VOCAL-BASED MUSICAL GENRE CLASSIFICATION

Bryan Huh **Arun Miduthuri**

Abstract

Musical genres are labels assigned to pieces of music. Features of a genre generally include lyrical structure, rhythmic structure, instrumentation, and harmonic content. In recent years, efforts have been made to classify genres of music using machine learning techniques on the above mentioned aspects of a song. Our project attempts to find what audio features are more important when classifying vocal tracks into different genres. Two vocal datasets are examined: one containing Indian vocal genres with no accompaniment, and another with accompaniment filtered out with the help of a recent vocal separation technique.

Introduction

As the digital database for music grows, so does the demand for its organization. Currently, much of music classification is still done by individual users who store the artist name, song title, genre, and album in the metadata of the music file. However, the vast quantity of music files on the Web is making the manual classification of music libraries more and more infeasible. The automatic classification of music has thus become an important problem, and is one of the goals of Music Information Retrieval (MIR). Automatic musical genre classification is particularly sought after, but it is also a great challenge. The boundaries of a genre are generally not well defined, and even humans will often disagree on the genre of a song. Indeed, human performance on genre classification has shown that automatic genre classification is fundamentally limited by the subjectivity of genre [0].

The vast majority of research on musical genre classification has targeted feature-extraction. Previous studies using "timbral features" such as Mel-Frequency Cepstral Coefficients (MFCCs) and rhythmic features such as beat features, chroma features, and other pitch-related features have shown that by far the most useful feature in genre classification is the MFCCs. Surprisingly, those features which humans seem to rely on in genre classification such as rhythm, harmony, and vocal content have not yet made an impact in automatic genre classification. In particular, the role of vocals separately in genre classification has never been formally addressed before. Since vocals play a large role for humans in genre classification, it is an important question how much they are being utilized in automatic genre classification. Perhaps features which have had little contribution in the past, when extracted from vocals, can make a greater impact in musical genre classification. In this

paper, we determine the relative importance of various features in the genre classification of a standard Genre dataset used in [1] and a primarily-vocal dataset.

Related Work

Musical genre classification has been explored in detail in the seminal work by Tzanetakis et al [1],[2]. Related problems include singer identification [3], finding similar music using unsupervised methods[4], and locating singing voice segments within musical pieces[5]. Related to our task of vocal genre classification is the broader idea of separating vocals from accompaniment in monoaural recordings. This has been done in several different ways, for example, independent component analysis [6], mixed Gaussian models [7], and peak clustering [8]. In our work we make use of three primary features: Mel-frequency cepstral coefficients, chroma features, and linear spectral pair features (LSPs). MFCCs and LSPs are widely used for speech discrimination, but have proven useful in music classifiers, for example, see [5]. There are additional spectral centroid, rolloff, and flux features added on the basis of features used in [1]. In addition, linear spectral pair coefficients have been added as they are widely used in speech coding. Chroma features are commonly used to capture pitch content and harmony. Since LSPs are primarily used for speech modeling, it would not be surprising if LSPs contribute most to genre classification when they are extracted from pure vocals.

Design

Datasets

Three datasets (two vocal datasets) were used for genre classification. The first is a standard genre dataset used in [1], which we will call the original/general Genre dataset. We selected those genres which had vocal content: Country, Disco, HipHop, Rock, Blues, Reggae, Pop, Metal. The second dataset was generated from this Genre dataset with the aim of making the vocal component of the songs more prominent. A peak clustering algorithm (Marsyas) was used to isolate the vocals of the Genre dataset. We call this the Vocal-separated dataset. Finally, we used a dataset of traditional Indian music since Indian music has a large databank of purely vocal songs. This was divided into eight genres: Female Bollywood, Female Carnatic, Female Hindustani, Female Mantra, Male Carnatic, Male Hindustani, Male Qawwali, and Male Rajasthani. Male and female vocals were separated for better training.

Features

The genre classification was done using combinations of the MFCC features, spectral features (spectral centroid, spectral rolloff, spectral flux, time domain zero crossings) [1], chroma features and LSPs. We adopt the name “Timbral features” [1] for the collection of MFCCs and spectral features combined. It is standard to include Timbral features as a baseline. We then classified using Timbral features combined with either chroma features or LSPs. Feature extraction was done using MARSYAS, an open-source software used for audio analysis [9]. For each audio file, a

single feature vector was computed.

Training and Classification

Previous results [2] have shown that a Gaussian SVM classifier is a successful classifier in genre classification. Thus for each of our data sets a Gaussian SVM classifier was used, and our results were obtained using 10-fold cross-validation.

Results and Discussion

The figures on the next page (figure 2, figure 3) summarize the results for the three datasets and the various combinations of features. For the original Genre dataset, adding chroma features significantly improves the classification accuracy. However, linear spectral pair features made no contribution. For the Indian music dataset, Timbral and Timbral/Chroma features alone gave poor classification accuracy, but here linear spectral pair features make a significant contribution. The same is true for the Vocal-Separated set.

The significant contribution of the linear spectral pair features in the genre classification of the Indian music dataset is not surprising considering that the Indian music consisted entirely of vocals, and LSPs have traditionally been used for speech coding. Examination of the confusion matrices (Figure 1) in classification runs where the linear spectral pairs were included and left out shows that without them, a number of the male vocal genres were misclassified as belonging to Male Qawwali. Interestingly, one of the female genres, Female Carnatic, which had melodies in the same pitch range as some male voices, was also misclassified as Male Qawwali in every test example. The genres are intrinsically somewhat similar to one another in terms of harmonic quality, and are generally mainly different in tone of voice. Evidently linear spectral pairs capture this tone difference. However, their failure to improve the classification accuracy for the original Genre dataset suggests that their contribution can only be made when vocals are at least moderately isolated. This can be seen in the case of separated vocals, where LSPs cause a dramatic improvement in classification accuracy. In contrast, the chroma features, which capture pitch and harmony, are assisted by background instrumentation as they only make contributions in the original Genre dataset and marginally in the vocal-separated dataset, in which there continued to be some residual background accompaniment.

Figure 1(a): Confusion Matrix: w/LSP features

a	b	c	d	e	f	g	h	<-- classified as
14	0	0	0	0	0	0	0	a = Female_Bollywood
1	0	0	1	0	0	6	0	b = Female_Carnatic
6	0	0	1	0	0	0	0	c = Female_Hindustani
0	0	0	12	0	0	0	0	d = Female_Mantra
0	0	0	0	7	0	5	0	e = Male_Carnatic
0	0	0	0	0	0	7	0	f = Male_Hindustani
0	0	0	0	0	0	15	0	g = Male_Qawwali
0	0	0	0	0	0	4	7	h = Male_Rajasthani

Figure 1(b): Confusion Matrix: w/o LSP features

a	b	c	d	e	f	g	h	<-- classified as
13	0	0	0	0	0	1	0	a = Female_Bollywood
1	0	0	0	0	0	7	0	b = Female_Carnatic
6	0	0	0	0	0	1	0	c = Female_Hindustani
0	0	0	1	0	0	11	0	d = Female_Mantra
0	0	0	0	0	0	12	0	e = Male_Carnatic
0	0	0	0	0	0	7	0	f = Male_Hindustani
0	0	0	0	0	0	15	0	g = Male_Qawwali
0	0	0	0	0	0	11	0	h = Male_Rajasthani

FIGURE 2

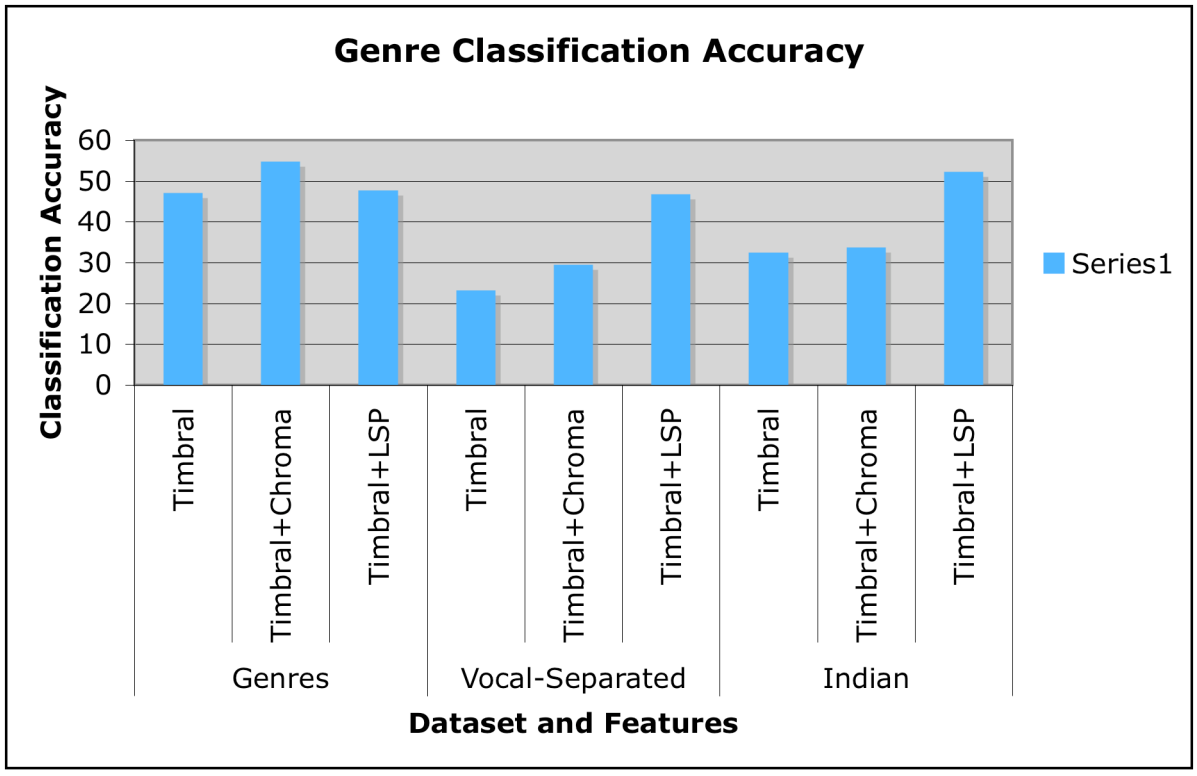

FIGURE 3
CLASSIFICATION RESULTS AND FEATURE COMPARISON

General genre classification (Dataset 1)	
Features	*Accuracy*
Timbral	47.1%
Timbral + Chroma	54.8%
Timbral + LSP	47.7%
Indian vocal music (Dataset 2)	
Features	*Accuracy*
Timbral	32.6%
Timbral + Chroma	33.7%
Timbral + LSP	52.3%
Separated vocals (Dataset 3)	
Features	*Accuracy*
Timbral	23.2%
Timbral + Chroma	29.5%
Timbral + Chroma + LSP	46.9%

Conclusion and Future Work

We found that linear spectral pair features are most useful in genre classification when they can be extracted from vocals without the interference of background instrumentation. The opposite seems to be true with Chroma features. This is consistent with intuition, and it suggests that first isolating the components of the music file, and then extracting features from the isolated components (in particular the vocals) may be an important preceding procedure for improved musical genre classification.

The most immediate avenue for future work would include improving upon general genre classification by including linear spectral pairs from the separated vocals. It would also be interesting to see how the relative importance of features changes with gradual attenuation of the background accompaniment. In particular we would like to see if any other features commonly used in genre classification behave differently when the vocals are more prominent (for example, LPCCs are also a common voice feature like LSPs). We could also test other methods for separating vocals, such as by means of ICA or mixed Gaussian models. Finally, we would like to expand our vocal dataset to include Western genres.

References

[0] S. Lippens, J. P. Martens, M. Leman, B. Baets, H. Meyer, and G. Tzanetakis. "A Comparison of Human and Automatic Musical Genre Classification," in *Proceedings of the IEEE International Conference on Audio, Speech and Signal Processing,* 2004.

[1] G. Tzanetakis, P. Cook, "Musical Genre Classification of Audio Signals," *IEEE Transactions on Speech and Audio Processing*, Vol 10, No 5, July 2002.

[2] T. Li, G. Tzanetakis, "Factors in Automatic Musical Genre Classification of Audio Signals," *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[3] Y.E. Kim, Brian Whitman, "Singer Identification in Popular Music Recordings Using Voice Coding Features," *Proceedings of the 3rd International Conference on Music Information Retrieval, 2002*.

[4] X. Shao, C. Xu, M.S. Kankanhalli, "Unsupervised Classification of Music Genre Using Hidden Markov Model," *2004 IEEE International Conference on Media and Expo (ICME)*.

[5] A Berenzweig, D. Ellis, "Locating Singing Voice Segments within Musical Signals," *IEEE Workshop on the Application of Signal Processing to Audio and Acoustics, 2002*.

[6] S. Sofianos, A. Ariyaeeinia, R. Polfreman, "Towards Effective Singing Voice Extraction from Stereophonic Recordings," *IEEE Int Conf on Acoustics, Speech and Signal Processing 2008*.

[7] Tsai et al., "Blind clustering of popular music recordings based on singer voice characteristics," *4th International Conference on Music Information Retrieval, 2003*.

[8] Lagrange et al, "Normalized cuts for predominant melodic source separation," *IEEE Transactions on Audio, Speech, and Language Processing, 2008*.

[9] Marsyas toolkit for audio classification tasks, available http://marsyas.sourceforge.net