Speech/Speaker Recognition Using a HMM/GMM Hybrid Model

Elena Rodríguez, Belén Ruíz, Ángel García-Crespo, Fernando García

Universidad Carlos III de Madrid
c/ Butarque, 15, 28911 Leganés (Madrid), Spain
e-mail: pdgar@rioja.uc3m.es

Abstract. In this paper, a speaker recognition voice based system is presented [5]. We have implemented it in a Sun platform.We train (and test) the system using a Database recorded in several sessions in order to repair the huge effects that the speech variability with time has in the recognition rate system. Several experiments have been made in order to achieve the best configuration in the system set up. This is an important point to take into account in a real world system in which users train the system once and the models generated in the training process are not updated for strategic reasons. The recognition rate obtained for the proposed system is around 93% if the speech came from a microphone is around 90% when the speech came from a phone line.

1 Introduction

Nowadays security when accessing services and applications through the network communications it is an important topics and needs to be improved. For this reason, it is necessary to include some constrains in order to reject impostors and accept only authorised users when they try to access services or applications [1] [5]. The most reliable method is to use own user characteristics as the image, the speech, the fingerprints, etc. The user voice is the simplest way for user recognitionfor two main reasons: it is the most useful communications way between people and it can be transmited by phone line, multimedia networks, etc that are widely extended. From the speech recorded by microphone or phone and following in this approach, the spectral features (characteristics and variations) from a utterance spoken by the user are extracted and modelled in a statistical way. The system extracts the spectral characteristics from the speech recorded. Those characteristics are represented in a mel-scale generating a Hidden Markov Model. The HMM tries to isolate the speech from the environment noise and modells the spectral representation by a Fdp (function density of probability) composed by a mixture of Gaussian function. The spectral envelope is fitted to a number of Gaussian decided in the system set up. When a model for each user is calculated in the initial process (training process), it is stored in the Database and it is ready to be used when he/she tries access to the system. Then, the system generates the user spectral features, searching in the Database the model that better fits these features (identification system), or comparing with the pre-stored model for this user (verification system). When the distance of the user model is less than a threshold defined in the system set up, the user will be accepted to the service or application, otherwise is rejected.

The paper is organised as follows: A description of HMM and GMM algorithms are presented first [1] [2] [3] [6]. Next, the system description will be shown continuing with the system implementation and results obtained with the database used. The conclusions close our presentation.

2 System Description

In this paper, a speaker verification voice system based on a HMM/GMM hybrid model (Hidden Markov Model/Gaussian Mixture Models) is presented.

2.1 HMM Description

A HMM is defined by the initial probability π_i and the transition and observation probabilities (a_{ij} ,b_i (O_t)). A statistic model (HMM) with three states and continuos transition model is used in our system. The first and third states represent the beginning and end of the utterance and the second state represents the speech itself [1].

In each state the symbol distribution follows a statistic that can be modelled by a Gaussian function or a mixture of Gaussians functions. In each state a different configuration of mixture has been defined. In this sense the first and latter states could be modelled only by one Gaussian, being the number of Gaussian T in the intermediate state much higher. In each state the observation's probability b_i (O_t) is calculated through a Gaussian mixture model (GMM). Each Gaussian is defined by mean and variance. The mixtures number depends on the amount of data and its distribution.

We have implemented a text independent identification/verification system. The system works in two process. The first one is the training process and the second one is the recognition process.

Training process

To develop the system, the first task that must be accomplished is the trainning process. Generally, this process is performed once only in order to generate a speaker model. But, when the user's characteristics have drastically over a period of new training is needed for the user model.

In this process, we must analyse the speech signal, making a model for each speaker and generating the speaker Data Base that contains all possible users of the system. In the figure 1 we can see a diagram that shows the three main blocks of this process.

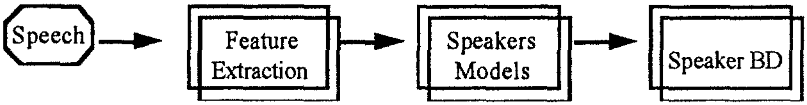

Fig. 1. Training process

The first block named feature extraction analyses the spectral speech characteristics. The input to this block is the sampled speech using PCM a 8 KHz being previously filtered with a low pass filter with a 3.8 KHz of cutoff frequency. A preemphasis of speech signal is made in order to flat the spectrum. Therefore, a windowing of the signal is made with a rectangle window of 25 ms overlapping the windows (speech frames) 12,5 ms. This is made in order to reduce the high changes in the spectral features between speech frames. Then we make a vectorial representation of speech signal, generating the feature vector of 22 components composed by 10 Mel-cepstrum, 10 differential Mel-cepstrum, the energy and differential energy.

The objective of the second block is the generation of individual speaker models. In this work we use a statistic models, a Hidden Markov Model (HMM) with 3 states and the observation probability defined by Gaussian mixture models characterised by a means and a variance.

When we have all speaker models we create the Speaker Data Base that contains all of users, and which will be used in the recognition process.

Recognition process

The objective of this process is the speaker verification, comparing if the speech characteristics obtained by the user are adjusted to the model generated in the training process for the user claimed. When the distance of the speech characteristic to the model are less than a predifined threshold, the user is accepted, otherwise is rejected.

When the system is a user identification,the recognition process calculates the distance of the speech characteristics of user to the possible candidates contained in the DB, obtaining the model more similar to the input model. If the two models correspond to the same speaker the acceptance is produced, otherwise the speaker is rejected. Figure 2 and figure 3 presents an overview of this process (verification /identification).

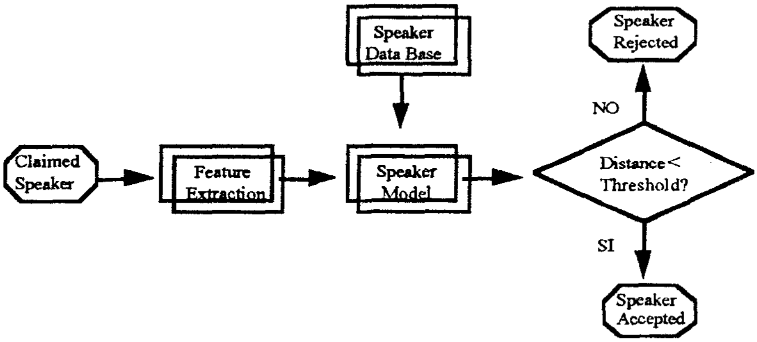

Fig. 2. Speaker verification system.

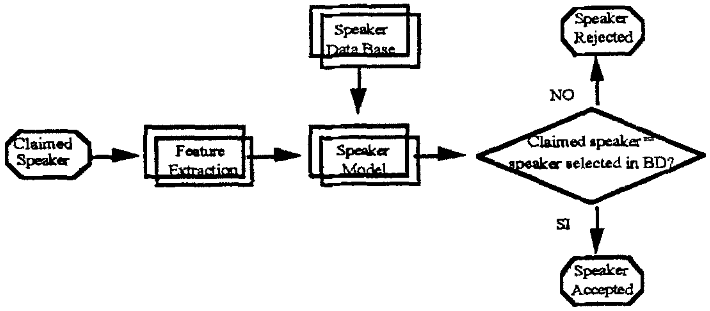

Fig. 3. Speaker identification system.

2.2 GMM Description

One of the most widely used method to model speaker speech characteristics, are Gaussian functions or mixture of Gaussians functions [2]. There are two principal motivations for using Gaussian mixture densities as a representation of speaker identity. First one is the possibility to represent the speaker dependent vocal tract configurations and the second one is the observation that a linear combination of Gaussian basis functions is capable, representing a large class of distributions.

GMM works in different ways depending on the choice of covariance matrices. The model can have one covariance matrix per Gaussian component. This covariance is named nodal covariance. The second one is a grand covariance that consist in a single covariance matrix for all Gaussian components in a speaker model and the third kind of covariance is a global covariance that is a single covariance matrix shared by all speaker models.

A Gaussian mixture density is a weighted sum of M component densities and given by the equation :

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

where i=1...M, $\vec{x}$ is a random vector, $b_i(\vec{x})$ are the component densities and p_i are the mixture weight.

Each component density is a Gaussian function of the form:

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

where $\vec{\mu}_i$ is the mean vector, $\sum_i$ is the covariance matrix and the mixture weights satisfy the constraint $\sum_{i=1}^{M} p_i = 1$.

The Gaussian mixture density is parameterised by the mean vectors, covariance matrices and mixture weights from all component densities. The most useful notation is:

$$\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\}\ i = 1, \ldots, M.$$

For speaker identification, each speaker is represented by a GMM and is referred to by his λ.

3 Data Base Definition

In order to evaluate the system we use a multi-session Data Base recorded in the University Carlos III from Spain in a supervised form. The DB contains 40 speakers recorded in 12 sessions separated in time from 3 days to 1 week each. The speech has been recorded with a Sound Blaster of 16 bits. The sample frequency is 8 KHz making a previous filtering of the speech to 3.8 KHz. The format of the speech files are .WAV containing a header o 44 bytes in which is described the length of data segment, number of record channels (mono or stereo), etc.;

We have used also a Database known as Polycode II. This Data Base has been recorded from IDIAP in an automatic way by phone, with previous filtering from 300 to 3400 Hz to eliminate the noise of the telephone line. The number of speakers is 35 and the number of sessions is 12, the speech files have no header and they are recorded in 16 bits.

4 System Configuration

We have developed a prototype in high level implementation that tries to evaluate system performance. Our system has three main blocks, pre-process, training and recognition task that we describe in the following points. To test the system we use the Data Base previously specified.

4.1 Pre-Process

To prepare the speech files we must do the following task:

- Eliminate the WAV header.
- Detect the noises and silences implementing a speech activity detector (SAD).
- Extraction of speech characteristic windowing the signal with a rectangle window of 25 ms overlapping the windows (speech frames) 12,5 ms.
  - ⇒ The feature vectors are created with 22 coefficients composed by 10 Mel-Cepstrum, 10 Differential Mel-Cepstrum , Energy and Differential Energy.

This process is common in both working phases: Training and recognition.

4.2 Training Process

This process generate a statistical model for each speaker based in a HMM with 3 states where the observation probability is given by a GMM model from the spectarl representation.

After performing a lot of tests and experiments, the best results are obtained training the system with the sessions from 1 to 6 in Carlos III Data Base and sessions from 1 to 3 in Polycode Data Base being the order of the mixture 32, 64 and 128. The process consists in making an hybrid model for each speaker.

4.3 Recognition Process

Once the speaker models had been generated and stored in the recognition base, we must test the reliability of the system, for that purpose we use some of the remaining voice in de Data Bases being the time of test more much smaller than the time of training, that is around 1 second of test voice and 1 minute of train voice.

To test the system we use sessions 7 and 8 of Carlos III and sessions from 4 to 8 of Polycode , the recognition process is made with a Viterbi algorithm calculating the error rate depending on rejects and false acceptances.

We test with independent sessions to notice the change in the recognition rate depending on time variability of the speech

5 Results

In this section we are going to present the results, after working with two Data Base, the first one is the Data Base Carlos III (CIII) and the second one is the telephone Data Base Polycode II.

The results are obtained using the training process 32, 64 and 128 mixtures in the GMM model.

In Figure 4 the results of recognition process are presented when are used 5 session to train the system and session 6 and 7 to text it. As you can see the recognition rate is smaller in the Polycode II, this is due to the noise introduced by the telephone line. The error rate is around 10% in telephone Data Base and around 8% in the Data Base CIII.

In Figure 5 a study of time evolution when the CIII are used in a system in whih 6 sessions had been used in the traing process, are showed.

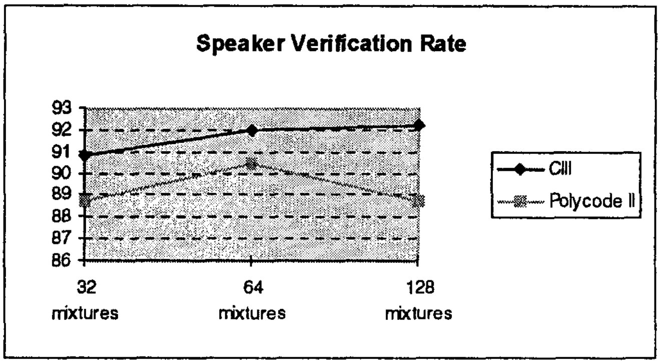

Fig. 4. Recognition rate in the test process.

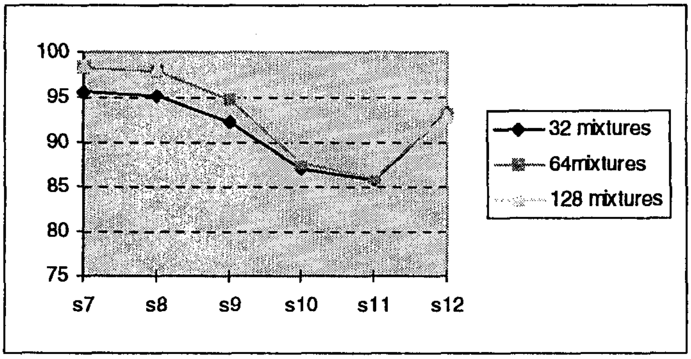

Fig. 5 Recognition rate in the test process.

6 Conclusions

Nowadays systems working under real time and real-world conditions are continuously used, and our implemented system tries to solve one of the most recent necessities, access control, this kind of systems are more sophisticated if the users access through networks. Observing the results in ours experiments we could conclude that:

- User identification through telephone line is most difficult and the recognition rate decreases. This is because telephone channel introduce a noises and distorsions in the communication.

- The time evolution of the speech is a very important question because a lot of factors must be considered, physical state that produce the modification of vocal tract (p.e. a cold, the age), psychological state (happiness, sadness, depression), etc. This is the reason for that the recognition rate is different in each session. When this factor is taked into account in the training process, using several session to train the system, the system performance increase higly. Our futur work will be wors in this sense introducing several techniques oriented to reduce this hughe effect.
- The choice of order model, that is, the number of gaussians mixtures are important to have the most adequate distribution. This is shown in Polycode II being 128 mixtures a worse distribution that 64 mixtures.
- Channel noise is another important factor to consider in the train proccess improving the system performance when the working system conditions are similar to train system conditions.

7 References

[1] Furui & Sondhi "Advances in Speech Signal Processing". Ed. MARCEL DEKKER, INC. 1989.

[2] Reynolds. (95) "Robust Test-Independent Speaker Identification Using Gaussian Mixture Speaker Models", Speech Communication 17 (1995) 91-108

[3] Ruiz-Mezcua, Lorenzo-Speranzini, García-Gomez. "Sistema de verificación automática de locutores", Internal Documente ALCATEL-SESA.

[4] Ruiz-Mezcua, Gerbolés-Espina, Escrihuela-Langa, Gomez-Mena, Veiga. (92) "Reconocimiento de grandes vocabularios independientes del locutor". URSI92 Conference.

[5] Ruiz-Mezcua, Hernadez, Domingo, Rodriguez. "Acceso a servicios multimedia a traves de la voz". URSI96 Conference.

[6] Veth & Bourlard "Comparison of Hidden Markov Model techniques for automatic speaker verification in real-world conditions", Speech Communication 17 (1995) 81-90