Music Information Retrieval with Polyphonic Sounds and Timbre

Angelina A. Tzacheva and Keith J. Bell

University of South Carolina Upstate, Department of Informatics
800 University Way, Spartanburg, SC 29303, USA
e-mail: atzacheva@uscupstate.edu, bellkj@uscupstate.edu

Abstract. With the fast booming of online music repositories, the problem of building music recommendation systems is of great importance. There is an increasing need for content-based automatic indexing to help users find their favorite music objects. Numerous approaches to acoustic feature extraction have already been proposed. However, most are based on pitch (fundamental frequency). Few methods focus on timbre, especially with polyphonic sounds (multiple distinct instruments), which occur often in the real music world. In this work, we perform a study on automatic classification of musical instruments with polyphonic sounds. We use timbre descriptors. We report high classification accuracy.

1 Introduction

To many people in many cultures music is an important part of their way of life. Greek philosophers and ancient Indian philosophers define music as horizontally ordered tones, or melodies and vertically ordered harmonies. Common sayings such as ”the harmony of the spheres” and ”it is music to my ears” point to the notion that music is often ordered and pleasant to listen to.

With 20th century music, there is a vast increase in music listening as the radio gained popularity and phonographs were used to replay and distribute music. The focus of art music is characterized by exploration of new rhythms, styles, and sounds.

Today, we hear music media in advertisements, in films, at parties, at the philharmonic, etc. One of the most important functions of music is its effect on humans. Certain pieces of music have a relaxing effect, while others stimulate us to act, and some cause a change in or emphasize our mood. Music is not only a great number of sounds arranged by a composer, it is also the emotion contained within these sounds (Grekow and Ras, 2009).

The steep rise in music downloading over CD sales has created a major shift in the music industry away from physical media formats and towards Web-based (online) products and services. Music is one of the most popular types of online information and there are now hundreds of music streaming and download services operating on the World-Wide Web. Some of the music collections available are approaching the scale of ten million tracks and this has posed a major challenge for searching, retrieving, and organizing music content. Research efforts in music information retrieval have involved experts from music perception, cognition, musicology, engineering, and computer science engaged in truly interdisciplinary activity that has resulted in many proposed algorithmic and methodological solutions to music search using content-based methods (Casey et al., 2008).

This work contributes to solving the important problem of building music recommendation systems. Automatic recognition or classification of music sounds helps user to find favorite music objects, or be recommended objects of his/her liking, within large online music repositories. We focus on musical instrument recognition, which is a challenging problem in the domain.

Melody matching based on pitch detection technology has drawn much attention and many music information retrieval systems have been developed to fulfill this task. Numerous approaches to acoustic feature extraction have already been proposed.

The original audio signals are a large volume of unstructured sequential values, which are not suitable for traditional data mining algorithms, while the higher level data representative of acoustical features are sometimes not sufficient for instrument recognition. This has stimulated the research on instrument classification and new features development for content-based automatic music information retrieval.

We focus on timbre related features, and perform a study on automatic classification of musical instruments with polyphonic sounds.

The rest of the paper is organized as follows: section 2 reviews related work, section 3 discusses timbre, section 4 describes features, section 5 illustrates our methodology, section 6 shows the experiment results, and finally section 7 concludes.

2 Related Work

2.1 Classification of Monophonic Music Sound

(Martin and Kim, 1998) employed the K-NN (k-nearest neighbor) algorithm to a hierarchical classification system with 31 features extracted from cochleagrams. With a database of 1023 sounds they achieved 87% of successful classifications at the family level and 61% at the instrument level when no hierarchy was used. Using the hierarchical procedure increased the accuracy at the instrument level to 79% but it degraded the performance at the family level (79%). Without including the hierarchical procedure performance figures were lower than the ones they obtained with a Bayesian classifier. The fact that the best accuracy figures are around 80% and that Martin and Kim have settled into similar figures shows the limitations of the K-NN algorithm (provided that the feature selection has been optimized with genetic or other kind of techniques). Therefore, more powerful techniques should be explored.

Timbre classification systems have also been developed based on auditory processing and Kohonen self-organizing neural networks; where, data is preprocessed by peripheral transformations to extract perception features, then fed to the network to build a map. Such method did not quite well distinguish the instruments.

Bayes Decision Rules and Naive Bayes classifiers are simple probabilistic classifiers, by which the probabilities for the classes and the conditional probabilities for a given feature and a given class are estimated based on their frequencies over the training data. They are based on probability models that incorporate strong independence assumptions, which may, or may not have a bearing in reality, hence are naive. The resultant rule is formed by counting the frequency of various data instances, and can be used then to classify each new instance. (Brown, 1999) applied this technique to 18 Mel-Cepstral coefficients by a K-means clustering algorithm and a set of Gaussian mixture models. Each model was used to estimate the probabilities that a coefficient belongs to a cluster. Then probabilities of all coefficients were multiplied together and were used to perform the likelihood ratio test. It then classified 27 short sounds of oboe and 31 short sounds of saxophone with an accuracy rate of 85% for oboe and 92% for saxophone.

(Tzacheva and Bell, 2010) used timbre related features with bayesian neural network and J48 decision tree for classification of musical instruments. Authors proposed new features for preservation of temporal information. Since classifiers do not distinguish the order of the frames, they are not aware that frame t1 is closer to frame t2 than it is to frame t3 . Their proposed features allow for that distinction to be made. Authors achieved good classification accuracy with monophonic sounds.

2.2 Classification of Polyophonic Music Sound

One approach to address multi-timbre estimation in polyphonic sound is to apply sound separation techniques (Zhang and Ras, 2007) along with the traditional classifiers. Each time when one classification label from a set is assigned, the sound separation module is applied to subtract the estimated timbre feature from the signal so that the signal of the single instrument is separate from the polyphonic sound signal. Then the classifier can be applied again on the residue of the signal to assign another label. The sound separation process for each frame continues until the remnant of the signal is too weak to give any further timbre estimation. However, there is one problem in this method. After each sound separation process, the timbre information of the remaining instruments could be partially lost due to the overlap of multiple timbre signals, which make it difficult to further analyze the remnant of sound signal.

Instead of giving one classification label at a time, multi-label classification assigns multiple labels from a set, to the target estimation object. Some research on multi-label classification has been done in the text categorization area (Joachims, 1998), (Schapire and Singer, 2000), and in scene recognition (Boutell et al., 2004). Authors approached the problem by training a classifier with samples with multiple labels. A multi-label classification system for polyphonic sounds was proposed by (Jiang at al., 2009).

Typically a digital music recording, in form of a binary file, contains a header and a body. The header stores file information such as length, number of channels, sampling rate, etc. Unless it is manually labeled, a digital audio recording has no description of timbre or other perceptual properties. Also, it is a highly nontrivial task to label those perceptual properties for every piece of music based on its data content.

In music information retrieval area, a lot of research has been conducted in melody matching based on pitch identification, which usually involves detecting the fundamental frequency. Most content-based Music Information Retrieval (MIR) systems query by whistling/humming systems for melody retrieval. So far, few systems exist for timbre information retrieval in the literature or market, which indicates it as a nontrivial and currently unsolved task (Jiang et al., 2009). In addition, typically, the timbre related methods use monophonic sounds (one distinct instrument). Such timbre estimation algorithms are rarely successfully applied to polyphonic sounds (multiple distinct instruments), which occur more often in the real music world. In this work, we perform a study on automatic classification of musical instruments with polyphonic sounds. We use timbre descriptors.

3 Timbre

The definition of timbre is: in acoustics and phonetics - the characteristic quality of a sound, independent of pitch and loudness, from which its source or manner of production can be inferred. Timbre depends on the relative strengths of its component frequencies; in music - the characteristic quality of sound produced by a particular instrument or voice; tone color. ANSI defines timbre as the attribute of auditory sensation, in terms of which a listener can judge that two sounds are different, though having the same loudness and pitch. It distinguishes different musical instruments playing the same note with the identical pitch and loudness. So it is the most important and relevant facet of music information. People discern timbre from speech and music in everyday life.

Musical instruments usually produce sound waves with frequencies, which are an integer (a whole number) multiples of each other. These frequencies are called harmonics, or harmonic partials. The lowest frequency is the fundamental frequency f0, which has close relation with pitch. The second and higher frequencies are called overtones. Along with fundamental frequency, these harmonic partials distinguish the timbre, which is also called tone color. The human aural distinction between musical instruments is based on the differences in timbre.

The body of a digital audio recording contains an enormous amount of integers in a time-order sequence. For example, at a sampling rate 44,100Hz, a digital recording has 44,100 integers per second. This means, in a one-minute long digital recording, the total number of the integers in the time-order sequence will be 2,646,000, which makes it a very large data item. The size of the data, in addition to the fact that it is not in a well-structured form with semantic meaning, makes this type of data unsuitable for most traditional data mining algorithms.

Timbre is rather subjective quality and not of much use for automatic sound timbre classification. To compensate, musical sounds must be very carefully parameterized to allow automatic timbre recognition.

4 Feature Descriptions

Based on latest research in the area, MPEG published a standard group of features for digital audio content data. They are either in the frequency domain or in the time domain. For those features in the frequency domain, a STFT (Short Time Fourier Transform) with Hamming window has been applied to the sample data. From each frame a set of instantaneous values is generated. We use the following timbre-related features from MPEG-7:

Spectrum Centroid - describes the center-of-gravity of a log-frequency power spectrum. It economically indicates the pre-dominant frequency range. We use *Log Power Spectrum Centroid*, and *Harmonic Spectrum Centroid*.

Spectrum Spread - is the Root of Mean Square value of the deviation of the Log frequency power spectrum with respect to the gravity center in a frame. Like Spectrum Centroid, it is an economic way to describe the shape of the power spectrum. We use *Log Power Spectrum Spread*, and *Harmonic Spectrum Spread*.

Harmonic Peaks - is a sequence of local peaks of harmonics of each frame. We use the *Top 5 harmonic peaks - Frequency*, and *Top 5 Harmonic Peaks - Amplitude*.

In addition, we use the *Fundamental Frequency* as a feature in this study.

5 Methodology

We start with a set of musical instrument recordings containing two instruments mixed. One instrument is predominant. A recording is divided into frames of 40 milliseconds. Each frame overlaps the previous frame by 1/3. Next, we extract the above described features on each frame (object). Then, we train and test a classifier. We use a single-label approach and assign the predominant instrument as the correct class.

Formally defined as: let $S = \{X, F \cup L\}$ be the training dataset, where X is the set of instances, $L = \{l_1, l_2, ..., l_n\}$ is the set of all class labels, and $F = \{f_1, f_2, ..., f_m\}$ is the $m-$dimensional feature vector used to build a classifier C, which estimates the target object $t = \{v_1, v2, ..., v_m\}$. Then, the estimation result would be $C(t) = \{d : d \in L\}$.

We use two classifiers - a bayesian neural network and a J48 decision tree, and compare the results. Neural networks process information with a large number of highly interconnected neurons working in parallel to solve a specific problem. Neural networks learn by example. Decision trees represent a supervised approach to classification. It is a simple structure where non-terminal nodes represent tests on one or more attributes and terminal nodes reflect decision outcomes.

One class label with the highest confidence is assigned to the estimation object and the other candidate classes are simply ignored. Figure 1. illustrates our polyphonic sound classification method.

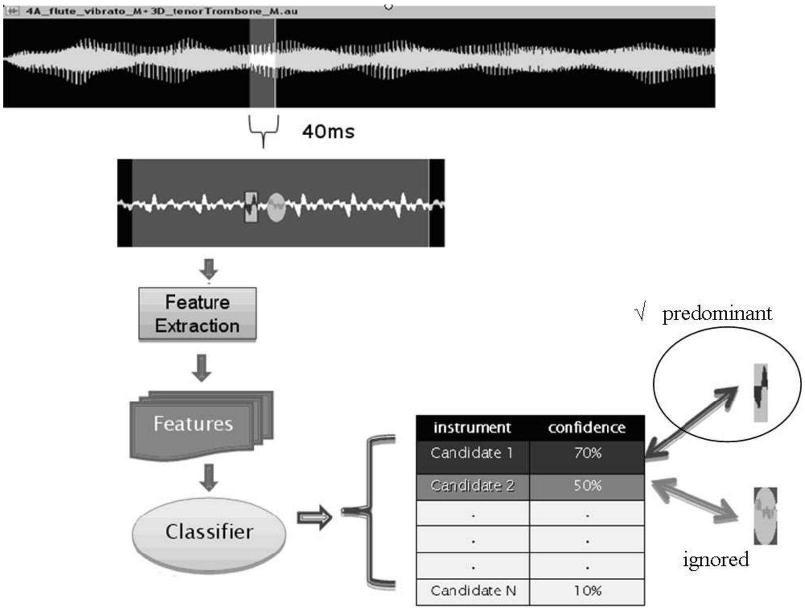

Fig. 1. Music information retrieval with polyphonic sounds and timbre

6 Experiment

We have chosen 6 instruments: viola, cello, flute, english horn, piano, and clarinet to our experiments. All recordings originate from MUMS CDs (Opolko and Wapnick 1987), which are used worldwide in similar tasks.

To create the analysis set a 0.2 second sample from 1.0 seconds to 1.2 seconds excerpt from the non-dominant instrument is mixed with the predominant instrument at every minute boundary (eg. 0.0s to 0.2s, 1.0s to 1.2s, 2.0s to 2.2s, etc.).

The mixed sound files used, with predominant instrument listed first, are: File1: Viola and Flute, File2: Flute and Englishhorn, File3: Viola and Englishhorn, File4: Cello and Piano, File5: Cello and Clarinet, File6: Clarinet and Piano.

Next, we split each recording into overlapping frames. Then, we extract the fatures described in the previous section 5. That produces a dataset with 1224 instances and 16 attributes.

We import the dataset into WEKA (Hall et al., 2009) data mining software for classification. We train two classifiers: Bayesian Neural Network and J48 Decision Tree. We test each one using 60% of samples for training and testing with remaining 40%, as well as testing using bootstrap. Results show Bayesian Neural Network has accuracy of 92.04% with 60/40 test and 94.28% via bootstrap testing. J48 decision tree has accuracy of 96.12% with 60/40 test and 98.93% via bootstrap testing. Summary results comparing the two classifiers are shown in Figure 2. The detailed results for Bayesian Neural Network and J48 Decision Tree are shown in Figure 3 and Figure 4 respectively.

	BayesNeuralNetwork	J48 DecisionTree
60/40 test	92.04 %	96.12 %
bootstrap test	94.28 %	98.93 %

Fig. 2. Correctly classified instances % - classifier comparison

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
60/40 test	0.912	0.015	0.964	0.912	0.937	0.996	2C_cello_bowed
	0.99	0.031	0.89	0.99	0.937	0.999	3A#_bflatclarinet
	0.867	0.034	0.907	0.867	0.886	0.975	3C_viola_bowed
	0.936	0.026	0.912	0.936	0.924	0.989	4A#_flute_vibrato
bootstrap	0.937	0.012	0.973	0.937	0.955	0.998	2C_cello_bowed
	0.996	0.027	0.899	0.996	0.945	1	3A#_bflatclarinet
	0.904	0.02	0.944	0.904	0.924	0.988	3C_viola_bowed
	0.951	0.016	0.944	0.951	0.948	0.994	4A#_flute_vibrato

Fig. 3. Bayesian neural network detailed accuracy by class

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
60/40 test	0.98	0.009	0.98	0.98	0.98	0.984	2C_cello_bowed
	0.98	0.008	0.97	0.98	0.975	0.983	3A#_bflatclarinet
	0.956	0.023	0.942	0.956	0.949	0.976	3C_viola_bowed
	0.927	0.013	0.953	0.927	0.94	0.963	4A#_flute_vibrato
bootstrap	0.992	0.004	0.992	0.992	0.992	0.998	2C_cello_bowed
	0.992	0.001	0.996	0.992	0.994	0.999	3A#_bflatclarinet
	0.982	0.004	0.988	0.982	0.985	0.998	3C_viola_bowed
	0.993	0.005	0.981	0.993	0.987	0.998	4A#_flute_vibrato

Fig. 4. J48 decision tree detailed accuracy by class

7 Conclusions and Directions for the Future

We produce a music information retrieval system, which automatically classifies musical instruments. We use timbre related features. Our proposed methodology successfully classifies polyphonic music sounds (multiple distinct instruments). We report high classification accuracy. This presents an improvement over previous timbre methods, which primarily focus on monophonic sounds.

This work contributes to solving the important problem of building music recommendation systems. Automatic recognition or classification of music sounds helps user to find favorite music objects within large online music repositories. It can also be applied to recommend musical media objects of user's liking. Directions for the future include automatic detection of emotions (Grekow and Ras, 2009) contained in music files.

References

1. M.R. Boutell, J. Luo, X. Shen, and C.M. Brown (2004). Learning multi-label scene classification, in Pattern Recognition, Vol. 37, No. 9, pp. 1757-1771.
2. J.C. Brown (1999). Musical instrument identification using pattern recognition with cepstral coefficients as features, Journal of Acousitcal society of America, 105:3, pp. 1933-1941
3. M.A. Casey, R. Veltkamp, M. Goto, M. Leman, C. Rhodes, M. Slaney, (2008). Content-Based Music Information Retrieval: Current Directions and Future Challenges. Proceedings of the IEEE, Vol. 96, Issue 4., pp. 668-696
4. J. Grekow and Z.W. Ras (2009). Detecting Emotion in Classical Music from MIDI Files, Foundations of Intelligent Systems, Proceedings of 18th International Symposium on Methodologies for Intelligent Systems (ISMIS'09), (Eds. J. Rauch et al), LNAI, Vol. 5722, Springer, Prague, Czech Republic, pp. 261-270.
5. M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I. H. Witten (2009). The WEKA Data Mining Software: An Update, SIGKDD Explorations, Vol. 11, Issue 1. New Zealand.
6. W. Jiang, A. Cohen, and Z. W. Ras (2009). Polyphonic music information retrieval based on multi-label cascade classification system, in Advances in Information and Intelligent Systems, Z.W. Ras, W. Ribarsky (Eds.), Studies in Computational Intelligence, Springer, Vol. 251, pp. 117-137.
7. T. Joachims (1998). Text categorisation with support vector machines: Learning with many relevant features, in Proceedings of Tenth European Conference on Machine Learning, pp. 137-142.
8. K.D. Martin and Y.E. Kim (1998). Musical instrument identification: A pattern recognition approach, in Proceedings of Meeting of the Acoustical Society of America, Norfolk, VA
9. F. Opolko and J. Wapnick (1987). MUMS-McGillUniversityMasterSamples.CD's.
10. R. Schapire and Y. Singer (2000). BoosTexter: A boosting-based system for text categorization, in Machine Learning, Vol. 39, No. 2/3, pp. 135-168.
11. A.A. Tzacheva and K.J. Bell (2010). Music Information Retreival with Temporal Features and Timbre, in Proceedings of 6th International Conference on Active Media Technology (AMT 2010), Toronto, Canada, LNCS 6335, pp. 212-219.
12. X. Zhang and Z.W. Ras (2007). Sound isolation by harmonic peak partition for music instrument recognition, in the Special Issue on Knowledge Discovery, Fundamenta Informaticae Journal, IOS Press, Vol. 78, No. 4, pp. 613-628.