Music Information Retrieval with Temporal Features 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. At a time when the quantity of music media surrounding us is rapidly increasing and the access to recordings as well as the amount of music files available on the Internet is constantly growing, the problem of building music recommendation systems is of great importance. In this work, we perform a study on automatic classification of musical instruments. We use monophonic sounds. The latter have successfully been classified in the past, with main focus on pitch. We propose new temporal features and incorporate timbre descriptors. The advantages of this approach are: preservation of temporal information and high classification accuracy.

1 Introduction

Music has accompanied man for ages in various situations. 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.

This has stimulated the research on instrument classification and new features development for content-based automatic music information retrieval. 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.

We propose new dynamic features, which preserve temporal information, for increased accuracy with classification.

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

2 Related Work

(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.

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.

Neural networks process information with a large number of highly interconnected processing neurons working in parallel to solve a specific problem. Neural networks

learn by example. (Cosi, 1998) developed a timbre classification system based on auditory processing and Kohonen self-organizing neural networks. Data were preprocessed by peripheral transformations to extract perception features, then were fed to the network to build the map, and finally were compared in clusters with human subjects' similarity judgments. In the system, nodes were used to represent clusters of the input spaces. The map was used to generalize similarity criteria even to vectors not utilized during the training phase. All 12 instruments in the test could be quite well distinguished by the map.

Binary Tree is a data structure in which each node contains one parent and not more than 2 children. It has been pervasively used in classification and pattern recognition research. Binary Trees are constructed top-down with the most informative attributes as roots to minimize entropy. (Jensen and Amspang, 1999) proposed an adapted Binary Tree with real-valued attributes for instrument classification regardless of pitch of the instrument in the sample.

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 exists for timbre information retrieval in the literature or market, which indicates it as a nontrivial and currently unsolved task (Jiang et al., 2009).

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.

3.1 Challenges in Timbre Estimation

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 and Instruments

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 Design of New Temporal Features

Describing the whole sound produced by a given instrument by single value of a parameter which changes in time, may be omitting a large amount of relevant information encoded within the sound. For example, calculating the average of the values taken in certain time points. For this reason, we design features, which characterize the changes of sound properties in time.

5.1 Frame Pre-processing

The instrument sound recordings are divided into frames. We pre-process the frames, in way that each frame overlaps the previous frame by 2/3 as shown on Figure 1. In other words, if frame1 is abc, then frame2 is bcd, frame3 is cde, and so on. This preserves temporal information contained in the sequential frames.

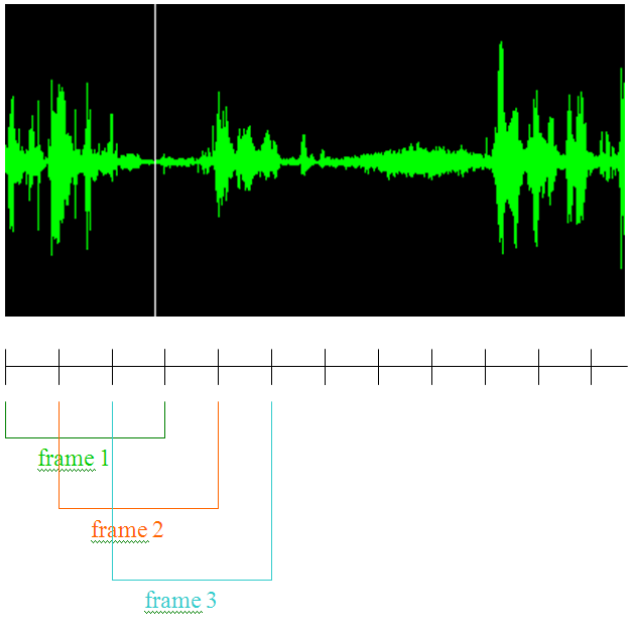

Fig. 1. Overlapping frames

5.2 New Temporal Features

After the frames have been pre-processed, we extract the timbre related features described in section 4 for each frame. We build a database from this information, shown in Table 1. $x_1, x_2, x_3, ..., x_n$ are the tuples (or objects - the overlapping frames). Attribute **a** is the first feature extracted on them (log power spectrum centroid). We have a total of 7 attributes, 2 of which in a vector form.

Next, we calculate 6 new features based on the attribute **a** value for the first 3 frames t_1, t_2, and t_3. The new features are defined as follows:

$$d_1 = t_2 - t_1$$

$$d_2 = t_3 - t_2$$

$d_3 = t_3 - t_1$

$tg(\alpha) = (t_2 - t_1)/1$

$tg(\beta) = (t_3 - t_2)/1$

$tg(\gamma) = (t_3 - t_1)/2$

This process is performed by our Temporal Cross Tabulator. $y_1, y_2, y_3, ..., y_n$ are the new objects created by cross tabulation, which we store in a new database - Table 2. So, our first new object y_1 in Table 2 is created from the first 3 objects x_1, x_2, x_3 in Table 1. Our next new object y_2 in Table 2 is created from x_2, x_3, x_4 in Table 1. New object y_3 in Table 2 is created from x_3, x_4, x_5 in Table 1.

Since classifiers do not distinguish the order of the frames, they are not aware that frame t_1 is closer to frame t_2 than it is to frame t_3. With the new features α, β, and γ, we allow for that distinction to be made. $tg(\alpha) = (t_2 - t_1)/1$ takes into consideration that the distance between t_2 and t_1 is 1, while $tg(\gamma) = (t_3 - t_1)/2$ because the distance between t_3 and t_1 is 2.

This temporal cross-tabulation increases the current number attributes 6 times. In other words, for every attribute (or feature) from Table 1, we have $d_1, d_2, d_3, \alpha, \beta$, and γ in Table 2. Thus, 15 current attributes (or features: log power spectrum centroid, harmonic spectrum centroid, log power spectrum spread, harmonic spectrum spread, fundamental frequency, top 5 harmonic peaks amplitude - each peak as a separate attribute, and top 5 harmonic peaks frequency - each peak as a separate attribute) multiplied by 6 = 90. The complete Table 2 has 90 attributes, which comprises our new dataset.

6 Experiment

We have chosen 6 instruments: viola, cello, flute, english horn, piano, and clarinet for our experiments. All recordings originate from MUMS CD's (Opolko and Wapnick 1987), which are used worldwide in similar tasks. We split each recording into overlapping frames, and extract the new temporal features as described in the previous section 5. That produces a dataset with 1225 tuples and 90 attributes.

We import the dataset into WEKA (Hall et al., 2009) data mining software for classification. We train two classifiers: Bayesian Neural Network and J45 Decision Tree. We test using bootstrap. Bayesian Neural Network has accuracy of 81.14% and J45 has accuracy of 96.73%. The summary results of the classification are shown in Figure 3 and the detailed results in Figure 4.

7 Conclusions and Directions for the Future

We produce a music information retrieval system, which automatically classifies musical instruments. We use timbre related features. We propose new temporal features. The advantages of this approach are preservation of temporal information, and high classification accuracy. This work contributes to solving the important problem of building music recommendation systems. Automatic recognition or classification of music sounds

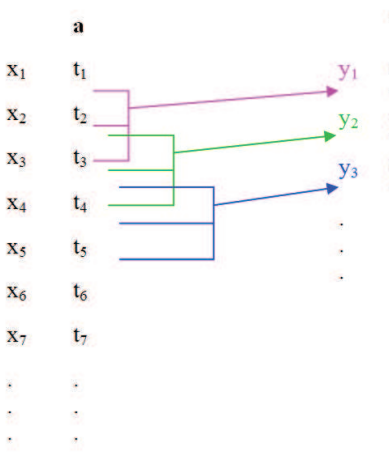

Table 1. Features extracted on overlapping frames

	d_{1a}	d_{2a}	d_{3a}	α	β	γ
y_1	$d_1 = t_2 - t_1$	$d_2 = t_3 - t_2$	$d_3 = t_3 - t_1$	$tg(\alpha) = (t_2 - t_1)/1$	$tg(\beta) = (t_3 - t_2)/1$	$tg(\gamma) = (t_3 - t_1)/2$
y_2	$d_1 = t_3 - t_2$	$d_2 = t_4 - t_3$	$d_3 = t_4 - t_2$	$tg(\alpha) = (t_3 - t_2)/1$	$tg(\beta) = (t_4 - t_3)/1$	$tg(\gamma) = (t_4 - t_2)/2$
y_3	$d_1 = t_4 - t_3$	$d_2 = t_5 - t_4$	$d_3 = t_5 - t_3$	$tg(\alpha) = (t_4 - t_3)/1$	$tg(\beta) = (t_5 - t_4)/1$	$tg(\gamma) = (t_5 - t_3)/2$

Table 2. Temporal cross tabulation

Fig. 2. New Temporal Features

	Correctly Classified	Incorrectly Classified	Correct %	Incorrect %
Bayesian Neur. Net.	994	231	81.1429 %	18.8571 %
J45	1185	40	96.7347 %	3.2653 %

Fig. 3. Results Summary

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC	Class
BNNet							
	0.957	0.141	0.548	0.957	0.697	0.98	2A#_piano
	0.931	0.016	0.911	0.931	0.921	0.994	2C_cello_bowed
	0.783	0.011	0.945	0.783	0.856	0.988	3A#_bflatclarinet
	0.712	0.046	0.732	0.712	0.722	0.947	3A#_englishhorn
	0.693	0.008	0.934	0.693	0.796	0.971	3C_viola_bowed
	0.792	0	1	0.792	0.884	0.994	4A#_flute_vibrato
J45							
	0.973	0.015	0.919	0.973	0.945	0.996	2A#_piano
	0.973	0.009	0.953	0.973	0.963	0.999	2C_cello_bowed
	0.996	0.002	0.992	0.996	0.994	1	3A#_bflatclarinet
	0.935	0.006	0.966	0.935	0.95	0.996	3A#_englishhorn
	0.926	0.003	0.981	0.926	0.953	0.999	3C_viola_bowed
	0.981	0.004	0.985	0.981	0.983	1	4A#_flute_vibrato

Fig. 4. Results - Detailed Accuracy by Class

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. 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
2. 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
3. P. Cosi (1998). Auditory modeling and neural networks, in Course on speech processing, recognition, and artificial neural networks, LNCS, Springer
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. K. Jensen and J. Arnspang (1999). Binary decision tree classification of musical sounds, in Proceedings of International Computer Music Conference, Beijing, China
7. 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.
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.