Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 15 (2011) 4031 – 4035

Procedia Engineering

www.elsevier.com/locate/procedia

Advanced in Control Engineering and Information Science

Design and Implementation of an Audio Classification System Based on SVM

Wang Shuiping[a,b,c], Tang Zhenming[a] ,Li Shiqiang[b] , a*

[a]School of Computer Science & Technology,Nanjing University of Science & Technology, Nanjing,210094, China
[b]School of Computer Science & Technology,Nanjing University of Information Science & Technology Nanjing,210044 China
[c]Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, 210044 China

Abstract

Time-domain and frequency-domain features were extracted. The research of process and architecture of an audio classification system based on SVM was done, and the SVM audio classifier was designed. The results of experiments show that the audio classification system designed in the paper can classify audio signal effectively, and the average identification accuracy is about 90%.

*Keywords:*Audio classification ; MFCC; SVM

1. Introduction

Content-based audio classification and recognition technology research began in the late 20th century. It has great application value in distance learning, digital libraries, news search, and other fields.

Lu Jian, Nanjing University, proposed an audio classification method based on Hidden Markov Model [1]. It can be used for voice, music, and their hybrid sound classification, and the best classification accuracy of this algorithm is about 90.28%. Zhao Xueyan et al, Zhejiang University, proposed an audio classification and retrieval system based on unsupervised mechanism [2]. In this method, audio features can be extracted from compressed domain and feature dimension reduction is completed by a time and space constraint fuzzy clustering. The speed of this retrieval method is fast, and the accuracy is increased

* Corresponding author. Tel.: 13913333513; fax: 025-58731323.
E-mail address: shuipingw@126.com.

1877-7058 © 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2011.08.756

greatly. Li, S.Z et al used MFCC(Mel Frequency Cepstral Coefficients) as audio features[3], which reflect the characteristics of human audio perception well, and designed an audio multi-classification system based on SVM. Erlin Wold et al analyzed the audio distinctive features, which include loudness, pitch and harmonicity, and then designed an audio classifier with Nearest Neighbor criterion. The data they used include 16 types, such as laughter, ringtones, phone ring and etc. Chih-Chieh Cheng et al used ellipsoid distance [4] to identify musical instrument sounds, male voices, female voices and environmental sounds. The features detected from these sounds include Short-Time Energy, Zero-Crossing Rate, Centroid and Bandwidth of the voice frequency spectrum. The optimize symmetric matrix was used in audio feature selection experiments.

In this paper, we used Short-Time Average Zero-Crossing Rate, Short-Time Energy, Centroid of audio frequency spectrum, Sub-Band Energy and MFCC as the characteristic parameters and designed an audio classification system based on SVM. The experimental results are satisfactory.

2. Time-domain Feature Extraction

2.1. Short-Time Average Zero-Crossing Rate

Short –Time Average ZCR stands for the times of crossing the zero signal in a unit time. As to the discrete audio signals, it means the sign changes of the audio signal. Short-Time Average ZCR can reflect the nature of the signal spectrum to a certain extent, so it can be used to estimate the signal spectral characteristics roughly.

Short –Time Average ZCR can be calculated as follow:

$$Z_n = \frac{1}{2}\sum_{k=-\infty}^{\infty}\left|\mathrm{sgn}[s(k)]-\mathrm{sgn}[s(k-1)]\right|w(n-k)=\frac{1}{2}\sum_{k=n}^{n+N-1}\left|\mathrm{sgn}[s_w(k)]-\mathrm{sgn}[s_w(k-1)]\right| \quad (1)$$

Where $w(n)$ is the window function, and $s_w(k)$ is the signal $s(k)$ after windowing processing. N stands for the length of window function, and $\mathrm{sgn}[\cdot]$ means the sign function.

2.2. Short-Time Energy

As to an audio signal $\{s(n)\}$, Short-Time Energy can be defined as follow:

$$E_n = \sum_{k=-\infty}^{\infty}[s(k)w(n-k)]^2 = \sum_{k=-\infty}^{\infty}s^2(k)h(n-k) = s^2(n)*h(n) = \sum_{k=n}^{n+N-1}s_w^2(k) \quad (2)$$

Where $h(n)=w^2(n)$. Short-Time Energy can be used to measure the strength of the audio signal, and it can be used for sound/silent determine.

3. Frequency-domain Feature Extraction

3.1. Centroid of Audio Frequency Spectrum

The Centroid of an audio frequency spectrum means the average points of the spectral energy. It reflects the center of audio frequency distribution, it is a measure of the audio signal brightness, and it can be defined as follow:

$$SC = \frac{\int_0^{\bar{w}}\omega\left|F(\omega)\right|^2 d\omega}{E} \quad (3)$$

When the frequency is fixed as ω_k , that means $\omega = \omega_k$, where ω_k is the center frequency, E means the Energy, and $|F(\omega)|^2$ means the power spectrum of the audio signal.

3.2. Sub-Band Energy Ratio

Sub-Band Energy Ratio is used to measure the different Sub-Band Energy Ratio of the total band energy. The Sub-Band Energy of the music signal is distributed uniform, while the energy spectrum of voice signal is mainly in the first sub-band. The energy of every sub-band can be calculated as follow:

$$D = \frac{1}{E}\int_{Lj}^{Hj} \left| F(\omega)^2 \right| d\omega \tag{4}$$

3.3. Mel Frequency Cepstral Coeffiencents

Mel Frequency Cepstral Coeffiencents are the acoustic characteristics derived from human hearing mechanism [5]. Studies have shown that it is approximate linear relationship between people's feeling and the sound frequencies below 1000Hz, and that it is linear relationship not in sound frequencies but in logarithmic frequency coordinates.

4. SVM-based Audio Classification System

4.1. System Design

The system flow chart of audio classification system designed in this paper is shown as Fig.1. The first step is pre-processing. After doing so, we can get audio signal frame data. Several frame-level features such as Short-Time Average Zero-Crossing Rate, Short-Time Energy, Centroid of audio frequency spectrum, and Sub-Band Energy and MFCC. We also calculate some statistical characteristics, such as mean, variance, High Zero-Crossing Rate Ratio and Low Short-Time Energy. After that, we can get complete set of feature vectors.

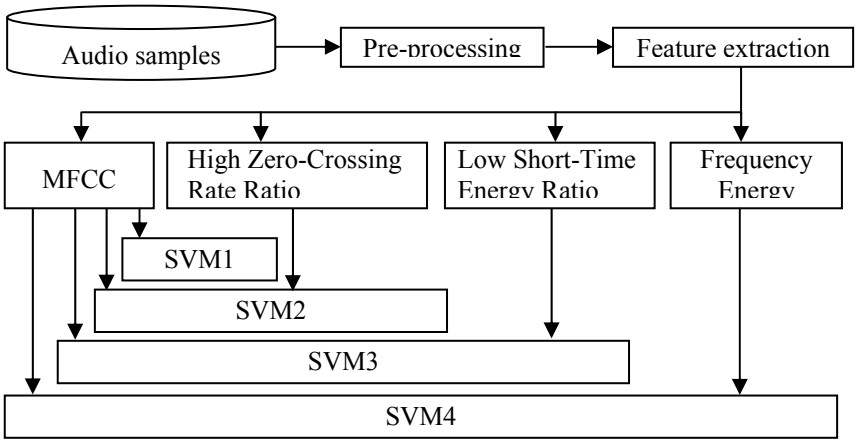

Fig.1 Flow chart of audio classification system

The training samples and test samples are sent to SVM to begin training and testing. The block diagrams of classifier training and classification subsystems are designed as Fig.2 and Fig.3.

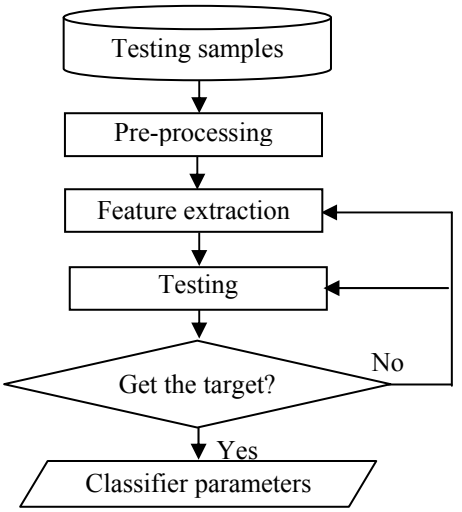

Fig.2 Block diagram of SVM training

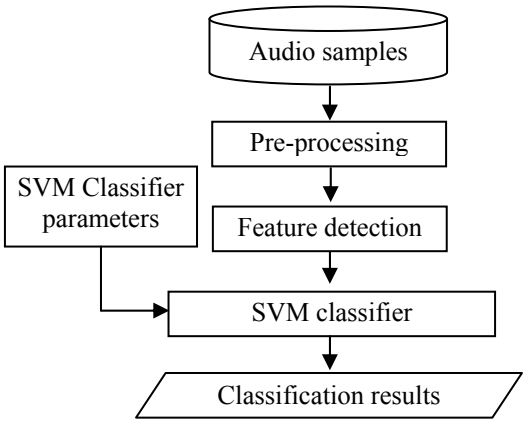

Fig.3 Block diagram of SVM classification

The classification accuracy can be calculated as equation 5, which is shown as follow:

$$\text{Classification Accuracy} = \frac{\text{Number of correct classification audio clips}}{\text{Number of total audio clips of audio sample}} \quad (5)$$

4.2. SVM Classifier Processing

The processing of SVM classifier includes 6 steps, which is shown as follow:

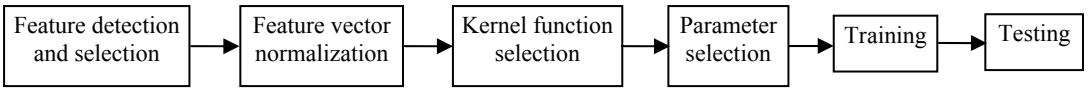

Fig. 4 SVM classifier work processing

In this paper, the mean and variance of MFCC are selected to build the basic feature set. Some clip based features are chosen to add to the basic feature set one by one, and several times of trainings and tests are done. RBF kernel function is selected in kernel function selection module.

5. Experiment and Analysis

In the experiments, the original audio data include 2500 clips. 1200 of them are voice data, and the other 1300 clips are music clips. 800 voice clips and 800 music clips are chosen to form the training set. The rest 300 voice clips and 400 music clips form the test set. The results are shown in Table.1 and Table.2.

Table.1 Result of MFCC

Sample		Result		Accuracy（%）
Category	Clip Number	Voice	Music	
Voice	300	**274**	26	91.33%
Music	400	41	**359**	89.75%
Average Recognition Rate				90.43%

Table.2 Result of MFCC and Other features

Feature	MFCC (SVM 1)	MFCC and HZCRR (SVM 2)	MFCC and LSTER (SVM 3)	MFCC and Frequency Energy (SVM 3)
Accuracy（%）	90.43%	91.29%（+0.86%）	91.86%（+1.43%）	92.14%（+1.71%）

Experimental results show that the average identification accuracy of MFCC is about 90.43%, and 3 other clip-based features can also improve the recognition rate. The accuracy may be improved by 1.71% with Frequency Energy features.

Acknowledgements

This study is supported in part by Jiangsu Provincial Government Scholarship Foundation, and by Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

[1]Lu Jian, Chen Yisong, Sun Zhenfxing. Automatic Audio Classification by Using Hidden Markov Model. *Journal of software;* 2002, p. 1593-1597.

[2] Zhao Xueyan, Wu fei, Liu Junwei. Audio Clip Retrieval and Relevance Feedback based on the Audio Representation of Fuzzy Clustering. *Journal of Zhejiang University*; 2003,p. 264-268.

[3] Li S Z, Guo Guodong. content-based audio classification and retrieval using SVM learning. *Proceedings of the 1st IEEE Pacific-Rim Conference on Multimedia*. Sydney, Australia. 2000, p.1507-1510.

[4] Chih-Chieh Cheng,Chiou-Ting Hsu. Content-Based Audio Classification with Generalized Ellipsoid Distance.*Proc,PCM*. Hsinchu, Taiwan. 2002, p.328-335.

[5] Han Jiqing, Feng Tao, Zheng Guibing, Ma Yiping. Audio Information Processing Technolygy. *Tsinghua University Press;* 2007.

[6] Theodoros Giannakopoulos, Dimitrios Kosmopoulos. Violence Content Classification Using Audio Features. *SETN Springer-Verlag Berlin Heidelberg* 2006, LNAI 3955, p.502-507.

[7] Bai Liang; Hu Yaali; Lao Songyang; Chen Jianyun; Wu Lingda; Feature analysis and extraction for audio automatic classification. *IEEE International Conference on Volume 1*. 2005: 767-772.