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 coarse 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 LPCCs (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).

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 \( \nabla = 70\% \) for training and \( \nabla = 20\% \) for testing work well compared with other values of \( \nabla \) (i.e. \( 0 < \nabla < 100\)).

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 row in 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).

### 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).

### Table 3: Comparison Results

<table>
<thead>
<tr>
<th>Class</th>
<th>SVM</th>
<th>MNN</th>
<th>GMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male Vocals</td>
<td>95.78</td>
<td>94.10</td>
<td>93.59</td>
</tr>
<tr>
<td>Female Vocals</td>
<td>90.34</td>
<td>88.34</td>
<td>86.58</td>
</tr>
<tr>
<td>Harmonica</td>
<td>91.75</td>
<td>90.57</td>
<td>89.03</td>
</tr>
<tr>
<td>Vocal mixed</td>
<td>92.82</td>
<td>91.22</td>
<td>89.56</td>
</tr>
</tbody>
</table>

### References


