F-15

A Robust GMM Based Speaker Recognition System with Features Extracted from Voiced Parts

Eric Simancas Acevedo[1], Takayuki Nagai[2], Akira Kurematsu[2],
Mariko Nakano-Miyatake[1] and Hector Perez-Meana[1]
1. ESIME Culhuacan, National Polytechnic Institute of Mexico
Av. Santa Ana 1000, 04430 Mexico City, Mexico
2. The University of Electro Communications
Chofu-Shi, Tokyo, Japan

ABSTRACT

A robust text independent speaker identifications system is proposed in which the speaker is characterized using the pitch period and LPC-CEPSTRAL extracted from voiced parts. These features are feed into a Gaussian Mixture Models used for the speaker identification. The proposed system was evaluated using a data-base of 80 speakers, with phrases of 3-5s in Japanese language stored during 4 months. Evaluation results show that proposed system achieves a global recognition rate of more than 95%.

Keywords: GMM Model, speaker identification, pitch period, LPC-Cepstral, speaker recognition, voiced part.

1. INTRODUCTION

Several methods have been proposed for speaker identification such as: Vector Quantization (VQ), Dynamic Time Warping (DTW), Hidden Markov Models (HMM) and the Networks Neural (ANN) and Gaussain Mixture Model [9]. Among them, a very good method used for text independent (TI) speaker identification tasks is the Gaussian Mixture Model (GMM), which is similar to the HMM but the GMM omit the transition time between the states, so, the GMM uses only a unique Gaussian Distribution Matrix to represent to each speakers in the system [2].

In all pattern recognition task, the feature extraction is very important, then to do the system robust, the LPC-Cesptral coefficients are used, because their computation load is low and used together with the Gaussian Mixture Model lead to the development of robust the systems. However, when the LPC-Cepstral coefficients are calculated useful information such as the pitch is ignored, and this is an specific feature of the individual speaker identity.

Thus to improve previously proposed GMM TI speaker recognition systems, this paper proposes a GMM based speaker recognition system [2], using a speaker features vector derived from a combination of the pitch and LPC-Cepstral, extracted only from voiced parts, which enhance the. Computer evaluation results show that proposed system improve the recognition rate shown by the previously proposed GMM systems.

2. PROPOSED SYSTEM

The proposed system shown in Figure 1, estimate the features vector using a combination of the pitch period information, obtained derived by using he autocorrelation method, and the LPC-Cepstral coefficients derived from the LPC-analysis technique using only the voiced segments of speech signal which are given by

$$c(k) = -a_k + \frac{1}{n}\sum_{n=1}^{k-1}\frac{n}{k}c(n-k)a_k \text{ ,k=2,3,...M,} \qquad (1)$$

where a_k are the LPC coefficients vector of voiced parts, estimated using the Levinson method. Here, instead of the pitch value it is used ***log (F0)***, where F_0 is the inverse of the pitch period.

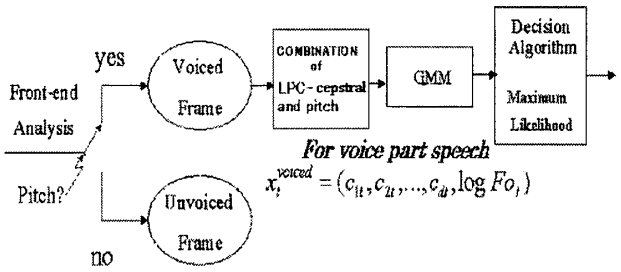

Figure 1. Proposed Speaker recognition System.

.It has been found that the pitch value carries specific speaker information, however it only appears in the voiced segments. Then in order to use the pitch information together with the LPC-Cepstral features, keeping a relatively low computational complexity, the combination of both features is done only if the signal frame is considered as a voiced part, an then the feature vector is given by equation (2). Thus the feature vector of proposed system consists of 17 data, 16 LPC-cepstral and the log(Fo) to represent to the pitch information.

$$x_t^{voiced} = (c_{1t}, c_{2t}, ..., c_{dt}, \log Fo_t). \qquad (2)$$

Identification Stage

This stage uses the Gaussian Mixture Model (GMM) shown in figure 2, which provides a model of the main speaker voice sound. This model only has one state with a Mixture Gaussian distribution that represents the different acoustic classes, and ignores the temporal information of the acoustic observation sequence. Thus the GMM in proposed system uses a full covariance matrix with 17 mixtures to obtain the optimum model for each speaker.

In GMM the features distributions of the speech signal are modeled for each speaker by using the sum of weighted Gaussian distribution of the speech signal of the speaker,

$$p(\mathbf{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\mathbf{x}), \text{ with } \sum_{i=1}^{M} p_i = 1 \quad (3)$$

where $\mathbf{x}$ is a random vector of D-dimension, $p(x/\lambda)$ is the speaker model, p_i are the mixture weights, $b_i(\mathbf{x})$ are the density components, that are formed by the mean μ_i and covariance matrix σ_i to i = 1,2,3,....M., and each density component is a D-Variate-Gaussian distribution of the form

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\} \quad (4)$$

where the mean vector, μ_i, variance matrix, σ_I, and mixture weights p_i of all the density components, determines the complete Gaussian Mixture Density. To obtain an optimum model for each speaker, the parameters μ_i, σ_i and p_i should be estimated iteratively until convergence is achieved. The initial condition $p(i|x_t, \lambda)$ is obtained using the Viterbi algorithm. Subsequently, the parameter p, the Mean vector $\boldsymbol{\mu}$ and Variance σ_i^2 should be updated until find the approximated optimum model.

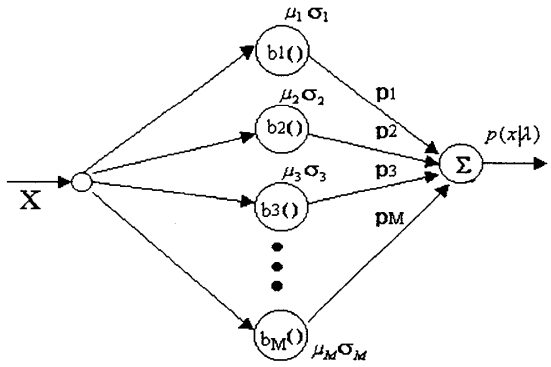

Figure 3. Gaussian Mixture Model

Decision Algorithm

After the models GMM for each speaker have been estimated, the target is to find the model wuth the maximum likelihood a posteriori for a observation sequence. Usually

$$\hat{S} = \arg \max_{1 \le k \le S} \Pr(\lambda_k | X) = \arg \max_{1 \le k \le S} \frac{\Pr(X|\lambda_k)\Pr(\lambda_k)}{p(X)} \quad (5)$$

where eq. (5) is given by the Bayes rule. Then assuming that all speaker have the same probability and noting that p(x) is the same one for all models of the speakers, the classifiers rule is reduced to

$$\hat{S} = \arg \max_{1 \le k \le S} \Pr(X|\lambda_k), \quad (6)$$

3. EVALUATION RESULTS

The proposed system was evaluated using a database of 80 different speakers with the pronunciation 10 different texts that contain 25 different Japanese phrases of 3-5s stored 4 times; each time every one month. Figure 3 shows the recognition results obtained using the whole speech signal without pitch information, and Figure 4 shows the recognition results obtained using only the voiced segments incorporating the pitch information. Evaluation results show that although the system proposed in [2] provides fairly good recognition performance, the recognition rate of some speakers may be lower than 80%. This difficulty was solved using only the voiced including pitch information, achieving, in all cases, recognition rates higher than 90%.

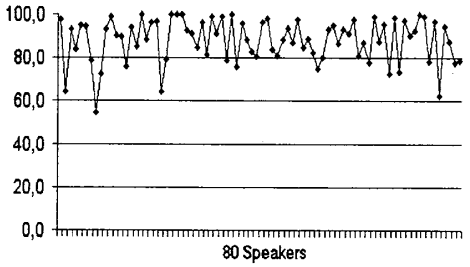

Figure 4. Recognition performance using voiced and unvoiced segments

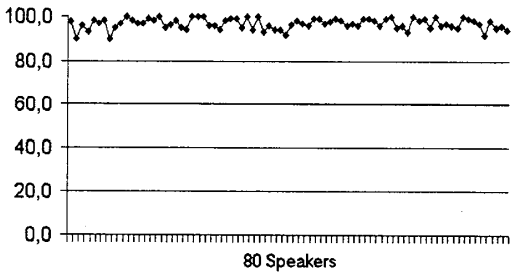

Figure 5. Recognition performance using voiced segments and pitch information

4. CONCLUSIONS

This papers proposed a TI Speaker identification system using the LPC-cepstral and pitch information as speaker features. Evaluation results shown that the use of this combination at vector level improves the recognition rate. To reduce the computational complexity of proposed system, only the voiced part is used. These modifications allowed us to improve the system performance obtaining a global recognition rate of about 98%.

REFERENCES

[1] Konstantin P. Markov and Seiichi Nakagawa. "Integrating pitch and LPC-residual information with LPC-cepstrum for text-independent speaker recognition", J. Acoust Soc. Jpn (E), 20, 4, pag 281-291,1999.

[2] 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, pag. 287-294, Springer Verlag, Berlin, 2001.