USABILITY EVALUATION OF VISUALIZATION INTERFACES FOR CONTENT-BASED MUSIC RETRIEVAL SYSTEMS

Keiichiro Hoashi† **Shuhei Hamawaki**‡ **Hiromi Ishizaki**† **Yasuhiro Takishima**† **Jiro Katto**‡
† KDDI R&D Laboratories Inc.
{hoashi, ishizaki, takisima}@kddilabs.jp
‡ Graduate School of Science and Engineering, Waseda University
{hamawaki, katto}@katto.comm.waseda.ac.jp

ABSTRACT

This research presents a formal user evaluation of a typical visualization method for content-based music information retrieval (MIR) systems, and also proposes a novel interface to improve MIR usability. Numerous interfaces to visualize content-based MIR systems have been proposed, but reports on user evaluations of such proposed GUIs are scarce. This research aims to evaluate the effectiveness of a typical 2-D visualization method for content-based MIR systems, by conducting comparative user evaluations against the traditional list-based format to present MIR results to the user. Based on the observations of the experimental results, we next propose a 3-D visualization system, which features a function to specify sub-regions of the feature space based on genre classification results, and a function which allows users to select features that are assigned to the axes of the 3-D space. Evaluation of this GUI conclude that the functions of the 3-D system can significantly improve both the efficiency and usability of MIR systems.

1. INTRODUCTION

The popularity of online music distribution services have provided an opportunity for users to access to millions of songs. Furthermore, the rapid spread of portable devices with large storage, *e.g.*, the iPod and 3G mobile phones, has also enabled common users to carry around music collections, which may consist of thousands of songs. These developments have prompted the need for effective music information retrieval (MIR) technologies, in order to ease the user burden to find songs which they want to listen to.

It is obvious that, for any MIR system, the usability of its interface is essential for the user to efficiently search for songs which match their preferences. However, while various GUIs for MIR systems have been proposed, reports on user evaluations of such GUIs are scarce, hence, the effectiveness and/or problems of visualizing MIR systems are yet to be formally clarified.

The objective of this research is to evaluate the effectiveness of MIR visualization, and derive what functions are necessary to improve its usability. In order to accomplish this objective, we first conduct a comparative user experiment between a typical 2-D visualization MIR interface, against the traditional list-based format. Through the analysis of this experiment, we verify if visualization actually contributes to improve the efficiency of MIR, and also derive potential problems of visualization in general. Based on the knowledge obtained from this analysis, we next propose an extended 3-D interface with several new functions, which aim to resolve the problems that have become apparent from the results of the prior experiment. Comparative experiments with the previous GUI indicate that the additional functions contribute to improve the efficiency and entertainability of MIR systems.

2. RELATED WORK

One of the initial research efforts to visualize content-based MIR is the *Islands of Music* application, developed by Pampalk [1]. This application utilizes self-organized maps to plot songs on a two-dimensional feature space, and expresses populated clusters in the feature space by illustrating "islands" of music. Furthermore, Lamere *et al.* have developed a visualization application called *Search Inside the Music* [2], which calculates audio-based similarity between songs in the music collection, and locates highly similar songs adjacently in a 3-D feature space.

Recently, there have been numerous reports to collect meta-information of songs and/or artists from the Web, and display the collected information on the user interface of MIR systems, to support the users' music searching process. *MusicRainbow* [3] is an application designed to discover artists, which maps similar artists on a circular "rainbow." The artist similarity is calculated based on acoustic analysis of the artists' songs. Labels of the artists are applied by analyzing Web information retrieved by using the name of the artists as the search query. Other applications to visualize music with web-based metadata include *MusicSun* [4], an extended application of *MusicRainbow*, as well as the work presented in [5, 6], etc.

As clear from the above descriptions of existing work,

many proposals of user interfaces for MIR systems have been made in recent years. However, thorough user evaluations of such interfaces are scarce. Therefore, it is not apparent that the various functions, user interface designs, etc. which have been proposed in previous work, actually contribute to improve the overall usability of MIR systems.

Considering these problems, we set the motivation of our research to first evaluate the effectiveness of typical visualization interfaces, by comparison with the traditional list-based format to output MIR results. Furthermore, through the analysis of user logs of the experiment, we will clarify the advantages and drawbacks of visualization, and utilize this analysis to further improve usability of MIR visualization interfaces.

3. EVALUATION OF 2-D INTERFACE

Our first experiment is to evaluate a prototype 2-D interface for content-based MIR systems, by comparison with the traditional list-based interface. For the following experiment, 16 subjects (all Japanese students in computer science) have participated to work on an experiment task using the MIR systems. Details of the systems, and the evaluation experiment are as follows.

3.1 Systems

3.1.1 Feature extraction

For both MIR systems used in the experiment, the songs in the music collection are vectorized, based on the TreeQ algorithm developed by Foote [7]. For the initial training data of TreeQ, we use the songs and sub-genre metadata of the RWC Genre Database [8]. MFCCs are extracted as the features used for the TreeQ method. In order to optimize the feature space to suit the characteristics of the experimental music collection, we re-construct the feature space based on the algorithm proposed by Hoashi *et al* [9]. The final song vectors are generated based on the re-constructed feature space.

3.1.2 2-D interface

Based on the song vectors generated by the previous process, we have developed a prototype 2-D interface for our MIR system. In order to plot the song vectors to the 2-D space, the vectors are compressed to two dimensions, by conducting principal component analysis (PCA), and extracting the first two components of the PCA results.

A screenshot of this system is illustrated in Figure 1. The 2-D interface consists of two major components: the *macro feature space viewer*, which displays the entire "universe" of the music feature space, and the *local feature space viewer*, which displays a close-up view of the area where the user is interested in. In this system, users can first select their area of interest, by clicking on the macro feature space viewer. Next, users can listen to songs in the selected area, by clicking on the plots displayed in the local feature space viewer.

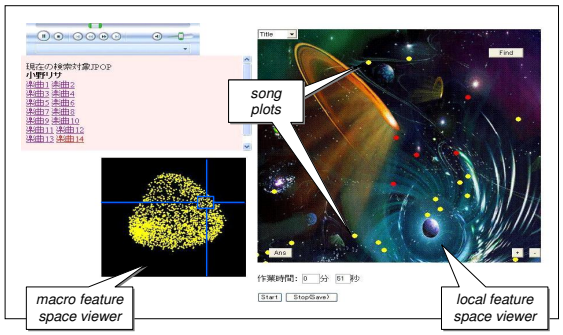

Figure 1. Screenshot of 2-D interface

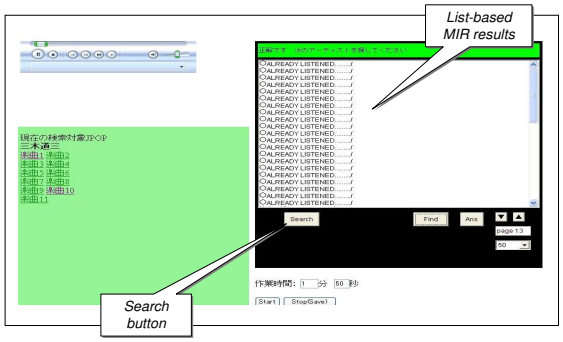

Figure 2. Screenshot of list-based interface

3.1.3 List-based interface

For comparison with the 2-D interface, we have also developed a list-based MIR system, which outputs the MIR results in a list-based format. This interface resembles the list format to present search results for typical Web search engines. A screenshot of this system is shown in Figure 2.

In this system, a random list of songs in the music collection is initially presented, with the title and artist information hidden to the user. Users are to search for the target songs, by first listening to the songs in this initial list to select a query, and clicking on the "Search" button to execute MIR. The vector similarity between the query song and all other songs are calculated, and the songs with high similarity are presented in the list, sorted accordingly to the similarity to the query song. Users can listen to the songs in the list and continue searching, by repeating the MIR procedure for the songs listed in the MIR results.

3.2 Experiment task

The task given to the subjects of the experiment is to use the two previous MIR systems, to search for songs performed by specific artists. For each experiment, a set of target songs, which are performed by a pre-specified Japanese artist, are added to the base collection, which consists of 723 Korean pop songs (hereafter referred to as *K-pop* songs). Naturally, the *K-pop* songs are unfamiliar to the subjects. The number of target songs ranges from 10 to 14 songs, depending on the target artist.

Prior to each experiment, the subjects are provided with a set of sample songs, which are performed by the target

artist, but are different from the target songs added to the base collection. Based on the impression of listening to the sample songs, the subjects then use the MIR system to search for the target songs. A single task session finishes, when the subject has successfully discovered one of the target songs. Each experiment consists of three sessions. For each session, a different artist is assigned as the target. Each subject performs this experiment, for both the 2-D and list-based systems.

3.3 Evaluation measures

Comparative evaluation of the two systems are conducted by the following objective and subjective measures.

3.3.1 Objective evaluation measures

For objective evaluation, we measure the efficiency of the MIR experiment task, based on the following two measures:

- **Operation time** *(OTime)*: Time required to complete experiment task.
- **Number of song plays** *(PlayNum)*: Number of songs played to complete task.

3.3.2 Subjective evaluation measures

For subjective evaluation, each subject is asked to apply a five-ranked rating to the systems, for the following elements (1:Bad $\sim$ 5:Good):

- **Operability** *(Oper)*: Evaluation of the usability of the interface
- **Accuracy** *(Acc)*: Evaluation of the accuracy of MIR results
- **Explicitness** *(Expl)*: Easiness to grasp relationship between songs in the MIR results
- **Enjoyability** *(Enj)*: Evaluation of overall entertainability of system

3.4 Results

3.4.1 Comparison of evaluation measure results

In order to directly compare the two systems, we summarize the evaluation results by an election-like approach. For each subject, the evaluation value for each measure is compared, and a "vote" for the subject is cast to the system with the higher score. For the objective measures (*OTime*, *PlayNum*), the vote is cast to the system with the smaller value, since the efficiency of a superior system should result in lower *OTime* and *PlayNum*.

The resulting number of votes of all evaluation measure for the two systems, are illustrated in Figure 3. Note that, the votes of the subjects who could not complete the experiment task are omitted from the results of the objective measures (*OTime*, *PlayNum*).

Comparison of the number of votes for the objective measures in Figure 3, show that the number of subjects

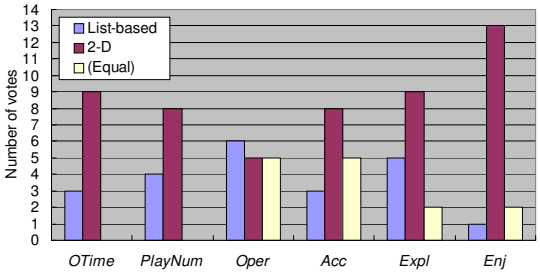

Figure 3. Number of votes for 2-D and list-based systems

who have completed the experiment task more efficiently with the 2-D interface is approximately twice as high as the list-based interface. This result indicates that the 2-D interface contributes to improve MIR efficiency. However, the vote results for *Oper* show that, subjectively, the operability of the two systems do not differ as much as the objective evaluation results suggest. A reason for this result is assumed to be that, many of the subjects are generally used to the traditional list-based interface, while the 2-D interface requires time to get acquainted with. Further analyses to discuss this result will be presented in the following section.

The results for the other subjective evaluation measures show that the 2-D interface has been more well-accepted than the list-based interface. The reason why the 2-D system has received more votes for *Expl* is obvious, since the list-based interface can only present the similarity between the query songs and other songs in the MIR results, while the 2-D interface not only displays the similarity to the query, but also the relative position between other songs that are plotted in the interface. An interesting observation is that the accuracy (*Acc*) of the 2-D system has received more votes, despite the fact that the feature extraction and vectorization methods of the systems are the same. This result indicates that the visualization of the feature space not only improves efficiency of MIR, but also applies a better impression about the accuracy of MIR results.

3.4.2 Analysis of user logs

For further analysis of the 2-D system, we analyze the user logs of the experiment. Figures 4 and 5 illustrate the *PlayNum* of all subjects, for the three target songs (hereafter referred to as *J-pop*$\{1,2,3\}$, respectively). For the first target song (*J-pop1*), *PlayNum* is generally lower for the list-based system, compared to that of the 2-D system. This indicates that the initial efficiency of MIR is higher for the list-based system, which is assumed to be the reason for the close voting results for *Oper* in Figure 3. However, as the experiment proceeds to the second and third target songs, a decreasing trend of *PlayNum* can be observed for the 2-D system, while no such trend is apparent for the list-based system. These observations show that users are able to search for music with high efficiency by using the 2-D system, as soon as they get acquainted with its interface.

Next, we analyze the search logs of 2-D system users,

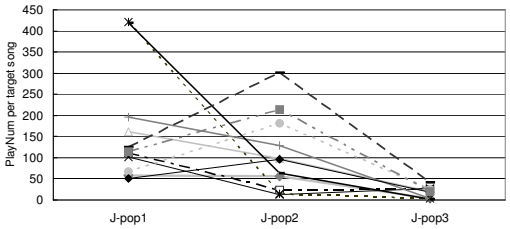

Figure 4. *PlayNum* per target song (2-D system)

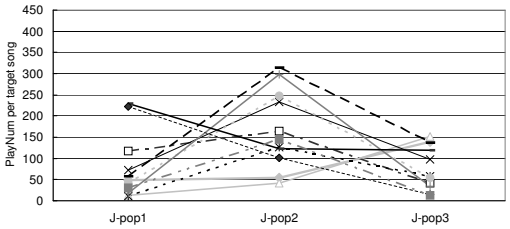

Figure 5. *PlayNum* per target song (list-based system)

to identify drawbacks of the 2-D system. Figure 6 illustrates the distribution of the *K-pop* and target song plots in the 2-D feature space, and Figure 7 shows the search path of a typical subject when using the 2-D interface. The search path for the first target song *J-pop1* shows that the subject first tries to grasp the characteristics of the feature space, by listening to songs that are plotted in various locations. In the next session, *i.e.*, the search for *J-pop2*, the log shows that the subject initially focuses on the edges of the feature space, where the acoustic features of the songs are assumed to be characteristic from the others, and gradually moves towards the center area. Finally, in the third session, it is clear that the subject is able to discover the target songs (*J-pop3*) with ease, assumably based on his/her experience from the previous sessions.

Overall, the above experimental results provide proof that the visualization interface of the prototype 2-D system contributes to improve the efficiency and usability of the MIR process. However, analysis of user logs also show that a major problem of the visualization interface is that users are required to experience the system sufficiently, in order to grasp its characteristics. Another interpretation of this result may be that, many users have experienced difficulty to comprehend the features, *i.e.*, the timbral features of the MFCC-based TreeQ vectors, that are used for the visualization of the songs in the MIR system.

4. 3-D MIR SYSTEM INTERFACE

In order to resolve the problems that have become apparent from the previous experiment, we next develop an extended visualization interface for content-based MIR. This system features a 3-D visualization interface with the following additional functions: (1) Selection of sub feature spaces based on genre classification, and (2) User selection of features which define the axes in the 3-D feature space. A screenshot of the 3-D system is illustrated in Figure 8. The objective of the first function is to provide an intuitional way to choose specific areas within the visualization feature space, in order to reduce the time required to grasp its characteristics. The second function aims to improve the understandability of the visualization interface, by allowing the users to select features which they prefer to use for MIR. Details of the system are presented in the following sections.

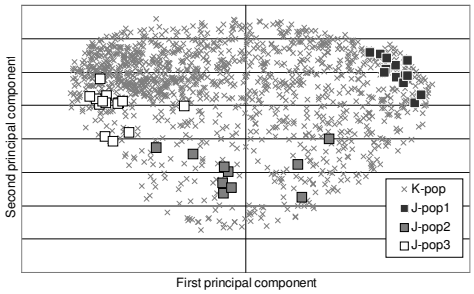

Figure 6. Song plots in the 2-D interface

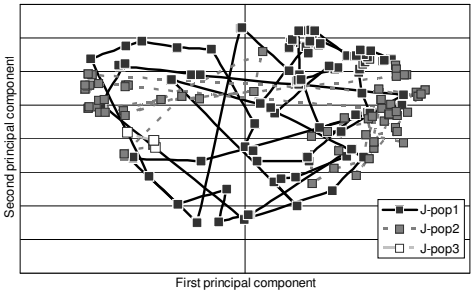

Figure 7. Search path of typical user of 2-D interface

4.1 Selecting subspaces based on genre classification

A major problem of visualization, as clarified from the previous experiment, is the experience time required for the user to get acquainted to the visualization interface. In order to reduce this acquaintance time, we have added a function to the 3-D system, which enables the user to select "sub-spaces" in the macro feature space viewer, by utilizing genre classification results.

The sub-spaces are generated by conducting k-means clustering on all vectors of the songs included in the music collection, plus the song vectors of the RWC Genre Database [8], which were used as the initial training data to generate the tree-based vector quantizer. Next, labels are applied to the clusters, by referring to the sub-genre metadata of the RWC songs that are classified to each of the clusters. The number of clusters k was set to $k = 7$, based on preliminary experiments. The labels applied to the clusters are listed in Table 1.

Each sub-space is represented as a sphere in the macro feature space viewer of the 3-D system. The labels for each sub-space are listed below the viewer. Users can drag in the viewer to rotate the feature space, and click on the sphere which best represents their area of interest. This

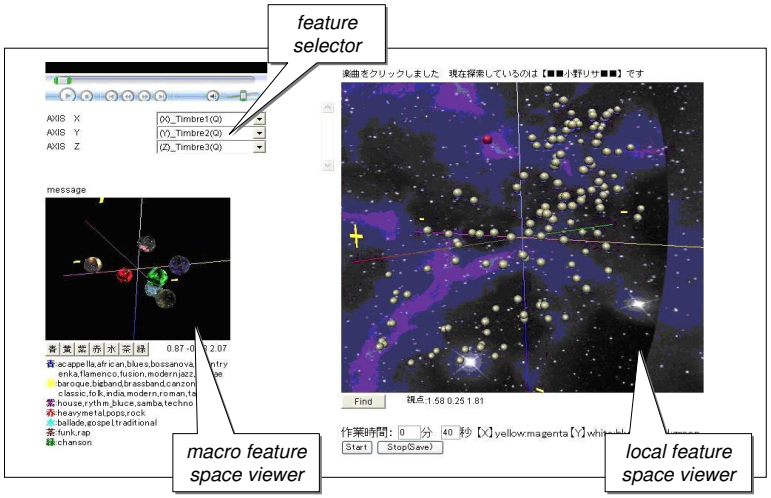

Figure 8. Screenshot of 3-D interface

Cluster ID	Labels
C_1	blues, bossanova, enka, flamenco, reggae
C_2	baroque, bigband, classic, folk, india
C_3	house, rhythm_blues, samba, techno
C_4	heavymetal, pops, rock
C_5	ballad, gospel, traditional
C_6	funk, rap
C_7	chanson

Table 1. Sub-space labels of 3-D system

function enables users to select areas in the feature space intuitively, thus is expected to resolve the try-and-error approach necessary to get acquainted with the previous 2-D system.

4.2 Selection of features for visualization

By clicking on a sub-space sphere in the macro feature space viewer, users can view the plots of the songs which belong in the selected sub-space (cluster), on the local feature space viewer. As illustrated in Figure 8, the song plots are displayed in a 3-D feature space.

Another problem of visualization is the unfamiliarity of the features used to visualize music. However, it is also clear that the appropriateness of features differ among individual users. For example, users with musical experience may consider keys or chords as adequate features for MIR, while casual music listeners may not be able to discriminate such high-level features of music.

In order to resolve this problem, we have added the *feature selector* function, which allows the user to define the features to be used in the 3-D feature space. By using the feature selector, system users can select a feature which corresponds to the x, y, z axes in the feature space. The list of features from which users can select from, are written in Table 2, along with the tools used to extract the features: TreeQ [10], MIRtoolbox [11], jAudio [12], and COFViewer [13]. Note that the "Timbre" features are equivalent to the PCA results of the TreeQ vectors utilized in the previous 2-D system.

Feature	Description	Tool
Timbre(1,2,3)	PCA results (first 3 components) of TreeQ vectors	TreeQ
Man-Woman	Categorization score of Male/Female vocal TreeQ classifier	TreeQ
Tempo	Average tempo	MIRtoolbox
Dynamics	Overall power of song	jAudio
Key	Basic key of song	COFViewer
TransKey	Transition ratio of keys	COFViewer
Mode	Degree of major/minor	MIRtoolbox
Stat(1,2,3)	PCA results (first 3 components) of jAudio features	jAudio

Table 2. List of features in 3-D system

5. EVALUATION OF 3-D INTERFACE

We have conducted a user experiment to evaluate the 3-D system. The objective of this experiment is to examine whether the problems of the 2-D system have been resolved by the additional features of the 3-D system. The task of the experiment, and the subjects are the same as the experiment described in Section 3. The measures used for evaluation are also the same as the previous experiment.

5.1 Results and discussions

First, we compare the usability of the 2-D and 3-D interfaces, by counting the votes of the subjects for all evaluation measures, similar to the evaluation in Section 3.4.1. Figure 9 illustrates the number of votes for the 2-D and 3-D systems.

The voting results for the objective measures, *OTime* and *PlayNum*, indicate that the 3-D system has contributed to improve the efficiency of MIR. Furthermore, the votes for the subjective measures show that the subjects have considered the 3-D system to be superior than the 2-D system, for all evaluation measures.

As mentioned in Section 3.4.2, a problem of the 2-D system is the time required to get used to the interface. In order to examine if the 3-D system has resolved this

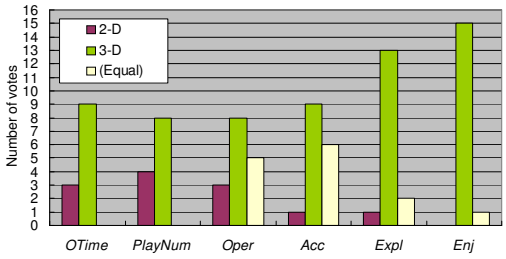

Figure 9. Number of votes for 2-D and 3-D systems

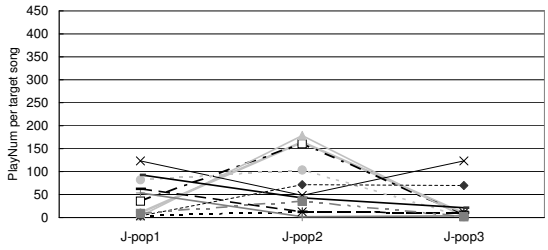

Figure 10. *PlayNum* per target song (3-D system)

problem, we analyze the *PlayNum* of the 3-D system for the three target songs, similar to the analysis presented in Figure 4. This result is illustrated in Figure 10.

Comparison of the results in Figures 4 and 10 show that, the *PlayNum* of the first target song *J-pop1* is lower for the 3-D system. This result shows that users have required less time to get acquainted with the interface of the 3-D system.

Next, in order to evaluate the usability of the feature selection function, we asked the subjects to select which features they considered useful for the experiment task, after the experiment (subjects are allowed to select multiple features in this questionnaire). The number of selections for each feature is listed in Table 3.

The results in Table 3 show that system users have selected features other than the timbre features that were used in the 2-D system. Furthermore, in this questionnaire, 11 of the 16 subjects have selected more than one feature as useful. These results indicate that users have proactively selected various features during the task, meaning that users have favorably accepted the feature selection function.

6. CONCLUSION

In this research, we have evaluated the effectiveness of visualization methods for content-based MIR, by conducting comparative user experiments of various MIR interfaces. The first experiment, which is based on a GUI with a typical 2-D visualization approach, has proved that visualization contributes to improve overall usability compared to the list-based interface, but also indicate problems for inexperienced users to comprehend the characteristics of the visualized feature space. Evaluation of the extended GUI, which is added new functions to resolve the previous problems, makes clear that the additional functions further contribute to improve the usability of MIR systems. Through

Feature	No of selections
Man-Woman	14
Tempo	8
Timbre	4
Dynamics	4
Key	1
Total	31

Table 3. Number of selections per feature in 3-D system

this research, we consider ourselves to have contributed to clarify general user requirements for the visualization of MIR systems, and also have proposed a successful approach to satisfy such needs.

7. REFERENCES

[1] E. Pampalk: Islands of Music: Analysis, Organization and Visualization of Music Archives, Master's thesis, Vienna University of Technology, 2001.

[2] P. Lamere, D. Eck: Using 3D Visualizations to Explore and Discover Music, Proc. of ISMIR 2007, 2007.

[3] E. Pampalk, M. Goto: MusicRainbow: A New User Interface to Discover Artists Using Audio-based Similarity and Web-based Labeling, Proc. of ISMIR 2006, 2006.

[4] E. Pampalk, M. Goto: MusicSun: A New Approach to Artist Recommendation, Proc. of ISMIR 2007, 2007.

[5] P. Knees, M. Schedl, T. Pohle, G. Widmer: An Innovative Three-dimensional User Interface for Exploring Music Collections Enriched with Meta-information from the Web, Proc. of ACM Multimedia 2006, 2006.

[6] M. Torrens, P. Hertzog, J.-L. Arcos: Visualizing and Exploring Personal Music Libraries, Proc. of ISMIR 2004, 2004.

[7] J. Foote: Content-based Retrieval of Music and Audio, Proc. of SPIE, Vol. 3229, pp. 138-147, 1997.

[8] M. Goto, H. Hashiguchi, T. Nishimura, R. Oka: RWC Music Database: Music Genre Database and Musical Instrument Sound Database, Proc. of ISMIR 2003, 2003.

[9] K. Hoashi, K. Matsumoto, F. Sugaya, H. Ishizaki, J. Katto: Feature Space Modification for Content-based Music Retrieval based on User Preferences, Proc. of ICASSP 2006, Vol. V, pp. 517-520, 2006.

[10] "TreeQ software," http://treeq.sourceforge.net/

[11] O. Lartillot: MIRtoolbox, http://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox

[12] D. McEnnis, C. McKay, I. Fujinaga, P. Depalle: jAudio: A Feature Extraction Library, Proc. of ISMIR 2005, 2005.

[13] T. Inoshita, J. Katto: Key Estimation Using Circle of Fifth, Proc. of MMM 2009, 2009.