A SVM-Based Classification Approach to Musical Audio

Namunu Chinthaka Maddage
Institute for Inforcomm Research
21, Heng Muikeng, Terrace
Singapore 119613
maddage@i2r.a-star.edu.sg

Changsheng Xu
Institute for Inforcomm Research
21, Heng Muikeng, Terrace
Singapore 119613
xucs@i2r.a-star.edu.sg

Ye Wang
School of Computing
National University of Singapore
Singapore 117543
wangye@comp.nus.edu.sg

Abstract

This paper describes an automatic hierarchical music classification approach based on support vector machines (SVM). Based on the proposed method, the music is classified into coursed classes such as vocal, instrumental or vocal mixed with instrumental music. These main classes are further sub-classed according to gender and instrument type. A novel method, Correction Algorithm for Music Sequence (CAMS) has been developed to improve the classification efficiency.

1 Introduction

With the growing need for multimedia applications, audio analysis has become an important issue in the signal processing area. Content-based audio retrieval depends on classification of intrinsic properties of the audio. Automatic music transcription is another important application, which depends upon a method of audio analysis and is related to post processing and editing phases of actual recordings (Eronen, et.al., 2000).

The goals of this research are: (1) to explain whether there exist significant statistical differences between vocal melody structure, music instruments (string type-acoustic guitar, blowing type-harmonica) and mixture of vocal and instrumental music without taking time dependent characteristics into consideration; (2) to study how support vector machines (SVM) performs for music classification as a time series analysis problem; and (3) to compare the classification performance with multilayer neural networks (MNN)and Gaussian mixture model (GMM).

2 Musical Audio Features

We consider features that are often used in audio / speech analysis including linear prediction coefficients (LPC), LPC derived cepstrums (LPCC), mel-frequency cepstral Coefficients (MFCC), spectral power (SP), short time energy (STE), and zero crossing rates (ZC) (Rabiner, et.al, 1993).

The different features have different strengths distinguishing one class from other class of music. MFCCs are more effective in identifying different vocal structures as well as instrumental music. The SP and ZC features perform better in identifying vocal related music and blowing type of instrumental music. The LPC and LPCC are highly correlated with each other and performance wise LPCCs are much better in identifying vocal music. The selective frequency band LPCC can improve the performance over full band LPPCs (Maddage, et.al., 2002).

3 Experimental Setup

We have recorded 10 Sri Lankan songs (2~3 minutes long), sung in middle scale with major chords composition, by both male and female singers at different time periods with a stereo 16-bit wave format and a 44.1 KHz sampling frequency. In order to generate testing and training samples, we mixed vocal tracks (female and male) with instrumental tracks (acoustic guitar and harmonica) without distorting the melody characteristics of the songs, as shown in Figure 1 (the positions of time T1, T2 and T3 are changed in generating audio samples).

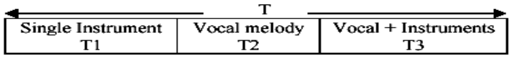

Figure 1: Song Composition

In Figure 2**,** the classification steps of musical audio are shown. Here we use six SVM classifiers (SVM 1~6) and all the classifiers are trained for 2-class classification. The training and testing data sets are shown in table 1**.** Initially the musical audio is segmented into 20ms frames with variable ∇ percentage overlapping. Else where (Xu, et.al., 2002) we have shown the ∇ =70% for training and ∇ = 20% for testing work well compared with other values of ∇ (i.e. 0 < ∇ <100).

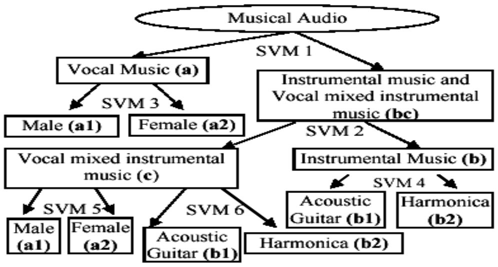

Figure 2: Classification Steps

In order to find effective orders of LPCCs, MFCCs and SPs, we vary the order of the feature and note down the classification accuracies respective. Then select the order according to best classification accuracy. Both effective orders and the classification accuracies of LPCCs are noted in row 1 and row 2 of Table 1, respectively. The joint feature efficiencies are noted in last four rows. Since this classification task (Figure 2) is non-linear (Xu, et.al., 2002), we train radial basis kernel function (Vapnik, 1998) with different variable setting which varies with vector dimension and the common length scale constant (CLSC). The values for CLSCs are selected via cross validation (Table 1).

Main Classes	Detailed Individual Classes	Training Set(minutes)	Evaluation Set (minutes)	Kernel -RBF	
				Classifier	CLSC
Vocal music (a)	Male Vocals (a1)	5.12	7.54	SVM 1	13
	Female Vocals (a2)	6.36	8.23	SVM 2	45
Instrumental music (b)	Acoustic Guitar (b1)	4.58	6.11	SVM 3	24
	Harmonicas (b2)	5.17	6.89	SVM 4	18
Vocal mixed instrumental music (c)	Male or Female Vocals (a1/a2)	15.78	17.57	SVM 5	21
	Acoustic Guitar or Harmonica (b1/b2)	13.19	15.24	SVM 6	18

Table1: Training and Evaluation Date Set

Classes / Features	a-bc	b-c	a1-a2	b1-b2	c	
					a1-a2	b1-b2
LPCC	22	15	19	20	21	25
	90.57	84.55	87.52	88.34	88.69	87.43
MFCC	12	23	24	22	25	12
	89.80	86.21	88.77	87.97	86.75	34.17
SP	17	25	12	12	12	13
	76.12	79.81	51.75	79.94	80.52	85.40
STE	18.34	61.09	52.35	64.07	80.13	34.17
ZC	26.78	69.90	85.22	85.62	73.76	66.64
MFCC+LPC	92.86	88.34	90.22	90.57	90.04	
LPC+SP					89.03	90.34
MFCC+SP					88.38	
MFCC+ZC			89.29	88.24		

Table 2: Feature Analysis

3.1 Correction Algorithm for Music Sequence (CAMS)

Since the temporal features of the music signals are not taken into consideration while making feature vectors, it is noticed that SVM misclassifies the class boundaries of different music, which is more pronounced in separating class b and c. We have developed an algorithm that exploits the rhythm of the musical score in order to do better classification.

The prominent periodicities of the melody of many types of music may be extracted by using rhythm of the music. The main beat frequency (1/ rhythm period) of the musical score is calculated using beat histogram described in (Tzanetakis, et.al., 2001) and the assumption we made is that minimum duration of a class of music is more than the 1/2 of the rhythm period. The key points of the CAMS are summarized below.

- *$(n_x, \underline{n}_x)$, n_y, and n_z are number of feature vectors (i.e.- feature frames) in classes C_x, C_y, and C_z defined by the SVM classifier .*
- *$\underline{C}_x$, $\underline{C}_y$, and $\underline{C}_z$ are the mean vectors in the classes; C_x, C_y, and C_z*
- *f()is the frame index*
- *N is the total number of frames in the musical score.*
- *N_{th} is pre-define integer and it depends on the beat /rhythm of the musical scale N_{th} = (1/beat frequency)*0.5*

Since number of frames (n_y) in class C_y is less than N_{th} , we merge those frames with either class C_x or class C_z according to two cases describe below.
$n_y<N_{th}<< (n_x, \underline{n}_x, n_z) \leq N \quad n_x\, \underline{n}_x\, n_y\, n_z \neq 0$

```
Case 1: $[\{f(i+j+1)\sim f(i+j+n_x)\}\&\{f(i-\underline{n}_x)\sim f(i)\}]\in C_x \quad f(i+j)\in C_y \quad j= 1\ldots n_y \leq N_{th}$
        Then $f(i+j)\in C_x$

Case 2: $\{f(i-n_x)\sim f(i)\}\in C_x$, $\{f(i+j+1)\sim f(i+j+n_z)\}\in C_z$, $f(i+j)\in C_y$, $j= 1\ldots n_y \leq N_{th}$
        If $\{eudist(\underline{C}_y-\underline{C}_x)\geq eudist(\underline{C}_y-\underline{C}_z)\}$ ; Euclidean distance between
           Then $f(i+j)\in C_z$                                mean vectors in class $C_x\ C_y\ C_z$
        Else
           $f(i+j)\in C_x$
```

3.2 Comparison

To further illustrate the advantage of the proposed approach, we compare the performance of the SVM method with other methods including MNN (Haykin, 1998) and GMM (Bilmes. 1998). For MNN, we use 6 hidden layers with 32 nodes in each layer. The classification results in Table 3 prove that hierarchical classification (Figure 2), is ideal for multi class classification problem and CAMS improves post classification efficiency of SVM, MNN and GMM) by (2~4) %. The SVM performs better in all the classifications than MNN and GMM. Both gender (a1-a2) classification and instruments (b1-b2) classification in vocal mixed instrumental music (c) are difficult tasks compared with other classes. This is because of the complexity of vocal structure and it is more pronounced when female vocals are mixed with instrumental music (female vocals have higher order harmonics than male vocals).

Classes / Classifiers		a-bc	b-c	a1-a2	b1-b2	C	
						a1-a2	b1-b2
Classifiers without CAMS	SVM	92.86	88.34	90.22	90.57	90.34	90.34
	MNN	87.22	84.19	85.78	82.35	79.56	82.87
	GMM	88.56	82.26	80.45	87.21	81.24	83.11
Classifiers with CAMS	SVM	95.78	91.21	94.10	93.59	94.22	94.72
	MNN	91.45	87.81	89.25	86.76	81.77	86.58
	GMM	91.98	88.56	83.18	90.13	83.46	88.95

Table 3: Comparison Results

4 Conclusion and Future Work

Although the test data sets we used in our experiments are not sufficient to generalize the very high performance of both the features and the SVM classifier, it can be seen that musical score is statistically separable with good performance (over 85 %); specially main 3 classes (i.e. a, b & c). The classification complexity can be reduced by hierarchical classification steps. By introducing CAMS we could be able to increase the overall performance by (3 ~4) %. One of the drawbacks of this system is high computational complexity in calculating different feature orders for different classification steps.

Several problems need to be tackled in the future. The reduction of feature dimension with good overall performance and developing uncorrelated features are challenging tasks. The other direction is to improve the CAMS. By taking mutual information between frames in to consideration, we can improve stability of the CAMS.

References

Bilmes, J. A. (April 1998). A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. International Computer Science Institute Berkeley CA, 94704.

Eronen, A. & Klapuri A. (2000). Musical Instrument Recognition Using Cepstral Coefficients and Temporal Features, *Proc. ICASSP.*

Haykin, S. (1999). *Neural Networks*: 2nd edition, Prentice Hall.

Maddage, N.C., Xu, C.S, LEE, C.H., Kankanhalli, M. S. & Tian, Q. (2002). Statistical Analysis of Musical Instruments. *3rd IEEE PCM, Taiwan* (pp 581-588).

Rabiner, L. R. & Juang, B. H. (1993). *Fundamentals of Speech Recognition,* Prentice-Hall.

Tzanetakis, G., Essl, G. & Cook, P. (2001). Automatic Music Genre Classification of Audio Signals. *ISMIR.*

Vapnik, V. (1998). *Statistical Learning Theory*: Wiley.

Xu, C.S, Maddage, N.C. & Tian, Q. (2002). Support vector Machine Learning for Music Discrimination. *3rd IEEE PCM, Taiwan* (pp 928-935).