Semi-Automatic Approach for Music Classification

Tong Zhang

1. INTRODUCTION

Digital music are becoming more and more popular in people's life. It is quite common for a person to own thousands of digital music pieces these days, and users may build their own music library through music management systems or software such as the Music Match Jukebox. While for professional music databases, labors are often hired to manually classify and index music assets according to predefined criteria; most users do not have the time or patience to browse through their personal music collections and manually index the music pieces one by one. On the other hand, if music assets are not properly categorized, it may become a big headache when the user wants to search for a certain music piece among the thousands of pieces in a music collection, or when one simply wants to browse through his collections to find out what he has.

The efficiency and ease-of-use of music management systems can be greatly improved by adding automatic music classification and retrieval functions, that is, to automatically categorize music in the database and retrieve music pieces according to certain requests. However, a complete automation in categorizing music into proper classes for browsing and searching is not yet supported by today's technology. Also, the issue of music classification is subjective to some extent as each user may have his own preferences for categorizing music. Therefore, automatic music categorization has rarely been used in existing music management systems or software.

In this research, we propose an approach of semi-automatic music classification. That is, music assets are automatically classified into a number of basic categories first, which is achievable by analyzing features of the music signal. For example, vocal music has distinguished temporal and spectral patterns from those of pure instrumental music; while within instrumental music, different instrument families (string, wind, percussion, etc.) may be differentiated by analyzing music timbre. Based on this brief classification of music, further categorization and retrieval within each class will be much more convenient. The user may further classify within certain categories or change some classification results manually with the assistance of a well-designed user interface. Overall, this semi-automatic approach greatly improves efficiency in music management including classification, browsing and retrieval. At the same time, it also adapts to the users' individual preferences and the gradual progress of music processing techniques.

The rest of this report is organized as follows. The framework of the proposed system is presented in section 2. The automatic music classification procedures are described in section 3. Then, the manual adjustments of classification results are introduced in sections 4. Finally, open issues of the system are discussed in section 5, followed by conclusion remarks and future research plans.

2. SYSTEM OVERVIEW

A brief classification of music pieces within a music collection are conducted based on extracting and analyzing audio features. An illustration of the proposed scheme is shown in Figure 1. First of all, human voice can be detected by checking several audio features in both the time and frequency domains. Therefore, vocal music (singing) and pure instrumental music (no human voice) can be separated. Then, within vocal music, chorus and solo may be separated by checking some spectral features, and solo songs can be further divided into male solo and female solo based on pitch features. Within pure instrumental music, symphonies can be distinguished according to their featured patterns. Other instrumental music are analyzed in terms of harmonic, timbre and rhythmic features, and they can be further divided into the string instruments (violin, cello, etc.), the wind instruments (such as trumpet, horn, flute), the keyboard instruments (e.g. piano, organ), the percussion instruments (drums, etc.), and others.

Next, with the help of a specially designed GUI, the user may view the classification results and browse through music items within a certain music category. The system provides functionalities for the user to add new classes, to merge or remove existing classes, and to move a music piece from one category to another. For example, the user may set up a new folder of piano music within the keyboard instrument class, or set up a folder for the songs of a particular singer within the male or female solo class. In case of misclassifications, the user may easily move a misclassified music piece to the right category.

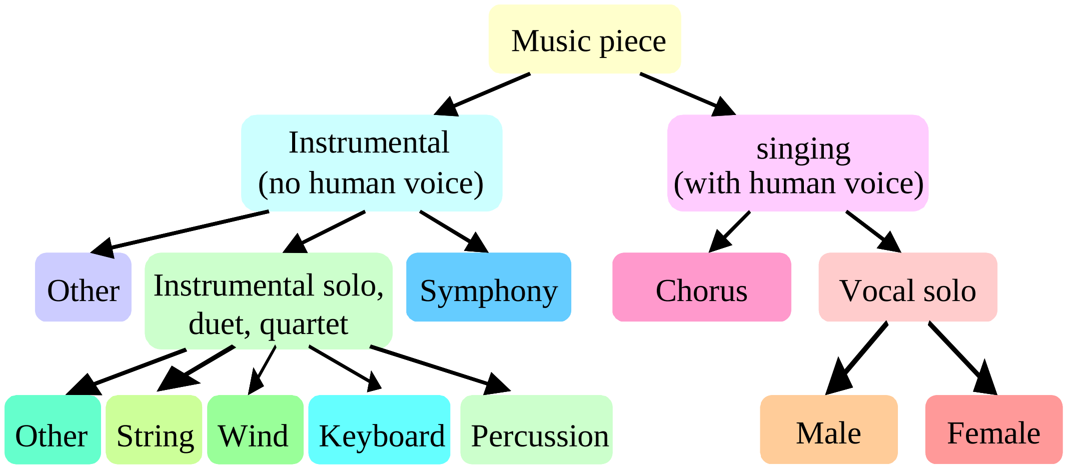


Figure 1: Brief classification of music assets based on automatic content analysis.

3. AUTOMATIC MUSIC CATEGORIZATION

The flowchart for the automatic music categorization is shown in Figure 2, and processing blocks of this flowchart are described in the following subsections.

3.1. Distinguishing singing from pure instrumental music

According to Fig.2, the first step is to separate singing (i.e. with human voice) from pure instrumental music (i.e. without human voice). Zhang and Kuo proposed a method in [1] to classify an audio signal into a number of basic audio types, including pure instrumental music (denoted as "pure music" in [1]) and singing (denoted as "song" in [1]). While instrumental music is normally indicated by stable frequency peaks in the spectrogram, i.e. spectral peak tracks which remain at fixed frequency levels, human voice(s) in singing is revealed by spectral peak tracks with changing pitches. Also, there are normally regular peaks and troughs in the energy function and the average zero-crossing rates of the singing signal. Some examples are shown in Figures 3 and 4.

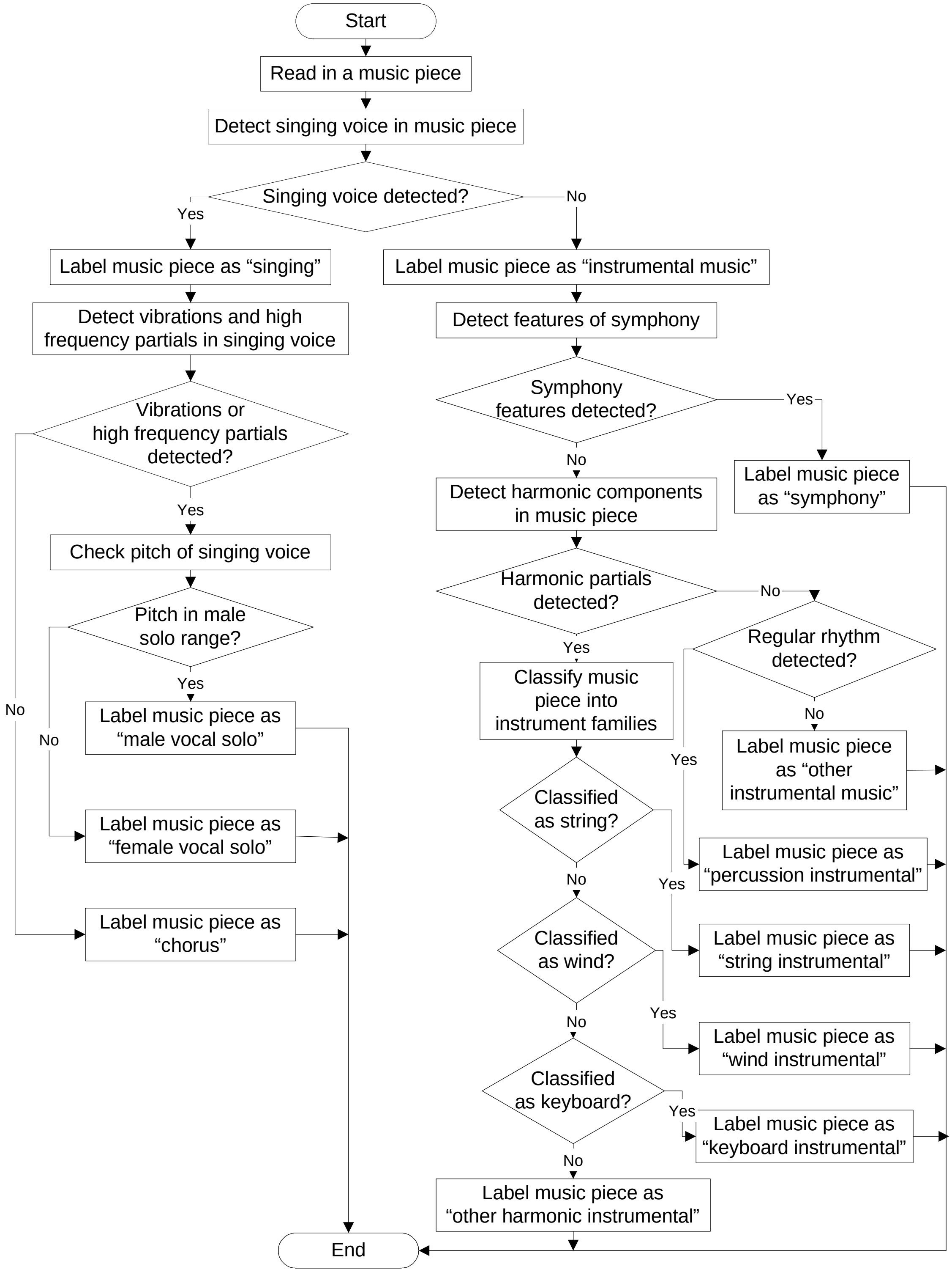
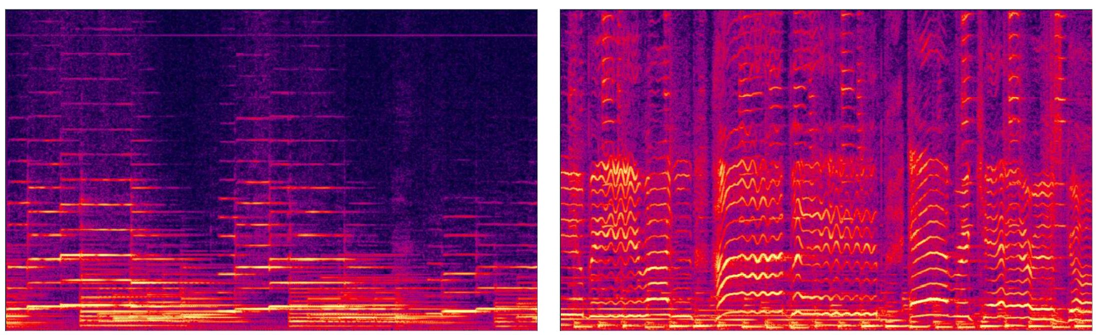


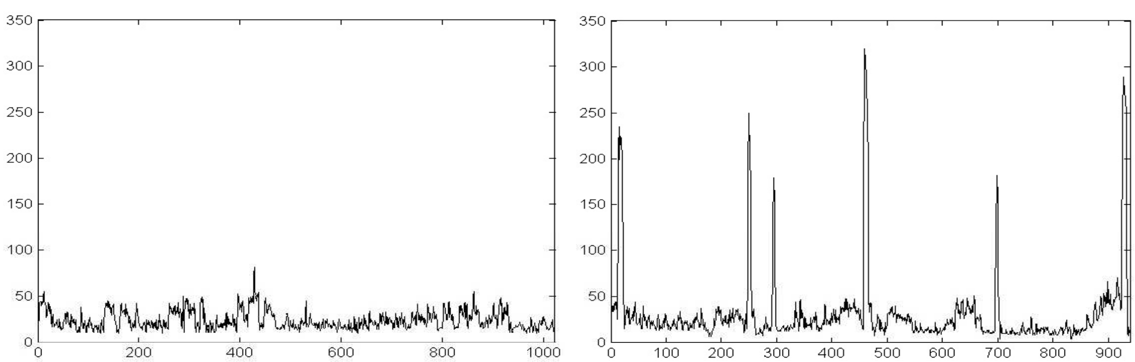
Figure 2: Flowchart for automatic music classification.

In Fig. 3, spectrograms of a pure instrumental music (clarinet-piano duet) and a female vocal solo are displayed. It can be seen that the frequency level of spectral peak tracks remain fixed during the duration of each note in instrumental music, thus in terms of image feature, the spectrogram contains mainly of flat horizontal lines. While in the case of vocal music, the pitch changes during the singing of one syllable, and in terms of image features, there are lines with up-and-downs and ripples. In Fig. 4, the average zero-crossing rate (ZCR) curves of a pure instrumental music (guitar) and a male vocal solo are plotted. Compared to the ZCR of instrumental music which mostly stays within a relatively small range of amplitude, there are high peaks in the ZCR of vocal music due to the pronunciation of some consonant components.



pure instrumental music female vocal solo

Figure 3: Spectrograms of pure instrumental music and singing.



pure instrumental music male vocal solo

Figure 4: Average zero-crossing rates of pure instrumental music and singing.

In [1], pure instrumental music was distinguished from other types of music, including singing, based on the average zero-crossing rate and the fundamental frequency properties. Four aspects were checked: the degree of being harmonic, the degree of the fundamental frequency's concentration on certain values during a period of time, the variance of the average zero-crossing rates, and the range of amplitudes of the average zero-crossing rates. For each aspect, there is one empirical threshold set and a decision value defined. If the threshold is satisfied, the decision value is set to 1; otherwise, it is set to a fraction between 0 and 1 according to the distance to the threshold. The four decision values are averaged with predetermined weights to derive a total probability of the music piece's being pure music. For a music piece to be labeled as "pure music", this probability should be above a certain threshold and at least three of the decision values should be above 0.5.

There is one prior art for locating singing voice segments within music signals proposed by Berenzweig and Ellis in [2], in which a speech recognition engine is applied to the music signal. The idea is that a speech-trained acoustic model would respond in a detectably different manner to singing (which shares some attributes of regular speech, such as formant structure and phone transitions) than to pure instrumental music. While this method achieves a relatively high level of accuracy (80% accurate at the frame level as reported in [2]), a speech recognition engine (which, however, is not always available, and can be computationally and economically expensive) is needed for this approach. A totally different scheme was described in [3], in which four audio features, namely, the energy function, the average zero-crossing rate, the harmonic coefficient, and the spectral flux are combined to detect the start of singing voice in a song. This method is rather efficient and can be easily applied to the purpose of separating vocal music pieces from pure instrumental music pieces.

In this paper, we differentiate instrumental music from vocal music by detecting features of human voice(s) in the music signal. Two kinds of audio features are checked. On one hand, the frequency level of spectral peak tracks is checked in the frequency domain (for definition and calculation of spectral peak tracks, please refer to [1]). If frequencies of a large percentage of the spectral peak tracks change significantly over time, i.e. change rate higher than a predefined threshold, due to pronunciations of vowels and vibrations of vocal chords, it is indicated that singing voice is detected. On the other hand, average zero-crossing rates (ZCR) of the music signal are computed. If there are a number of significant peaks (i.e. sharp and high, defined using a set of predefined thresholds) found in the ZCR curve which result from pronunciations of consonants (in both solos and choruses), singing voice is detected. Once singing voice is detected (with one or both of the features), the music piece is labeled as “singing”. Otherwise, it is labeled as “instrumental music”.

3.2. Distinguishing vocal solo from chorus

Then, within the category of “singing”, music pieces are further separated into the classes of “solo” and “chorus”. As illustrated in the upper plots of figure 5, in solo songs, vibrations of vocal chords of the singer during a long held note are reflected as ripples in the spectral peak tracks; while in choruses, voice vibrations of different singers offset each other and there are no significant ripples appearing. Also, the spectral peak tracks are thicker in chorus due to the mix of different singers’ voices, so that the partials in the mid to higher frequency bands can not be revealed (i.e. partials overlap with each other) in the frequency domain. Examples are shown in the lower plots of Figure 5, which are spectra at certain moments (indicated by the black lines in the upper plots) of the song. We can see in these examples that, while there are harmonic partials denoted by significant peaks (sharp and high, and have a common divisor in frequency – i.e. the fundamental frequency) up to about 6500Hz in the spectrum of the solo signal; they are not available above 2000Hz in the spectrum of the chorus signal.

Thus, solos can be distinguished by detecting ripples in the spectrogram and/or by detecting significant harmonic partials in frequency bands higher than a certain frequency level (e.g. higher than 2000Hz or 3000Hz). One way to detect ripples in the spectral peak tracks is to calculate the first-order derivative of the frequency value of each track, and ripples are reflected as a regular pattern in which positive and negative values appear alternately. In contrast, for tracks in a chorus, the derivative values are mostly around zero. On the other hand, a method for detecting significant peaks in a spectrum is proposed in [1]. This method can be applied to check whether there are significant harmonic partials in higher frequency bands of the signal, and decide whether it is a solo song or chorus.

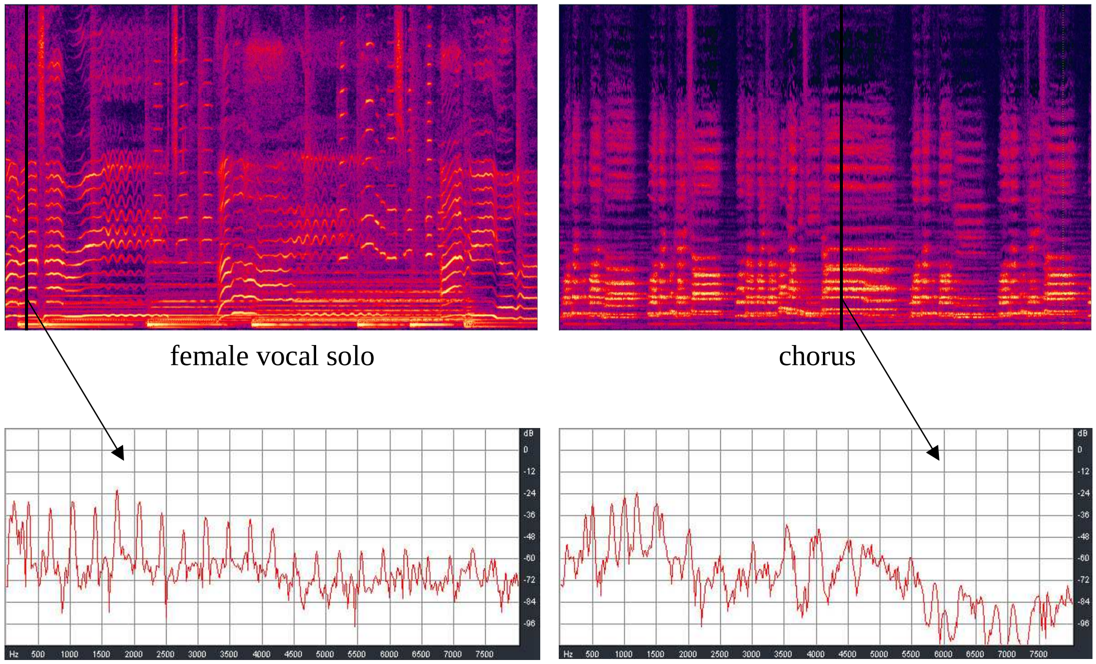
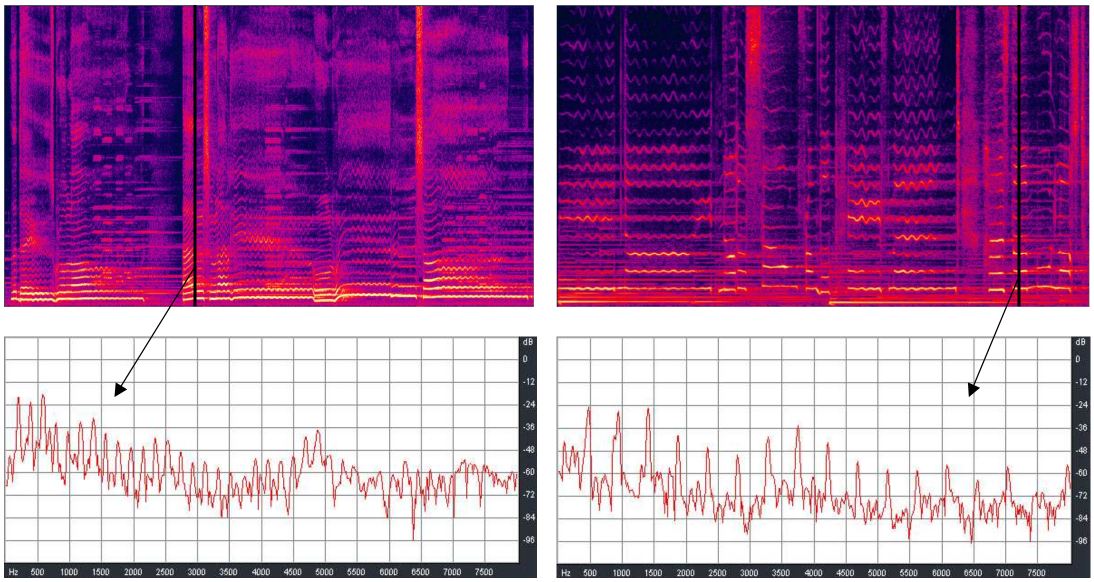


Figure 5: Spectrograms and spectra of vocal solo and chorus.

3.3. Discriminating female vocal solo vs. male vocal solo

In the next step, vocal solos are categorized further into male solo and female solo. This can be done by checking the range of pitch values in the song. There are various ways of estimating the pitch or the fundamental frequency (which is equivalent to pitch) of speech and music signals in the literature [1],[4]. The pitch of the singer's voice is estimated once every certain period of time along the song, e.g. every 500ms or every second. It is first checked whether there are harmonic partials of the singer's voice available at that moment – i.e. if there are spectral peak tracks (harmonic components) and if they have the features of singing voice (not musical instruments). Then, the pitch value is estimated when available. If most of the pitch values (e.g. over 90%) are lower than a predetermined higher threshold T_h (e.g. 250Hz), and at least some of the pitch values (e.g. no less than 10%) are lower than a predetermined lower threshold T_l (e.g. 200Hz), the song is labeled as "male vocal solo". Otherwise, it is labeled as "female vocal solo" which includes children's vocal solo as well. Examples of a male solo and a female solo are shown in Figure 6. The spectrograms of portions of the two songs are displayed in the upper plots of Fig. 6, and spectra at two selected moments (as indicated by the black lines in the upper plots) are shown in the two lower plots. It can be seen that the male solo has significantly lower pitch than that of the female solo, about 180Hz vs. 480Hz in these two examples.



male vocal solo female vocal solo

Figure 6: Spectrograms and spectra of male vocal solo and female vocal solo.

3.4. Recognizing symphonies within instrumental music

Within the category of instrumental music, symphonies are first distinguished. *Symphony* is defined as a music piece for large orchestra, usually in three or four movements. And *movement* is defined as a self-contained segment of a larger work, found in works such as sonatas, symphonies, concertos, etc. Another related term is *form*. It is the structure of a composition, the frame upon which it is constructed. Form is based upon repetition, contrast, and variation. Certain specific forms include sonata-allegro form, binary form, rondo, etc. There are some regularities in movements of symphonies. For example, the first movement of a symphony is usually a fairly fast movement, weighty in content and feeling. The vast majority of first movements are in Sonata Form. While in most symphonies, the second movement will be slow and solemn in character.

Since a symphony is composed of multiple movements and repetitions, there is an alternation between relatively high volume audio signal (e.g. performance of the whole orchestra) and low volume audio signal (e.g. performance of single instrument or a few instruments of the orchestra) along the music piece. Plotted in Figure 7 is the energy function, which represents the volume variation over time, of one symphony. Shown in boxes A and B are examples of high volume signal intervals which have two distinctive features:

- One is that the average energy of the interval is higher than a predetermined threshold level T_1, because the whole orchestra are performing;
- And the other is that there is no energy lower than a predetermined threshold level T_2 during the interval, because different instruments in the orchestra compensate each other (unlike the signal of a single instrument in which there might be a dip in energy between two neighboring notes).

Shown in boxes C and D are examples of low volume signal intervals which have average energy levels lower than a threshold T_3 (because only few instruments are playing) and the highest energy in the interval is lower than threshold T_4 . Also, the content in box F is apparently a repetition of that in box E with only minor variations.

Thus, by checking the existence of alternation between high volume and low volume intervals (with each interval longer than a certain threshold) and/or repetition(s) in the whole music piece, symphonies will be distinguished. Repetition(s) in a music piece might be detected in a number of ways in the time domain, the frequency domain or the tempo domain, etc. One simple method to detect repetition is to compute the autocorrelation of the energy function as illustrated in Fig. 7 and the repetition will be reflected as a remarkable peak in the autocorrelation curve.

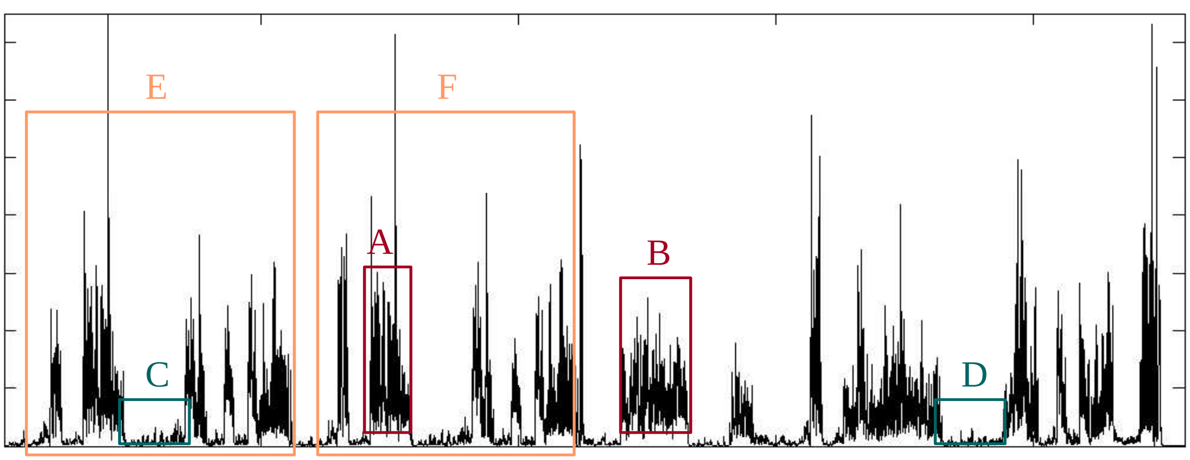


Figure 7: The energy function of a symphony music piece.

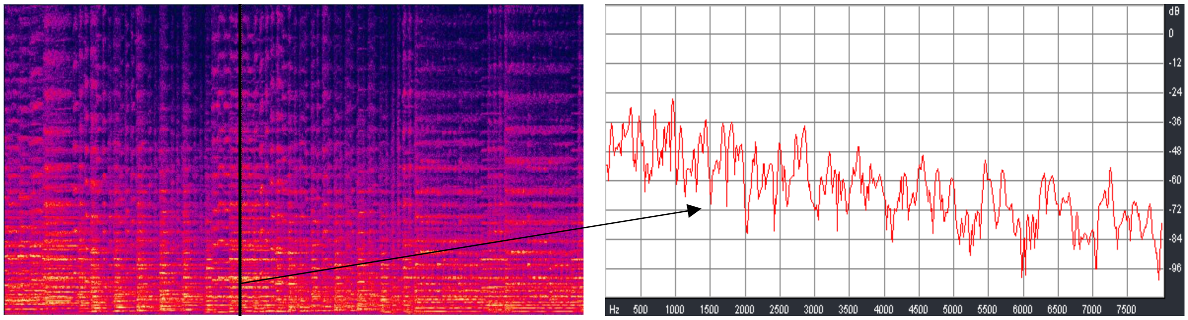


Figure 8: spectrogram and spectrum of a symphony (portion) music piece.

Besides, as shown in Figure 8, during the high-volume intervals, even though there are still remarkable spectral peak tracks which can be detected, the relation among harmonic partials of the same note is not as obvious (as illustrated in the right plot of Fig. 8) as in music which contains only one or a few instruments, because the mix of a large number of instruments makes the partials of different instruments overlap with each other in the frequency domain. Therefore, the lack of harmonic partials in the frequency domain (refer to [1] on how to detect harmonic partials) in the high-volume intervals is another feature of symphonies, which might be used in combination with the above methods in distinguishing symphonies from other types of instrumental music.

3.5. Music instrument classification

A system and method were proposed in [5] for the classification of music pieces according to the types of instruments involved. Whereas there were occasional misclassifications among instruments which belong to the same family (e.g. cello, viola and violin), rather reliable results could be obtained for categorizing music pieces into instrument families using the method in [5]. The instrument families include the string family (violin, viola, cello, etc.), the wind family (flute, horn, trumpet, etc.) and the keyboard family (piano, organ, etc.).

A music piece is first segmented into notes by detecting note onsets in the energy envelop of the music signal. After that, the pitch and harmonic partials are estimated in each note [5]. If harmonic components are not identified in most of the notes, then, it is detected whether there is a regular rhythm in the music piece by checking positions of note onsets. If a quasi-periodic relation exists among locations of note onsets in many portions of the music piece, it is labeled as “percussion instrumental” (e.g. drums). Otherwise, if note onsets could not be detected in most parts of the music piece (e.g. more than 50% of the total length of music) and/or harmonic partials are not detected in most notes (e.g. more than 50% of the notes) which might happen to music played with a number of different instruments (e.g. a band), the music piece is subsequently labeled as “other instrumental music”.

For harmonic music, temporal, spectral and partial features of each note are computed, such as rising and releasing speeds of the temporal envelope of the note; energy distribution among sub-bands in the spectrum of the note; as well as brightness, tristimulus parameters, odd partial ratio, dominant tones, and inharmonicity of the note. These features are normalized so as to be independent of the loudness, the length and the pitch of the note. Afterwards, the feature vector is sent to a pre-trained neural network for classification. Later, the classification results of all notes in the music piece are summarized, and the music piece is categorized as one of the following music types: “string instrumental”, “wind instrumental”, “keyboard instrumental” or “other harmonic instrumental”.

3.6. Music categorization based on tag information

In many audio formats, there are metadata of the music piece at the header or certain tag fields of the audio file. For instance, there is a TAG at the end of an MP3 music file (the last 128 bytes) which contains fields for information such as title, artist, album, year, genre, etc. However, in many MP3 songs the TAG may not exist or some fields might be empty. Nevertheless, when the information does exist, it may be extracted and used in the automatic music classification procedure. For example, samples in the “other instrumental” category might be further classified into the groups of “instrumental pop”, “instrumental rock” and so on based on the genre field of the TAG.

4. MANUAL ADJUSTMENT OF CLASSIFICATION RESULTS

For music classification in a professional database or personal collection, the above described automatic categorization is first applied. Then, the user could manually adjust folders and assign music pieces to folders. In the automation part, the user could also choose which classification tools to use. For example, if there are no symphonies in the collection, then the “symphony” folder could be removed and the tools for detecting symphony features could be disabled.

One of the most important components of this work is to design an easy-to-use user interface. One exemplary user interface is illustrated in Figure 9, in which results from the automatic classification are shown. The folders representing different music categories are displayed at the left side, each denoted by a button. The user may click on one button and the items within the folder are shown at the right side. For example, shown in this figure are items in the “female solo” folder (the selected folder is indicated by the solid color of its button). Each item is represented by the title of the music piece which might include the name of the song, the artist’s name when available, etc. There is also one thumbnail available for each song.

The thumbnail is a short audio clip (e.g. 10 to 30 seconds long) containing highlights of the music piece. It can be played by double-clicking the icon to the right of each song. The user may grasp major features of the music piece by listening only to the thumbnail, thus browse through the music collection in an efficient way. In this work, we generate audio thumbnails following a simple rule: for pure instrumental music, the first 10 seconds of the music piece is extracted and serves as the thumbnail; for vocal music, the first 10 seconds of singing is excerpted as the thumbnail, and for this purpose, the start of the singing voice is detected first [3]. There are other approaches for creating audio thumbnails in the literature. For example, a method was described in [6] which attempted to identify the chorus or refrain of a song by identifying repeated sections of the audio waveform. However, errors may occur in songs that do not meet the structural assumptions upon which the approach was built.

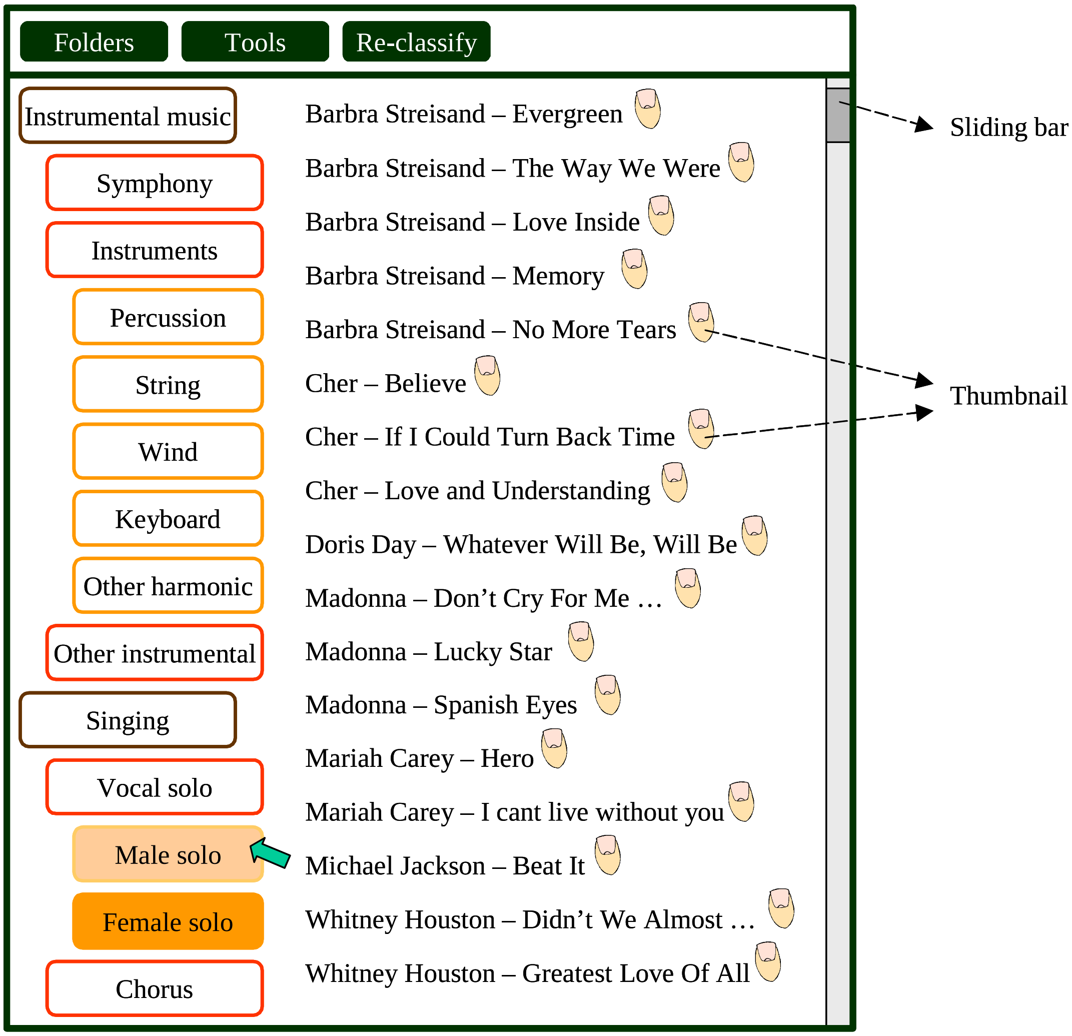


Figure 9: Exemplary user interface for semi-automatic music classification.

The following functions are provided to the user for manual adjustments of classification results:

- Manage folders. When clicking on the "Folders" button at the top of the GUI, a pull-down menu will appear which contains options including "add folder", "merge folders" and "delete folder". New folders may be added at any level of the hierarchy of categories. For example, a subfolder of "guitar" might be created within the "string" folder; and a new folder of "children's song" might be generated under either "singing" or "chorus". The user may also choose any two folders and merge them into one. A new name needs to be given to the merged folder. Furthermore, an empty folder (together with all its subfolders if it has any and everyone of them must be empty) might be selected to get deleted.

- Move music items. The user could always adjust results from the automated procedure by moving misclassified music items to the right group. One example is illustrated in Fig.9 in which the user clicks on the music file titled “Michael Jackson – Beat It” and moves it from the current folder to the “Male solo” folder (the destination folder is indicated by the semi-solid color of its button).
- Re-classify. When a large number of new samples have been put into the database, they can be classified into existing categories by clicking the “Re-classify” button of the task bar. Only new samples will be processed in such case. In another case, the user could click on the “Tools” button to select which music classification tools to use, i.e. he can disable some of the tools and enable the rest of the tools. All available tools are listed in a pull-down menu. The user may check the tools and then click “Re-classify”. Samples in the database which are affected by this change of tool combination are reclassified, and the results will be shown on the screen accordingly. For instance, if the tool for distinguishing symphonies is deselected, the “symphony” button will disappear from the GUI, and samples formerly under the symphony folder will be reclassified to other folders.

5. DISCUSSIONS, CONCLUSIONS AND FUTURE WORK

Currently, manual classification and indexing are basically used for managing professional music databases. In order to keep the consistency of music indexing among different people, a set of criteria are normally defined. While this is a time- and labor-consuming work for organizing professional music assets; it is literally impossible for managing people’s personal music assets once their collections become large. A semi-automatic approach for music collection management was presented in this paper, which integrates existing music analysis techniques with user’s preferences to achieve an efficient and easy-to-use system for music indexing, browsing and retrieving.

The proposed system aims at achieving a good balance among automation, accuracy and flexibility. First of all, it conducts an automatic classification of music assets based on available techniques and observations. This automated grouping of music assets is accomplished by utilizing a set of tools, with each tool undertaking one step of music classification, such as distinguishing vocal solo from chorus, or detecting symphonies. The user may decide which tools to select, and thus also determine the structure of the music collection. These tools are designed to be independent from each other, so that each tool could be modified without affecting other tools and the overall structure. Also, new tools can be easily inserted into the system. For example, there have been a couple of research work on singer identification published recently [3],[6], and once there is a reliable approach available, a tool could be added to the system which identifies the singer of a vocal solo and classifies songs according to the singer's information. Furthermore, existing tools in the system could always be replaced by new tools easily when new algorithms are developed which can achieve the same classification tasks more effectively and efficiently. On the basis of the classification result, the user is provided with a number of functions with the help of a specially designed user interface, including easily browsing through music pieces under each folder and browsing across different folders, manually adjusting the hierarchy and structure of folders, as well as moving music items from one folder to another.

Improvements of the proposed system could be done in several aspects. One of them is to add a feedback mechanism. Relevance feedback has been successfully used in image retrieval systems (one example was described in [8]), which exploits user's feedback to the retrieval result to tune the retrieval criteria. The same idea may be applied to music classification as well. The system could learn from the user's operations of moving music pieces across folders to adjust classification criteria in related tools. For example, when the user moves one misclassified music sample from the "keyboard" folder to the "string" folder, the system could find a bunch of music pieces in the "keyboard" folder which are similar to the moved sample, and ask the user whether to move these samples to the "string" folder as well. Or even further, the system could adjust its criteria for differentiating the two groups of instrumental music after a relatively large number of music pieces have been moved across the classes.

In general, we believe that by integrating music content analysis techniques with human interactions, this semi-automatic music classification approach will satisfy the urgent needs of managing music collections in an effective and efficient manner which exist nowadays, to the greatest extent as permitted by today's technologies. Especially, our proposed system has the ability to evolve along with progresses in the field by adding new tools or replacing existing tools to provide new functions and more accurate classification results.

6. REFERENCES

1. Tong Zhang and C.-C. Jay Kuo, *Content-Based Audio Classification and Retrieval for Audiovisual Data Parsing*, Kluwer Academic Publishers, 2001.
2. A. L. Berenzweig and D. P. W. Ellis, "Locating singing voice segments within music signals," *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp.119-122, New York, Oct. 2001.
3. Tong Zhang, "Automatic singer identification," *Proceedings of IEEE Conference on Multimedia and Expo*, vol.1, pp.33-36, Baltimore, July 6-9, 2003.
4. L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, Inc., New Jersey, 1978.
5. Tong Zhang, "Instrument classification in polyphonic music based on timbre analysis," *Proceedings of SPIE Conference on Internet Multimedia Management Systems II*, vol.4519, pp.136-147, Denver, Aug. 2001.
6. M.A. Bartsch and G.H. Wakefield, "To catch a chorus: using chroma-based representations for audio thumbnailing," *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp.15-18, New York, Oct. 2001.
7. Y. E. Kim and B. Whitman, "Singer identification in popular music recordings using voice coding features," *Proceedings of International Symposium on Music Information Retrieval*, Paris, France, Oct. 2002.
8. Y. Rui, T. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: A powerful tool in interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology,* Special Issue on Segmentation, Description, and Retrieval of Video Content, vol.8, no.5, pp.644-655, Sept. 1998.