Speaker Recognition and Speech Emotion Recognition Based on GMM

Shupeng Xu (*Corresponding author) ,Yan Liu and Xiping Liu
Computer Science and Engineering Department
Changchun University of Technology
Changchun, China
574790541@qq.com,
liuyan_5542@163.com,
759437907@qq.com

***Abstract*—This paper put forward a method for speaker recognition and speech emotion recognition based on GMM. The reason of using this model is its ability to the dependency among extracted speech signal feature vectors and the multi-modality in their distribution. Firstly, we extracted the Mel Frequency Cepstral Coefficients (MFCC) from each frame of the speech signal as speech features, and then apply Gaussian mixture model as a statistical classifier. The results of experiment show that we proposed method is effective than others.**

Keywords-Speaker recognition; speech emotion recognition; MFCC; GMM

I. Introduction

Speaker recognition and emotion recognition is an comprehensive research project which crosswise uses psychology, pattern recognition, speech signal processing and artificial intelligence. Speaker recognition is the identification of the people who is speaking by the characteristics of their voices. Speaker recognition is contains of speaker identification and speaker verification. The methods of speaker recognition can be divided into text independent and text dependent [1]. However, every person has different emotions when they are talking. The person's voice not only provides the meaning of spoken words, but also contains speaker dependent characteristics. For examples, the identity, the gender, the emotional state or the age of the speaker and so on. Especially in the criminal investigation of the police and military, it can get important applications. For example, in a variety of telephone extortion, kidnapping, telephone personal attacks and other cases, voiceprint recognition technology can find out suspect from record, and when the suspect is at trial, can analyze the implicit emotional of the suspect. Speaker recognition and emotion recognition plays an important role in not only security field but also human-computer interaction field. In the computer speech recognition system, the recognition accuracy rate for emotional speech is only 50% ~ 60%[2] , while an accuracy of 90% is achieved in neutral voice recognition. If we can consider the emotion information of the speech signal, the intelligentization and personification of artificial intellegence will be improved correspondingly, and the barriers between people and machines can be eliminated. Therefore, speaker recognition and emotion recognition research has gotten more and more attention in the field of speech signal processing.

Speaker recognition system contains feature extraction, model training, pattern matching, and logical judgments. In the earlier study, the speaker recognition experienced three stages: human ear recognition, spectrogram recognition and the machine automatical recognition. The research focuses have been transferred, from the feature parameters selection and extraction to the speaker recognition model. Representative of the characteristic parameters are: voice short-term energy, pitch, Mel Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and so on. The current in feature matching methods used in speaker recognition include Dynamic Time Warping (DTW), Principal Component Analysis (PCA), Vector Quantization (VQ), Hidden Markov Modeling (HMM), Artificial Neural Network (ANN), and the combination of these methods techniques. Similarly, the speech emotion recognition also needs to extract short acoustic and prosody's feature parameters reflecting emotion, and distinguish through a variety of classifier means. There are currently more popular method, Vector Quantization (VQ), Principal Component Analysis (PCA), Gaussian mixture model (GMM), K Nearest Neighbor method(KNN), Artificial Neural Networks(ANN) and so on. Although speaker identification and emotion recognition both belong to the scope of speech signal processing, their approach have different points and also the same place. For the same place, we can use the same features and the same method. However, the obvious differences exist in their different models, that is, speaker recognition part is to establish everyone's speak model, however, emotion recognition is to build the emotion model. In this paper, GMM model is used to deal with speaker identification and emotion recognition problems, which its simpleness, flexibility and robustness, has been widely applied to the field of speech recognition.

II. MFCC

It is show that MFCC can seize the acoustic characteristics for speech emotion recognition, speaker recognition, and so on[3-7]. Psychophysical studies have shown that people sensation of the frequency contents of sounds for speech signals follow a nonlinear scale. In a word, each tone with an actual frequency measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale. The mel-frequency scale follows a linear scale when frequency spacing between 0 and 1000HZ, but a logarithmic spacing above 1000Hz.As a reference point, the pitch which is a 1kHz tone and 40db above the perceptual hearing threshold, is defined as 1000mels. So we always compute the mels by using the following approximate formula for a given frequency f in Hz:

$$F_{mel} = 2595 \lg\left(1 + \frac{f}{700}\right) \quad (1)$$

First, x(n) is transferred into frequency domain form the time domain signal by an M point discrete Fourier transform (DFT).

$$X(K) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad n \ge 0, k \le N-1 \quad (2)$$

In the next step, for the log Mel-Frequency Cepstrum Coefficients (MFCC), the discrete cosine transform is done for transforming the mel coefficients back to time domain form frequency domain.

$$X(K) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi nk/N} \quad n \ge 0, k \le N-1 \quad (3)$$

Where, K=1, 2,…,K.

III. GMM MODEL SPEAKER AND EMOTION RECOGNITION

A. *The concept of GMM*

A Gaussian mixture model is a weighted sum of M component Gaussian densities as given by the equation,

$$P\left(x/\lambda\right) = \sum_{i=1}^{M} \omega_i b_i(x) \quad (4)$$

Where, data vector x, is a D-dimensional continuous-valued data vector, ω_i, i = 1, ..., M, are the mixture weights. $b_i(x_t)$, i = 1, ..., M, are the component Gaussian densities, and each component density is a D-variate Gaussian function ,

$$b_i(X) = \frac{1}{2\pi^{D/2} \left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(X-\mu_i)^t \Sigma_i^{-1} (X-\mu_i)\right\} \quad (5)$$

With mean vector μ_i and covariance matrix $\sum_i$. The mixture weights satisfy the stochastic constraint that

$$\sum_{i=1}^{N} \omega_i = 1 \quad (6)$$

The Gaussian mixture model is parameterized by the covariance matrices, mean vectors and mixture weights from all component densities. These parameters are common represented by the notation as following,

$$\lambda = \{\omega_i, \mu_i, \sum_i\} \quad i = 1, ..., M \quad (7)$$

For a T training vectors, $X = (\overrightarrow{X_1}, \overrightarrow{X_2}, \dots, \overrightarrow{X_T})$, The GMM likelihood, assuming independence of each other the vectors , can be written as following,

$$P\left(x/\lambda\right) = \prod_{t=1}^{T} P\left(x_t/\lambda\right) \quad (8)$$

In the log domain, it can be written as following:

$$\log P\left(x/\lambda\right) = \sum_{t=1}^{T} P\left(x_t/\lambda\right) \quad (9)$$

B. *GMM model speaker and emotion recognition model*

We consider using GMM recognition system to solve the problem of speaker recognition and speech emotion recognition. In the testing phase, it is essential to judge from a given set of emotions the one, which is most likely produced by a certain unknown testing utterances. Assuming all the emotions or speakers have same prior probability, the Bayesian decision rule reduces to the ML decision rule. X means the speech feature vectors which can be extracted from the test utterance, the index denotes the most likely emotion or speaker, i^* produced by the unknown testing utterance is given by the following formula,

$$i^* = \arg\max_i P(X/\lambda_i) = \arg\max_i \frac{P(\lambda_i/X)P(X)}{P(\lambda_i)}.$$

$$1 \le i \le N$$

$P(\lambda_i/X)$ is called the speaker acoustic model,

P(X) is the prior speaker information.GMM recognition system as show in Figure 1.

C. *Algorithm*

- First, the training speech samples after preprocessing and feature extraction, this paper mainly extracted MFCC and its first and second order differential for speaker recognition and emotion recognition.
- In These MFCCs which will be normalized can be used as the speech features for training a claimed speaker model or emotion model via GMM, and the GMM model is trained GMM1, GMM2, ..., GMMN

(Where, N means how many the number of speakers or emotion kinds) as reference model.

- Input the speech to be tested data for speaker recognition and speech emotion recognition.

IV. Simulation and Analysis

We use the CASIA emotional speech database to do the experiment. This database contains 1200 utterances which were collected from 4 speakers (2 male and 2 female), each of the 4 speakers read 50 sentences with the following six different emotions: happy, angry, surprise, neutral, fear and sad. The sampling frequency for all utterances is 16000 Hz. The simulations select a total of 120 samples of four different emotions (happy, angry, neutral, fear) as a test. In the speaker recognition part, because of the need for large amount of voice and data, some other speakers joined.

In speaker recognition part, we use vector quantization VQ method to do a comparison, in which, the feature parameters are MFCC, vector quantization recognition rate is 83.3%, and the proposed method is 88%. In Emotion recognition part, KNN be used as comparison, its average recognition rate is 61%, and the proposed method is 73%. From the experiment results recognition rate of this method, we can conclusion that based on a combination of MFCC and GMM model with speaker identification, are better than the vector quantization method, and its emotion recognition rate is better than the KNN method. Therefore, compared to the general, GMM model in speech recognition method has some advantages.

V. Summary

This paper we put forward a method for speaker recognition and speech emotion recognition based on GMM, the results show that the proposed method is feasible, but there are still too much room for improvement, for instance, the characteristic parameter selection is relatively simple, in the next step we will try to use more characteristic parameter identification.

References

[1] Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models [J]. Digital signal processing, 2000, 10(1): 19-41.

[2] Navratil J, Jin Q, Andrews W, Campbell J. Phonetic speaker recognition using maximum likelihood binary decision tree models.ICASSP2003

[3] Guiwei Ou and Dengfeng Ke,"Text-independent speaker verification based on relation of MFCC components," 2004 International Symposium on Chinese Spoken Language Processing, pp. 57-60, Dec. 2004.

[4] A. Mezghani and D. O 'Shaughnessy, "Speaker verification using a new representation based on a combination of MFCC and formants", 2005 Canadian Conference on Electrical and Computer Engineering, pp. 1461-1464, May 2005.

[5] M.M Homayounpour and I. Rezaian, "Robust Speaker Verification Based on Multi Stage Vector Quantization of MFCC Parameters on Narrow Bandwidth Channels," ICACT 2008,Vol 1, pp. 336-340, Feb.2008

[6] C.C. Lin, S.H. Chen, T. K. Truong, and Yukon Chang, "Audio Classification and Categorization Based on Wavelets and Support Vector Machine," IEEE Trans. on Speech and Audio Processing, Vol.13,No.5,pp.644-651,Sept.2005.

[7] L. Rabiner and B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall,1993

[8] Zhao Li. Speech Signal Processing [M].Beijing: Machinery Industry Press. 2003.

[9] Reynolds, Douglas A. Quatieri, Thomas F. and Dunn, Robert B. Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing10 (2000), 19–41.

[10] Reynolds D A. Gaussian Mixture Models [J].2009.

[11] Brunelli R, Falavigna D. Person identification using multiple cues [J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 1995, 17(10): 955-966.

[12] Hu H, Xu M X, Wu W. GMM supervector based SVM with spectral features for speech emotion recognition [C] Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. IEEE, 2007, 4: IV-413-IV-416.

[13] El Ayadi M M H, Kamel M S, Karray F. Speech emotion recognition using Gaussian mixture vector autoregressive models[C] Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. IEEE, 2007, 4: IV-957-IV-960.

[14] Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using GMMs[C] INTERSPEECH. 2006.

[15] Murty K S R, Yegnanarayana B. Combining evidence from residual phase and MFCC features for speaker recognition [J]. Signal Processing Letters, IEEE, 2006, 13(1): 52-55.

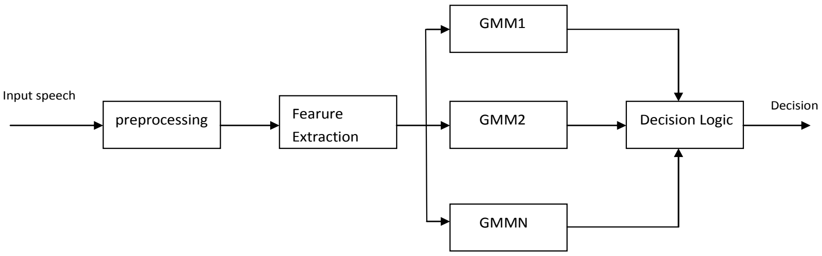

Figure 1. GMM recognition system.