http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 11 (7): 799-803, 2012
ISSN 1812-5638 / DOI: 10.3923/itj.2012.799.803

Application of Speaker Recognition Based on LSSVM and GMM Mixture Model

Haiyan Yang, Xinxing Jing and Ping Zhou
Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

Abstract: The technique of speaker recognition is becoming mature. A system of speaker recognition based on Least Squares Support Vector Machine (LSSVM) and Gaussian Mixture Model (GMM) mixture model is discussed in this study. The designed system will be considered to application in Internet environment. The performances of different feature parameters are compared such as LPCC, MFCC and WPTMFC in this study. After comparing recognition rate of LSSVM model and LSSVM-GMM mixture model, recognition rate of the system using the mixture model based on LSSVM-GMM is better and the mixture model is chose. The final results show that the system has good recognition rate and is potential for the practical applications.

Key words: WPTMFC, least squares support vector machine, Gaussian mixture model

INTRODUCTION

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech signal. These techniques are developed by Bell Laboratories for military intelligence purposes (Huang, 2002). Establishing classification model is the basic problems in speaker recognition system.

With the developing of the science and technology, the information systems move forwards intellectualization. Speaker recognition technology become more important and is applied to many fields, such as automobile voice lock, identification system, VoIP (Voice over Internet Protocol) telephone system and so on (Naik *et al*., 1989). In fact, real-time voiceprint identification system on the Internet is quite popular. Furthermore, the system designed in this study will adapt to the condition of the Internet environment.

In this study, we will apply LSSVM-GMM (Least Squares Support Vector Machine (LSSVM) and Gaussian Mixture Model (GMM)) mixture model to speaker verification (Suykens and Gestel, 2003). It can provide complimentary information to the Gaussian mixture model and the support vector machines and has high recognition rate and availability.

CLASSIFICATION MODEL

A system of speaker recognition consists of two stages: training and recognition (Rabiner and Juang, 1993). In the training phase, a number of training voice are processed to digital signals and the speaker's characteristic parameter is extracted, then the classification model is established and stored for each user. In the recognition stage, comparing features extracted from speaker's voice in the training stage, the recognition result is obtained based on certain similarity criteria (Reynolds *et al*., 2000).

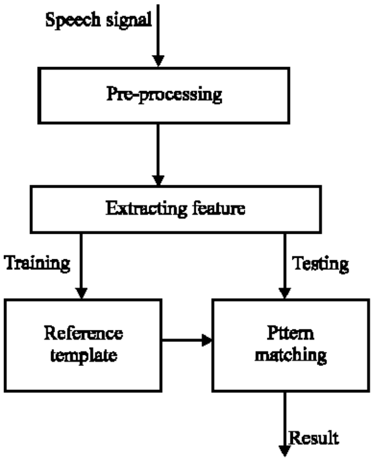

Fig. 1: The basic structure of speaker recognition system

The basic structure of the typical speaker recognition is shown in Fig. 1, among which model training has a great influence on the recognition rate.

For the text independent pattern recognition, studies have shown that GMM and SVM are the best recognition models now (Zhan and Jing, 2010). GMM is a generative model based on probabilistic, which can reflects the similarity of characteristics from a statistical point. SVM is a method based on discriminated model, which reflects the difference between heterogeneous data (Wu and Lin, 2009) by finding the optimum classification of surfaces between the different categories.

Corresponding Author: Haiyan Yang, Guilin University of Electronic Technology, Guilin 541004, Guangxi, China

Based on previous studies, this article will consider the mixture model of GMM and LSSVM to set up recognition system.

SYSTEM DESIGN

This study is to design a text-independent speaker identification system on the environment of Internet network. System framework is shown in Fig. 2.

Network construction: There are three solutions for internet voiceprint recognition System: server-only processing mode, client-only processing mode and client-server processing mode. Compared these patterns, the client-server processing mode, which extracting feature parameters in the client and identifying in the server, is more suitable for voiceprint identification based on internet.

Client-Server mode(C/S mode) based on the distributed concept. There is two different function parts in C/S mode. One is the server program that responds to and offer fixed services; the other is the client program that sends requirement to the server. C/S mode has some advantage for small bandwidth, such as a simple processing in the client, small account computation, convenient for a variety of clients. In the paper C/S mode is chosen: extracting parameters in the client and identifying in the server side and displaying the final recognition results in the client.

This network system will transmit a set of speech feature sequences which is pre-processed, ordered, high reliability characteristics data. TCP protocols has some advantages, such as connection-oriented, full duplex, controlled response and flow, reliability of data transmission. Therefore, we use TCP protocols to build a network transmission system. The system is developed by CSocket (Valyon, 2007) class of MFC (Microsoft Foundation Class) of Visual C^{++} 6.0 software.

Pre-processing and feature extraction in the client: In the system, processed speech feature is transmitted in the network, which avoiding system identification performance degradation caused by the original voice.

The voice data are chosen from the TIMIT speech database which is the standard speech database used to evaluate of speaker recognition systems. We choose 138 speaker's voice data. Each statement is 6 digits of English pronunciation, length of about 2 to 3 sec. Training speech for each speaker is 4 parts, each part has 24 statements. Test speech has 10 voice parts, each part has four statements. In order to test the system performance, we have used self-built voice database of 30-person. Using Cool Edit Pro software, 10 sec voice of random number strings is recorded for everyone in the general laboratory. When training and recognition, using 8 KHz sampled, 16 bit linear PCM coding, the length voice is selected.

First, there is a pretreatment of each speech signal, for instance noise reduction, muting, sampling, framing, windowing (the length of Hamming is 32 msec, shift is 16 msec) and pre-emphasis (using filter $H(z) = 1\text{-}0.97\ z^{-1}$).

After pretreatment, it is necessary to extract various parameters, including pitch, LPCC and differential parameters, MFCC and differential parameters and other derived parameters.

In the follow experiment, we will detect its performance.

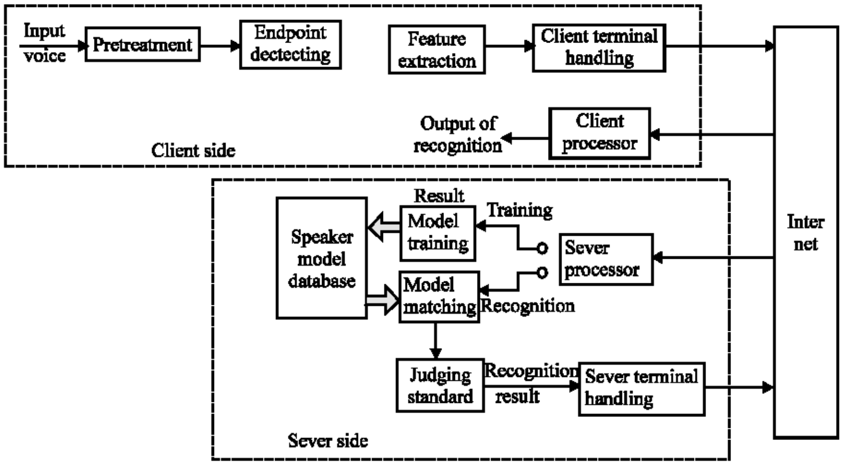

Fig. 2: The block diagram of voiceprint identification system based on internet

Table 1: Identification results under different kernel functions

Kernel type	Parameter settings	Recognition rate (%)
Polynomial	$\gamma = 0.5$, $c = 1$, $q = 3$, $C = 10^5$	88.9
RBF	$\sigma = 0.0625$, $C = 10^3$	93.0
Sigmoid	$\gamma = 0.5$, $c =1$, $C = 10^4$	91.2

Design of the classification on server: For training experiment, the length of voice is 20 sec. For testing experiment, the length of voice is 4 sec. The number of identified people is 40. The parameters is MFCC (13 dimensions) in this study.

First, this paper uses LSSVM and designs a SVM for trainer. The total number of SVM is 40. Recognition results are shown in Table 1.

From Table 1, we can see that LSSVM classifier, constructed by RBF kernel function, has higher recognition rate. So, RBF kernel function is used to construct LSSVM classifier in this article.

GMM is a method based on probabilistic. GMM can reflect similarity of same kind of data and can describe data distribution from statistical point. LSSVM is a method based on discriminate model. LSSVM can reflect difference of heterogeneous data. It is obviously that recognition principle for GMM and LSSVM is different and there are complementary aspects in reducing the errors. Therefore, this article considers combining GMM with LSSVM for recognition system.

The combination methods are as follows:

- Establishing Gaussian mixture model for each class of SVM. For a given vector x, output probability of the ith of GMM:

$$P_{GMM}\,(X/C_i)=\sum_{m=1}^{M} c_{im} N\left(x,\mu_{im},\sum{}_{im}\right) \tag{1}$$

Where:

$$N\left(x,\mu_{im},\sum{}_{im}\right)=\frac{1}{(2\pi)^{d/2}\left|\sum{}_{im}\right|^{1/2}}\exp\left[-\frac{1}{2}(x-\mu_{im})^T\sum{}_{im}^{-1}(x-\mu_{im})\right] \tag{2}$$

where, c_{im}, μ_{im} and Σ_{im} is, respectively, weights, mean and covariance of the mth Gaussian model for the ith GMM

- The probability output form of LSSVM using Sigmoid function:

$$\begin{cases} P(C_{+1}/\mathbf{x})=\dfrac{1}{1+\exp(-f(\mathbf{x}))} \\ P(C_{-1}/\mathbf{x})=\dfrac{1}{1+\exp(-f(\mathbf{x}))} \end{cases} \tag{3}$$

where, C_{+1} and C_{-1} is collection of samples of class corresponding to +1 and -1 class respectively. For the point at the surface of classification, if there is f (x) = 0, then the probability is 0.5 for the corresponding class -1 and +1

- If the probability output of GMM is embedded in the probability output of SVM, the output information will contain not only the information between different categories but also the information within each class. So mixture model based on LSSVM-GMM will be used. Assume that input vector is x, the posterior probability is:

$$\begin{aligned} P(C_{+1}/\mathbf{x})&=\frac{1}{1+\exp\left[-f(\mathbf{x})S(\mathbf{x},C_{+1})\right]} \\ P(C_{-1}/\mathbf{x})&=\frac{1}{1+\exp\left[-f(\mathbf{x})S(\mathbf{x},C_{-1})\right]} \end{aligned} \tag{4}$$

where, S(x, C_{+1}) and S(x, C_{-1}) is adjustment factors, their value are:

$$\begin{aligned} S(\mathbf{x},C_{+1})&=\begin{cases} P_{GMM}\,(C_{+1}/\mathbf{x})+0.5 & f(\mathbf{x})\geq 0 \\ 1.5-P_{GMM}\,(C_{+1}/\mathbf{x}) & f(\mathbf{x})<0 \end{cases} \\ S(\mathbf{x},C_{-1})&=\begin{cases} 1.5-P_{GMM}\,(C_{-1}/\mathbf{x}) & f(\mathbf{x})\geq 0 \\ P_{GMM}\,(C_{-1}/\mathbf{x})+0.5 & f(\mathbf{x})<0 \end{cases} \end{aligned} \tag{5}$$

Where:

$$P_{GMM}(C_{+1}/\mathbf{x})\frac{P_{GMM}(\mathbf{x}/C_{+1})}{P_{GMM}(\mathbf{x}/C_{+1})+P_{GMM}(\mathbf{x}/C_{-1})} \tag{6}$$

$$P_{GMM}(C_{-1}/\mathbf{x})\frac{P_{GMM}(\mathbf{x}/C_{-1})}{P_{GMM}(\mathbf{x}/C_{+1})+P_{GMM}(\mathbf{x}/C_{-1})} \tag{7}$$

Here, the purpose of the adjustment factors S(x, C_{+1}) and S(x, C_{-1}) is adjust the output of LSSVM depending to the results of GMM.

The simulation results are as follows. For training experiment, the length of voice is 20 sec. The number of identificated people is 40. The parameters is mixed features (13 dimensions) combined LPCC and WPTMFC. Classifier uses LSSVM-GMM, the parameter settings is such as: 32 order GMM, RBF kernel function for LSSVM. Recognition results are shown in Table 2.

As can be seen from Table 2, recognition rate increases when test speech length increases. Of course, the length of test speech will increase computation, the real-time processing will be affected. Consider the user requirements, it is necessary for a compromise consideration, so we choose the test speech length 4 sec.

Table 2: Recognition results under different test speech length

	1	2	4	6	8
Length	------------------------(sec)------------------------				
Recognition rate (%)	84.6	90.2	96.9	97.3	97.8

Table 3: Recognition results under different characteristic parameters

Feature	LPCC	MFCC	Mixed parameters
Recognition rate (%)	82.6	88.2	93.4

Table 4: Recognition results under different model

Model	LSSVM	LSSVM-GMM
Run time for test (sec)	1.204*40	2.976*40
Recognition rate (%)	92.5	96.9

For training experiment, the length of voice is 20 sec. For testing experiment, the length of voice is 4 sec. Choosing kernel function RBF, recognition results are shown in Table 3.

In Table 3, we can see that it have better recognition rate using LPCC and MFCC in the LSSVM-GMM mixture model. Mixed parameter has highest recognition rate.

From the above experiments, we determine all the parameters. Using the selected parameters we compared the effective of different models on the recognition results.

For training experiment, the length of voice is 20 sec. For testing experiment, the length of voice is 4 sec. The number of identificated people is 40. The parameters is mixed features (13 dimensions) combined LPCC and WPTMFC. 32 order GMM. RBF kernel function for LSSVM. Recognition results are shown in Table 4.

Analyzing Table 4:

- it was observed that run time for test is longer using LSSVM-GMM model than using LSSVM model. In fact, run time for test has only small difference for one sentence and human can accept the run time for test
- Recognition rate is higher using LSSVM-GMM mixture model than using LSSVM. Combinated GMM and LSSVM, the recognition rate was further improved. The two models have different mechanisms: SVM belong to discriminant model, while GMM is a model based on probability. If GMM model is used to adjust the output of LSSVM, it has good complementarity. So mixed model has better coverage for speaker feature space

Therefore, we select the identification method based on LSSVM-GMM in the server.

EXPERIMENT ANALYSES

The overall design of the system is based on two parts: text-independent speaker identification and Internet network transmission technology. The part of the network transmission, considering the requirements of voiceprint identification system and TCP/IP protocol, we choose C/S model based on TCP protocol which has high reliability. The part of text-independent speaker identification, considering the requirements of noise disturbing and real-time processing in network transmission, we improve the method of feature extraction and classifier to make it more suitable for network environments.

In this study, we consider two cases: no network transmission and network transmission.

For training experiment, the length of voice is 20 sec. For testing experiment, the length of voice is 4 sec. The number of identificated people is 40. LSSVM-GMM is chosen as classifier, the parameter set: 32 order GMM. RBF kernel function for LSSVM.

The line chart of recognition rate is shown in Fig. 3.

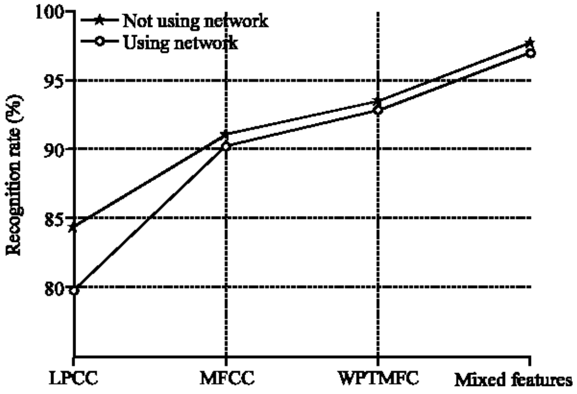

Fig. 3: Recognition rate for different characters under network environment

From Fig. 3, it can be seen that the recognition rate has decreased under network transmission. When the data is transmitted in the network, noise causes on bit errors, which reduces the recognition efficiency. From Fig. 3, it is shown that recognition rate is mostly affected using LPCC parameters, while recognition rate has small difference using MFCC, WPTMFC or mixed features. It is proved that MFCC parameters have good anti-noise performance.

CONCLUSION

This study analyzes the status of voice transmission and pattern recognition on the Internet and then chooses client/server model as the practical application and then design Internet-based voice pattern recognition system. This study also studies feature extraction techniques under the condition of network transmission and then gives a computation, robustness mixed parameters. On the server side, LSSVM is chosen as recognition model on the basis of comprising GMM and LSSVM. Finally, using CSocket of TCP/IP protocol under Windows network, it is completed by internet-based voice pattern recognition system.

As you know, the study of internet-based voiceprint identification system has challenge and practical. The designed system in this study is only a step in the laboratory, so it is necessary to study the security problem of the system, SVM multi-class problem, model database update problem on the server. If the three problems can be solved, the superior performance of Internet voice pattern recognition system will be applied to real life soon.

ACKNOWLEDGMENT

This study is supported by the project of Study of improved SVM implicated speech recognition system (Project Number: 609661002) of National Natural Science Foundation in China.

REFERENCES

Huang, Y., 2002. Internet Voice Communication Technology and Application. Posts and Telecom Press, Beijing.

Naik, J.M., L.P. Netsch and G.R. Doddington, 1989. Speaker verification over long distance telephone lines. Proceedings of the International Conferences on Acoustics, Speech and Signal Processing, May 23-26, 1989, Glasgow, pp: 524-527.

Rabiner, L. and B.H. Juang, 1993. Fundamentals of Speech Recognition. Prentice Hall, New Jersey.

Reynolds, D.A., T.F. Quatieri and R.B. Dunn, 2000. Speaker verification using adapted Gaussian mixture models. Digital Signal Process., 10: 19-41.

Suykens, J.A.K. and V.T. Gestel, 2003. Least Squares Support Vector Machines. World Scientific Publisher, Singapore.

Valyon, J., 2007. Extend least squares LS-SVM. Int. J. Comput. Intell., 3: 234-242.

Wu, J.D. and B.F. Lin, 2009. Speaker identification using discrete wavelet packet transform technique with irregular decomposition. Expert Syst. Appl., 36: 3136-3143.

Zhan, L. and X.X. Jing, 2010. Speaker recognition system based on the VQ-MAP and LS-SVM fusion. Electron. Technol. Appl., 6: 155-157.