telecommunications, Purkynova 118, 61200 Brno, Czech Republic.