Musical Genre Classification using Melody Features and Source Separated Audio from Polyphonic Signals

Dhaivat Shah, Columbia University in the City of New York

Abstract—High level melodic features extracted from the polyphonic music signals have been used to perform genre classification. A comparison of its performance with the same set of features from source separated audio of the polyphonic music signals is provided. A source/filter model is utilized for unsupervised main melody extraction. Classification and clustering uses standard machine learning libraries from scikit learn. Mel Frequency Cepstrum Coefficients (MFCC) provide a baseline model for evaluation. A combination of melody features and MFCC is also considered. The strength and weakness of melody features is analysed using different subsets of GTZAN dataset.

Keywords—Melody Extraction, Audio signal separation, Genre classification, pitch estimation, source/filter model

1 Introduction

The task of genre classification entails assigning labels based on the characteristics of music. The applications for it can be varied, for casual users, it can be for listening to similar kinds of music; academic/professional users may want to find related sequences to analyze and study, while the music industry needs it for efficient retrieval and storage. But as easy it may seem, it is no trivial task. While this comes naturally to people, it is quite difficult to automate it, due to polyphony. And due the volume of audio files generated, it is simply impossible to label everything by hand. The concept of genre is very subjective, a particular song could be a combination of two or more different styles; or it may be simple difficult to classify a particular song to a genre. This increases the difficulty in classification. Also, the use of various instruments along with the main melody leads to a polyphonic signal which is hard to analyse and represent by sufficiently descriptive features. The paper proposes a model based on melody features from source separated audio. While the approach does not succeed in the strictly general sense, it show promising results for a restricted conducive scenario, and signs of improvement for other features when used in combination.

2 Problem

In order to understand the problem itself and the subsequent formulation, it is necessary to understand a few concepts:

- Musical Genre: The term used to describe music that has similar properties
- Melody: The single (monophonic) pitch sequence that a listener might reproduce if asked to whistle or hum a piece of polyphonic music.
- Pitch: Perceptual property that allows the ordering of sounds on a frequency-related scale.
- Polyphony: Texture consisting of two or more simultaneous lines of independent melody, as opposed to music with just one voice (monophony) or music with one dominant melodic voice accompanied by chords (homophony).
- Source Separation: Estimation of main melody from a polyphonic signal.

Thus, the task at hand is to classify a given song into previously known genres, or to clus-

- *Dhaivat Shah is a Graduate Student in the Department of Computer Science at FFSEAS.*
E-mail: ds3267@columbia.edu

Manuscript received April 10, 2014.

ter a group of songs into possibly similar genres. The following pipeline is used:

- Perform melody extraction on polyphonic audio signal to obtain a source separated audio
- Extract melody features from the source separated audio
- Use standard classification and clustering algorithms on the feature set obtained.

Each of these steps is explained in great detail in the following sections.

3 State of the Art

A detailed survey of the techniques used in genre classification is provided in [1]. We provide a quick glimpse of the different approaches used in this section. For audio signals, features may be related to melody, rhythm, or timbre.

Main low-level features that can be viewed under timbre [19] are:

- Temporal: computed from audio signal source (zero-crossing rate, prediction coefficients, etc.)
- Energy: refers to energy content of the signal (Root Mean Square energy, energy of harmonic component of the power spectrum, etc.)
- Spectral Shape: centroid, shape, skewness, kurtosis, etc.
- Perceptual: loudness, sharpness, spread, etc.

Timbre features are the most widely used for genre classification, however, they are more suited to monophonic than polyphonic music.

Because of the failure of a single feature set being able to represent all the different genres, there is no clear state of the art technique. However, for the dataset that we are using 91% accuracy was obtained by [20] using sparse representations of auditory temporal modulations.

Also, [4] which uses melody features was able to obtain about 50-60% accuracy on the same dataset and about 90-95% accuracy on a synthesized dataset.

4 Approach

Each of the following steps are performed in the order mentioned to obtain the final result.

4.1 Source Separation

This is based on the source/filter model proposed in [2] and further explored in [3]. While the supporting code has been borrowed from the original authors, the main ideas underlying the model have been briefed here.

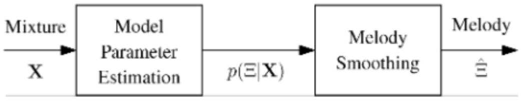

Fig. 1. System Outline

As shown in Fig. 1, the overall system is two step melody tracker which relies on parametrization of the power spectrogram. In the first step, parameters of a short time Fourier transform (STFT) of the mixture signal are estimated and the posterior probabilities are computed. The second step outputs the desired sequence after a melody smoothing block based on the Viterbi algorithm [22].

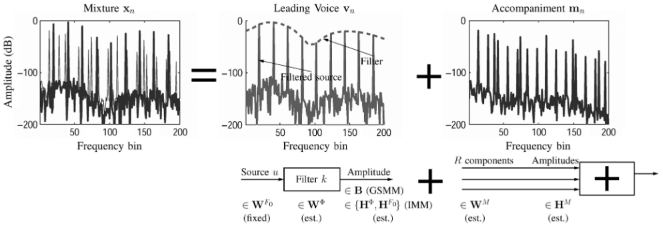

Fig. 2. Decomposition of frame into leading voice and accompaniment

Fig. 2 shows the general principle of the parametrization of the signal. A source filter model is fitted to the main melody part, to capture the variability of the lead voice in terms of pitch range and timbre; and the residual accompaniment, which consists of more stable pitch lines, is modelled in a non-negative matrix factorization (NMF) framework.

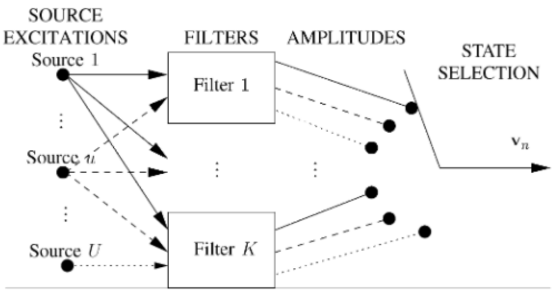

Fig. 3. Source/filter model

The source/filter model is depicted in 3. Each source excitation is filtered through each of the filters and the amplitudes for a frame are applied to each of the output signals. The final state selector sets the active state for a given frame based on its likelihood given the source, filter pairs; and each of the is scaled by the amplitude. This likelihood estimation is similar to a Gaussian Mixture Model (GMM) and Estimation Maximization (EM) is deployed to solve for the parameters.

In practice, the number of filters and number of accompaniment components have been derived from testing different combinations, and remains fixed during evaluation.

4.2 Feature Extraction

All the melody features have been derived from the fundamental frequencies (F0). For each of the 30 second audio track, F0 has been estimated at 5169 instances. The following have been used:

- **Pitch Based**
  - Mean F0
  - Variance
  - Max F0
  - Min F0
  - Skewness
  - Kurtosis
- **Vibration based**
  - Mean
  - Variance

Here, vibration has been calculated as the difference between consecutive recorded F0s. Skewness is the measure of asymmetry data around the sample, and can be calculated as:

$$s = \frac{E(x-\mu)^3}{\sigma^3} \tag{1}$$

Kurtosis is a measure of how outlier prone a distribution is, and can be calculated as:

$$k = \frac{E(x-\mu)^4}{\sigma^4} \tag{2}$$

All the pitch based features were also considered incrementally for vibration based ones, however they hampered the accuracy, so they were dropped. Features related to the topology were also tested to the same result.

Features based on Mel Frequency Cepstrum Coefficients (MFCC) were used for a baseline comparison. Its mean and variance provided a total of 26 features, considering 13 coefficients. A combined feature set of the melody features and the MFCC ones was also used.

Melody features were calculated for both the original audio as well as the source extracted audio separately.

4.3 Classification

Four different methods were used for classification:

- K-Nearest Neighbors (KNN) Classifier
- Gaussian Naive Bayes (GaussianNB)
- Support Vector Machine with linear kernel (linearSVC)
- Support Vector machine with rbf kernel (SVC)

4.4 Clustering

K-Nearest Neighbors (KNN) algorithm was used for clustering. The accuracy score was calculated using the following measures:

- Homogeneity: each cluster contains only members of a single class [21]
- Completeness: all members of a given class are assigned to the same cluster.
- V Measure: Harmonic mean of Homogeneity and Completeness.
- Adjusted Rand Score: measures the similarity of the two assignments, ignoring permutations and with chance normalization.
- Adjusted Mutual Info: measures the agreement of the two assignments, ignoring permutations.

5 Implementation

The following tools were used for carrying out the above mentioned tasks:

- **separateLeadStereo** from the pyFASST package [5] was used for source separation. pyFASST is a python implementation of the flexible Audio Source Separation Toolbox. The dataset used provided mono 16-bit audio files, so the code which was primarily for stereo had to be adapted a bit to suit the dataset.
- **Scikit Learn** has been used for implementing all the classification and clustering tasks. The classifiers have been used with their default configurations.
- **MFCC** implementation from Auditory Toolbox [18] is used for calculating the MFCC features.
- All the intermediate scripts for processing have been written in python and matlab. Python libraries scipy and numpy were used for all the calculations.

6 Dataset

A subset of the GTZAN dataset [6] is used for all evaluation criteria. The dataset consists of 1000 audio tracks each 30 seconds long. It contains 10 genres, each represented by 100 tracks. The tracks are all 22050Hz Mono 16-bit audio files in .au format. Of these the first 10 audio tracks from each of the 10 genres have been analyzed. The audio files had to be converted from .au format to the .wav format for further processing. A linux utility known as soxexam was used for the conversion.

The dataset contains tracks under the following categories (the description is from their respective Wikipedia articles):

- **Blues**: Cyclic musical form in which a repeating progression of chords mirrors the call and response scheme commonly found in African and African-American music [7]
- **Classical**: Classical music has been noted for its development of highly sophisticated forms of instrumental music [8]
- **Country**: Country music often consists of ballads and dance tunes with generally simple forms and harmonies accompanied by mostly string instruments [9]
- **Disco**: Includes a large pop band, with several chordal instruments [10]
- **Hiphop**: Consists of a stylized rhythmic music that commonly accompanies rapping [11]
- **Jazz**: Includes qualities such as swing, improvising, group interaction, developing an individual voice, and being open to different musical possibilities [12]
- **Metal**: A thick, massive sound, characterized by highly amplified distortion, extended guitar solos, emphatic beats, and overall loudness. [13]
- **Pop**: Includes common employment of repeated choruses, melodic tunes, and catchy hooks. [14]
- **Reggae**: It incorporates some of the musical elements of rhythm and blues [15]
- **Rock**: It has centered on the electric guitar, usually as part of a rock group with electric bass guitar and drums.[16]

While the minute details of each of the music form is not of particular interest for this project, the point to take away is that there is a lot of overlap in between these music forms, which makes it difficult to pin down some of the audio tracks to a specific genre. Indeed the boundaries for the different genres is also very fine. In fact, an audio track could very well in reality belong to two or more genres. Though the dataset has been used by many researchers for genre classification, there are still errors in classifying the audio tracks, which has been looked into by [17].

7 Results

The first set of observations has been done on the complete set of genres; whereas for the next set, we narrow down to classical, pop and rock genres, as their main melody is well suited for the feature set we are trying to capture.

Classification and clustering tasks (mentioned in section 4.3 and 4.4) have been performed on features extracted as explained in section 4.2. For classification, a 50:50 split for training and testing was used. Clustering was performed using all the audio tracks.

Classification	KNN	Gaussian Naive Bayes	Linear SVC	SVC
MFCC	**0.32**	0.52	**0.6**	**0.44**
Pitch Solo	0.28	0.36	0.22	0.22
Pitch Orig	0.22	0.3	0.28	0.2
Combined Solo	0.28	**0.54**	0.16	0.26
Combined Orig	0.22	0.3	0.1	0.2

TABLE 1
Classification Accuracy considering all genres

Clustering	Homogeneity	Completeness	V Measure	Adjusted Rand Score	Adjusted Mutual Info
MFCC	**0.411**	**0.437**	**0.424**	**0.173**	**0.262**
Pitch Solo	0.379	0.418	0.397	0.160	0.234
Pitch Orig	0.343	0.412	0.374	0.122	0.203
Combined Solo	0.372	0.416	0.393	0.134	0.228
Combined Orig	0.350	0.401	0.374	0.128	0.207

TABLE 2
Clustering Score considering all genres

Table 1 and 2 show the accuracies for classification and clustering respectively while considering all the genres. We see that our set of features perform very poorly in general for most cases, even when compared to the MFCC baseline. However, we do see some trends, the pitch features extracted from the source separated audio are performing better than those extracted from the original audio.

Classification	KNN	Gaussian Naive Bayes	Linear SVC	SVC
MFCC	0.46	0.73	0.93	0.6
Pitch Solo	0.46	1.00	0.93	0.8
Pitch Orig	0.46	0.93	1.00	0.6
Combined Solo	0.46	0.93	0.93	0.8
Combined Orig	0.46	0.86	1.00	0.6

TABLE 3
Classification Accuracy for specific genres

Clustering	Homogeneity	Completeness	V Measure	Adjusted Rand Score	Adjusted Mutual Info
MFCC	0.389	0.441	0.413	0.340	0.342
Pitch Solo	1.00	1.00	1.00	1.00	1.00
Pitch Orig	1.00	1.00	1.00	1.00	1.00
Combined Solo	1.00	1.00	1.00	1.00	1.00
Combined Orig	1.00	1.00	1.00	1.00	1.00

TABLE 4
Clustering Score for specific genres

Tables 3 and 4 cut down to only the classical, pop and rock genres. This is when we can actually see the power of our feature set. In both the cases, the pitch features are performing way better than MFCC, achieving perfect scores for clustering, and near perfect classification too. The MFCC feature set when combined with pitch features also provides promising results.

In some more experiments which have not been published here, the accuracy did not have a gradual drop as the number of genres were increased. However, it depended on which particular genre was taken into consideration. Certain genres lead to a rapid deterioration while some did not lead to much of a decline.

8 Conclusion

The above results show that the feature set is highly dependent on genres taken into account. Indeed, previous work on using solely pitch features [4] used a synthetic dataset which was suited to the melody features. Thus, though it is highly unlikely that the features be used in isolation, augmenting them to already used features can result in improvements as shown with the MFCC example. Also, using the features extracted from source separated audio provides a more stable prediction as compared to those from the original audio itself.

9 Future Work

Combining the melody related features with the existing state of the art and analysing the performance is something which should shed more of light on the dependability of these features. More complex features could be derived from the fundamental frequencies or from the pitch contours in general, which was not explored in detail in this paper.

References

[1] N. Scaringella, G. Zoia, and D. Mlynek, ”Automatic genre classification of music content: A survey,” IEEE Signal Process. Mag., vol. 23, no. 2, Mar. 2006

[2] J.L. Durrieu, G. Richard, B. Davis, C. Fevotte, ”Source/Filter Model for Unsupervised Main Melody Extraction From Polyphonic Audio Signals,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 3, Mar. 2010

[3] J.L. Durrieu, G. Richard, B. Davis, ”A Musically Motivated Mid-Level Representation for Pitch Estimation and Musical Audio Source Separation,” IEEE Journal of Selected Topics in Signal Processing, vol. 5, no. 6, Oct. 2011

[4] J. Salamon, E. Gomez, B. Rocha, ”Musical Genre Classifcation using Melody Features extracted from Polyphonic Music Signals”

[5] http://www.durrieu.ch/research/jstsp2010.html

[6] http://marsyas.info/download/data_sets/

[7] http://en.wikipedia.org/wiki/Blues

[8] http://en.wikipedia.org/wiki/Classical_music

[9] http://en.wikipedia.org/wiki/Country_music

[10] http://en.wikipedia.org/wiki/Disco

[11] http://en.wikipedia.org/wiki/Hip_hop_music

[12] http://en.wikipedia.org/wiki/Jazz

[13] http://en.wikipedia.org/wiki/Heavy_metal_music

[14] http://en.wikipedia.org/wiki/Pop_music

[15] http://en.wikipedia.org/wiki/Reggae

[16] http://en.wikipedia.org/wiki/Rock_music

[17] Sturm, Bob L. ”An analysis of the GTZAN music genre dataset.” Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies. ACM, 2012.

[18] Slaney, Malcolm. ”Auditory toolbox.” Interval Research Corporation, Tech. Rep 10 (1998): 1998.

[19] Peeters, Geoffroy. ”A large set of audio features for sound description (similarity and classification) in the CUIDADO project.” (2004).

[20] Panagakis, Yannis, Constantine Kotropoulos, and Gonzalo R. Arce. ”Music genre classification via sparse representations of auditory temporal modulations.” Proc. European Signal Process. Conf., Glasgow, Scotland. 2009.

[21] Rosenberg, Andrew, and Julia Hirschberg. ”V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure.” EMNLP-CoNLL. Vol. 7. 2007.

[22] Forney Jr, G. David. ”The viterbi algorithm.” Proceedings of the IEEE 61.3 (1973): 268-278.