Application of GMM in the speaker identification system

ZENG Chun[1]
Chongqing Technology and Business Institute,
Chongqing, 400052, China

LI Zhong[2]
Chongqing College of Electronic Engineering,
Chongqing, 401331, China

Abstract—**This paper discusses the application of speaker identification technology from several aspects such as speaker identification using GMM, optimization of identification method, system implement and experimental result, and analyzes application prospects of CMM in the speaker identification system.**

Keywords:* *Speaker Recognition, GMM model, parameter estimation

I. INTRODUCTION

Automatic Speaker Recognition (ASR) is the most attractive and challenging topic. It has the widely application prospects, such as security identity (confidential places access control), forensic investigation (crime monitoring and identity), military (battlefield environment monitoring and commander identity), information services (automatic information retrieval or electronic commerce), etc.

II. GAUSSIAN MIXTURE MODEL (GMM)

Gaussian mixture model (GMM) can be seen as one state CHMM. A GMM is a weighted sum of M component densities and is given by the form

$$p(x/\lambda)=\sum_{i=1}^{M}P_I b_i(x) \qquad (1)$$

Where x is a d-dimensional random vector, $b_i(x_t), i=1,....,M,$ is the component density. $P_i, i=1,....,M,$ is the mixture weight. Each component density is a D- dimensional Gaussian function of the form

$$b_i(x)=\frac{1}{(2\pi)^{D/2}\,|R_i|^{1/2}}\exp\{-\frac{1}{2}(x-\mu_i)^{'}R_i^{-1}(x-\mu_i)\}(2)$$

With mean vector μ_i and covariance matrix R_i .mixture weights satisfy the constraint that

$$\sum_{i=1}^{M}P_i=1, \qquad (3)$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights, which is defined as

$$\lambda=\{P_i,\mu_i,R_i\}, i=1,...,M \qquad (4)$$

In speaker recognition system, each speaker is represented by such a GMM and is defined as the reference model λ. Hence the likelihood functions $P(X\mid\lambda_i), i=1,..,N,$ in the sequence of X test vectors are estimated using Eqs (4-6). We can find the speaker whose mode λ_i maximizes a posteriori probability $P(X\mid\lambda_i)$.

The likelihood of sequence $X=\{x_t\}t=1,2,\cdots,T$ in the log domain is written as

$$L(X/\lambda)=\frac{1}{T}\sum_{t=1}^{T}\log p(x_t/\lambda) \qquad (5)$$

978-1-4244-6252-0/11/$26.00 ©2011 IEEE

III..MAXIMUM LIKELIHOOD (ML) ESTIMATION AND EXPECTATION-MAXIMIZATION(EM)

Maximum likelihood (ML) estimation is a most common and effective method. The aim of ML estimation is to find the model parameters which maximize the likelihood $L(\theta)$. Based on the properties of $L(\theta)$, the model parameters is selected by maximizing $L(\theta)$ in the log domain.

For the ML in speaker recognition system, let $X_1,\cdots,X_N$ training data, U,Σ are the estimation parameters (mean vector and covariance matrix). Therefore likelihood in the log domain is defined as

$$L=\ln f(X_1,\cdots,X_N\mid U,\Sigma) \qquad (6)$$

As the training data $X_1,\cdots,X_N$ are independent each other, hence

$$\begin{aligned} L&=\ln f(X_1,\cdots,X_N\mid U,\Sigma)\\ &=\ln\prod_{j=1}^{N} f(X_j\mid U,\Sigma)=\sum_{j=1}^{N} f(X_j\mid U,\Sigma) \end{aligned} \qquad (7)$$

Where

$$f(X_j\mid U,\Sigma)=\frac{1}{(2\pi)^{D/2}\mid\Sigma\mid^{1/2}}\exp\{-\frac{1}{2}(X_j-U)^T\Sigma^{-1}(X_j-U)\}$$

It can be simplified as

$$\begin{aligned} L&=\ln\{\frac{1}{(2\pi)^{D/2}\mid\Sigma\mid^{1/2}}\exp[\sum_{j=1}^{N}-\frac{1}{2}(X_j-U)^T\Sigma^{-1}(X_j-U)]\}\\ &=-\frac{D}{2}\ln(2\pi)-\frac{1}{2}\ln(\mid\Sigma\mid)-\frac{1}{2}\sum_{j=1}^{N}(X_j-U)^T\Sigma^{-1}(X_j-U) \end{aligned} \qquad (8)$$

The ML estimation of U,Σ are $\hat{U}_{ML},\hat{\Sigma}_{ML}$, which are obtained by

$$\begin{cases} \dfrac{\partial L}{\partial U}=-\dfrac{1}{2}\Sigma^{-1}\sum\limits_{j=1}^{N}(X_j-U)=0\\ \dfrac{\partial L}{\partial \Sigma}=-\dfrac{1}{2}\Sigma^{-1}+\dfrac{1}{2}\Sigma^{-2}\sum\limits_{j=1}^{N}(X_j-U)^2=0 \end{cases} \qquad (9)$$

Solving the Eq(9) and obtaining

$$\begin{cases} \hat{U}_{ML}=\dfrac{1}{N}\sum\limits_{j=1}^{N}X_j\\ \hat{\Sigma}_{ML}=\dfrac{1}{N}\sum\limits_{j=1}^{N}(X_j-U)^2 \end{cases} \qquad (10)$$

The mean and variances are unbiased, which are defined as follow

$$\begin{cases} U=\dfrac{1}{N}\sum\limits_{j=1}^{N}X_j\\ \Sigma=\dfrac{1}{N-1}\sum\limits_{j=1}^{N}(X_j-U)^2 \end{cases} \qquad (11)$$

Therefore, $\hat{U}_{ML}$ is unbiased and $\hat{\Sigma}_{ML}$ is biased. $\hat{\Sigma}_{ML}$ can be eliminated by multiplying a constant $N/(N-1)$.

EM is the expectation-maximization. The basic idea of the EM algorithm is, beginning with an initial model λ, to estimate a new model $\overline{\lambda}$, such that $p(X\mid\overline{\lambda})\geq p(X\mid\lambda)$. The new model then becomes the initial model for the next iteration and process is repeated until some convergence threshold is reached. This is the same basic technique used for estimating HMM parameters via the Baum-Welch reestimation algorithm. On each EM iteration, the following reestimation formulas are used which guarantee a monotonic increase in the model's likelihood value:

Mixture Weights:

$$C_m^i=\frac{\sum\limits_{t=1}^{T}\gamma_{tlm}^i}{\sum\limits_{t=1}^{T}\sum\limits_{m=1}^{M}\gamma_{tlm}^i} \qquad (12)$$

Means:

$$\mu_m^i=\frac{\sum\limits_{t=1}^{T}\gamma_{tm}^i o_t}{\sum\limits_{t=1}^{T}\gamma_{tm}^i} \qquad (13)$$

Variances

$$\Sigma_m^i=\frac{\sum\limits_{t=1}^{T}\gamma_{tm}^i(o_t-\mu_m^i)(o_t-\mu_m^i)'}{\sum\limits_{t=1}^{T}\gamma_{tm}^i} \qquad (14)$$

Where α is arbitrary element of the vectors; o_t is the observation vectors of t frame. m is the order number.

$$\gamma_{tm}^i=\frac{C_m^{i-1}N(o_t\mid\mu_m^{i-1},\Sigma_m^{i-1})}{\sum\limits_{m=1}^{M}C_m^{i-1}N(o_t\mid\mu_m^{i-1},\Sigma_m^{i-1})} \qquad (15)$$

3 GMM BASED VERIFICATION SYSTEM

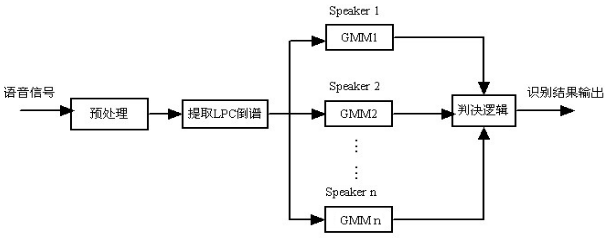

Fig. 1. Block diagram of our speaker verification system.

Based on the Bayes rule, the maximum posteriori probability is

$$p(\lambda_i / X) = \frac{p(X / \lambda_i) p(\lambda_i)}{p(X)} \tag{16}$$

Where

$$P(X / \lambda) = \prod_{t=1}^{T} p(x_t / \lambda) \tag{17}$$

In the log domain, it is expressed

$$\log P(X / \lambda) = \sum_{t=1}^{T} \log p(x_t / \lambda) \tag{18}$$

As the priori probability of $p(\lambda_i)$ is unknown, we assume that it is the same for all speaker models, namely

$$p(\lambda_i) = \frac{1}{N} \qquad 1 \le i \le N \tag{19}$$

For a given observation sequence, $p(X)$ is a determined constant and equal for all the speakers. Therefore, the maximum posteriori probability can be obtained by computing $p(X / \lambda_i)$. The classification rule simplifies to

$$i^* = \arg\max_i p(X / \lambda_i) \tag{20}$$

Where i^* is the identified speaker.

IV. EXPERIMENTAL RESULTS

The speech data is captured in laboratory circumstances. The database consists of 10 speakers (5 males and 5 females). Each one speaks 30 different words and each word repeats 6 times. Therefore, there are 1800 speech. The speakers are recognized by the 14th order LPC based on CMM. The first three words for all speakers (900 words) are used for training and obtain the parameter of each speaker. The remaining three words are used for testing. The experiment shows the max recognition rate is obtained when the dimension of GMM M=25.

Table 1 lists the recognition rate using normalized exponent transform GMM when M=25. Due to the influence of experimental circumstances, the recognition rate is less than that of [28]. From table 1, we can see that the recognition rate can be improved when n becomes smaller. When n>1/6, the recognition rate can not be significantly improved. In addition, it can be also improved by selecting an appropriate θ. When n=1/6, the error recognition rate is reduced to 5.3%. ML in the table is the log transformation.

Table 1 Identification rates (%) using normalized exponent transform

	ML	n=1/8	n=1/6	n=1/4	n=1/2	n=1	n=2
θ=0	0.869	0.912	0.913	0.902	0.893	0.888	0.858
θ=0.14	0.869	0.915	0.922	0.912	0.897	0.893	0.863

Table 2 Identification rates (%) using normalized exponent transform with different T

T	100	120	140	163
ML	0.8	0.824	0.858	0.869
n=1/6	0.834	0.873	0.909	0.922

改善	0.034	0.049	0.051	0.053

Table 2 shows the identification rates can be further improved when the testing frames number T increases.

In the text-independent speaker recognition system using GMM, actual circumstance and personal factors degrade the identification rate. Based on the statistical properties of the frame likelihood model, this paper analyzes the linear transformation can not improve identification rate, and proposed a novel transformation called normalized exponent transformation. The theoretical analysis and experimental results show, as compare to log transformation, the normalized exponent transformation can achieved to 5.3% identification rate.

V. CONCLUSIONS

This paper proposes a GMM based speaker recognition method using mixture feature parameters such as LPC cepstrum coefficient, difference cepstrum coefficient, fundamental tone and difference fundamental tone, as feature vector. To test the performance of speaker recognition system, a database consisting of 10 speakers with 1800 sentences is built. The speech database contains monosyllabic, two syllables and quadrisyllable words.

The speaker recognition system can recognize the speech in the database and test the performance of feature parameters. Meanwhile, it can identify the speaker in real time. Based on the statistical properties of the frame likelihood model, this paper analyzes that the linear transformation can not improve identification rate, and proposes a novel transformation called normalized exponent transformation using GMM based recognition system. The theoretical analysis and experimental results show, as compare to log transformation, the normalized exponent transformation can achieved to 5.3% identification rate.

REFERENCES

[1] Konstantin P. Markov, Seiichi Nakagawa, Text-independent speaker recognition using non-linear frame likelihood transformation, Speech Communication, pp193-209, 1998

[2] T. Matsui and S. Furui, Comparison of text-dependent speaker recognition methods using VQ-distortion and discrete/continuous HMMs, IEEE Proc ICASSP'93, Vol.II, pp391-394, 1993

[3] L. Zhao, speech signal processing, Machine Press, 2003

[4] H. HU, speech signal processing, Harbin University of Industry Press , 2000

[5] J.H Xie, HMM and its application to speech processing, Huazhong University of Science and Technology Press, 1995

Acknowledgement:
This research supported by the Project of Chongqing Education Commission (KJ. 101601).