Music Recommendation System Using Emotion Triggering Low-level Features

Kyoungro Yoon, *Senior Member*, IEEE, Jonghyung Lee, and Min-Uk Kim

**Abstract — ***Recently, many researches of modeling or measuring human feeling have been conducted to understand human emotions. However, researches on music-related human emotions have much difficulty due to the subjective perception of emotions. We selected low-level musical features which may trigger human emotions, based on TV music program's audience rating information and the corresponding music. In this program, audience was requested to rate music of the contestants and to select their preferred music based on their emotional feelings. In addition, we implemented personalized music recommendation system using selected features, context information and listening history. In the experimental results, we confirmed that selected features can be reliable features when these features are used in music recommendation systems.*[1]

Index Terms — Musical emotion, Low-level feature selection, Emotion triggering low-level feature, Personalized music recommendation system.

I. INTRODUCTION

Sound or music is a longitudinal wave phenomenon, where the pressure changes, caused by vibration of the sound source, propagate through media such as air or water. Music has been with human being in various fields serving religious, political, and/or social purposes, to name a few. National anthem can represent value and tradition of a country, school song or hymn has symbolic value for people who belong to the school or religion. A certain song may convey social atmosphere or emotions of past to people of today. Let alone such a symbolic meaning or special purpose, we encounter music anywhere in everyday life. When we are home, we cannot avoid music if you turn on a TV or radio, even though it is not a music program. When we step out of home, the situation is not much different. We hear music in restaurants, shopping malls, and even in public transportations. Music is everywhere in our everyday life.

One of the reasons why music is everywhere in our everyday life is that one of the fundamental functionality of music is triggering emotion. Most of the people empirically know that music triggers various emotions such as happiness, sadness, and joyfulness, to name a few [1]. Recent use of mood metadata still has troubles providing emotion-based analysis of music, because of the following reasons among others: 1. Music is very cultural; 2. Emotional perception of music depends on the personal exposure to specific cultural associations; 3. Recognition of emotion is a subjective matter [2] [3].

The advances in compression technology and the proliferative dissemination of various multimedia contents require evolutions in media services, even though traditional on-line music services provide search and recommendation services based on the textual metadata such as title, composer, musician, publication date, and genre. The needs of new types of services drive recent researches on personalized and/or emotion-based recommendation services [4][5].

We propose preference analysis method based on empirical evaluation scores for the selection of general and reliable low-level features of music that triggers emotions. For the experiments, songs performed at a weekly Korean TV music program and corresponding evaluation results of 12,000 audience panel are used, in hopes of reducing both influences by subjective perception of emotions and the need of psychological interpretation of the experimental result.

We also implemented personalized music recommendation system based on the low-level features selected by the proposed method. This recommendation system eliminates scalability problem of tag-based music recommendation systems, by employing automatically extracted low-level features of music which trigger emotions.

The organization of this paper is as follows: Section 2 explains how we selected low-level feature of music that triggers emotions. Section 3 provides the design of personalized music recommendation system based on the selected low-level features and the personal listening history combined with environmental information. Section 4 provides subjective evaluation results of several recommendation schemes to show subjective validity of the proposed recommendation system, and accuracy of the prediction of the audience panel's evaluation from the TV music program based on the selected low-level features to show objective validity of the selected low-level features. Section 5 draws conclusions based on the experimental results and provides future works.

II. EMOTION TRIGGERING LOW-LEVEL FEATURE SELECTION

In this section, we propose emotion triggering low-level feature selection method based on large number of subjective audience ratings.

[1] K. Yoon is with Department of Computer Science and Engineering, Konkuk University, Seoul 143-701, Korea. (e-mail: yoonk@konkuk.ac.kr).

J. Lee is with Department of Computer Science and Engineering, Konkuk University, Seoul 143-701, Korea. (e-mail: leejonghyung@hanmail.net).

M.-U. Kim is with Department of Computer Science and Engineering, Konkuk University, Seoul 143-701, Korea. (e-mail: kimminuk@gmail.com).

Contributed Paper
Manuscript received 03/23/12
Current version published 06/22/12
Electronic version published 06/22/12.

A. Database

We use the songs and evaluation results from a series of TV music competition program for emotion triggering low-level feature selection. At each round of competition which is held twice in every three weeks, each of seven singers sings a song of his/her choice. Until the time when we conduct these experiments, there were twenty-four rounds of competition with one hundred-sixty-eight songs sung. The audience panel consisting of five age groups of teens, twenties, thirties, forties, and fifties, each of which has fifty females and fifty males, are carefully selected not to have any bias for specific genre of song or for specific singer, at each round of competition. Therefore, 12,000 people participated in this experiment. At the end of each competition, the audience panel is requested to rate the performance by selecting the best three out of the seven performances, which impressed him/her the most.

For the experiments, we randomly selected two-thirds of the competitions, i.e., 16 competitions out of 24 competitions, for the training, and the rest one-thirds of the competitions, i.e., 8 competitions, for the test. By selecting competitions for the training randomly, we can reduce the temporal correlations of competitions and draw more generalized emotion-triggering low-level features. Every audio clip is normalized in volume and re-sampled at 16,000Hz, 16 bit depth, and mono channel.

B. Feature Extraction

Characteristics or features of sound source can be classified into three layers as follows: Low-level acoustic feature layer; Mid-level audio signature layer which can be used to separate distinct signal objects; High-level semantic layer which can be used to classify different class of sound sources. In our experiment, six high-level features such as rhythm, tonal, timbre, dynamics, fluctuation, and spectral, twenty-eight mid-level features such as RMS, peak, centroid, tempo, zero-cross, etc., and eight hundred-ninety low-level feature values are extracted from the songs in the database. Fig. 1 shows high-level and mid-level feature used in our experiments.

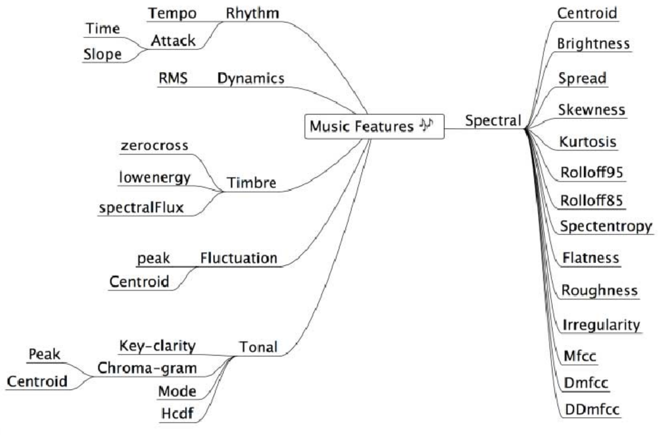


Fig. 1. High-level and mid-level music features

C. Feature Selection

Feature selection is a task to select the minimum number of features needed to represent the data in the intended way among the candidate features. In this paper, we focus on the audience rating information rather than data itself during feature selection, as the audience rating given by five hundred people for each competition is believed to be a measure of the emotional impression.

There are many feature selection methods such as: All possible regression which is a procedure considering all the possible independent variables; Forward selection which is a procedure adding most correlated variables rather than least; Backward elimination which is a procedure subtracting least correlated variables; Stepwise selection which is a combined procedure of forward selection and backward elimination [6].

In this experiment, we used the backward elimination method, which is implemented as the mirmap function in the MIRtoolbox [7]. The mirmap is a function implementing backward elimination method where the rating information is set as the dependent variable and feature values are set as the independent variables. The mirmap function first performs the Lilliefors test for all the features (independent variables) at five percent significance level to detect any feature that does not have sufficiently normal distribution [8]. Those features which are not sufficiently normal are then normalized using an optimization algorithm, so that their distributions become sufficiently Gaussian to apply correlation estimation. Then, the normalized features, which are still not sufficiently normal at one percent significance level, are excluded from any further consideration. Finally, the features are selected if their correlation with the ratings are statistically significant (with a p-value lower than 0.05) are selected by applying correlation estimation. The p-value represents the statistical significance. If the p-value is less than the significance level, the null-hypothesis is rejected and the result is said to be statistically significant [9].

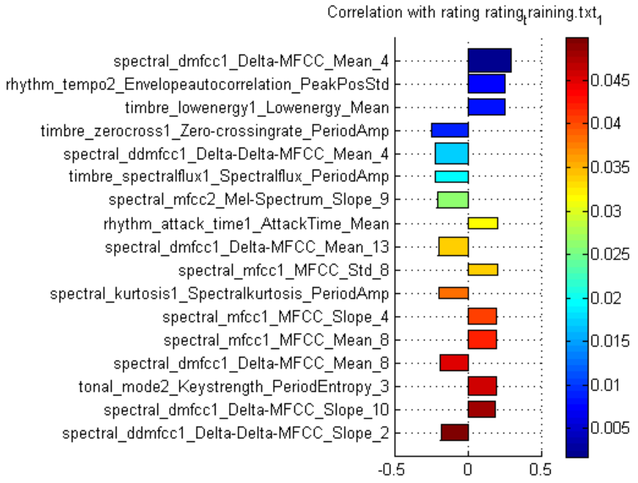


Fig. 2. Selected low-level features using backward elimination method from training set of competition.

Fig. 2 shows seventeen selected features from the training data set using the proposed method. In this experiment, any feature whose cross-correlation with higher correlation valued feature is larger than 0.6 is removed to keep only the independent features. The features are sorted in the decreasing order of correlation from top to bottom of the vertical axis. The length of the colored horizontal bars represents the magnitude of the correlation value. The color of the bar shows p-value, i.e., the significance of the corresponding feature. The correlation score of two variables ***x*** and ***y*** can be acquired by (1) where ***cov( )*** represents covariance, represents mean, and represents standard deviation.

$$\rho_{x,y} = corr(x,y) = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y} \qquad (1)$$

In this paper, it is assumed that the features selected through the proposed process based on the audience panel of various age groups are the general and reliable low-level features stimulating human emotions, as the large number of audience panel removes the possibility that the evaluation of each competition may reflect a certain personal preference of a singer or genre.

III. Personalized Music Recommendation System

In this section, we present a personalized music recommendation system, implemented using the proposed low-level features. This recommendation system eliminates scalability problem of tag-based music recommendation systems, by employing automatically extracted low-level features of music which trigger emotions. By analyzing user's listening history, we tried to reduce the semantic gap between low-level features and high level semantic classification information and to effectively reflect dynamic changes of user's behavior of selecting songs depending on listening environments. Fig. 3 shows overall architecture of the proposed music recommendation system. The details of the architecture are given in the following sub-sections.

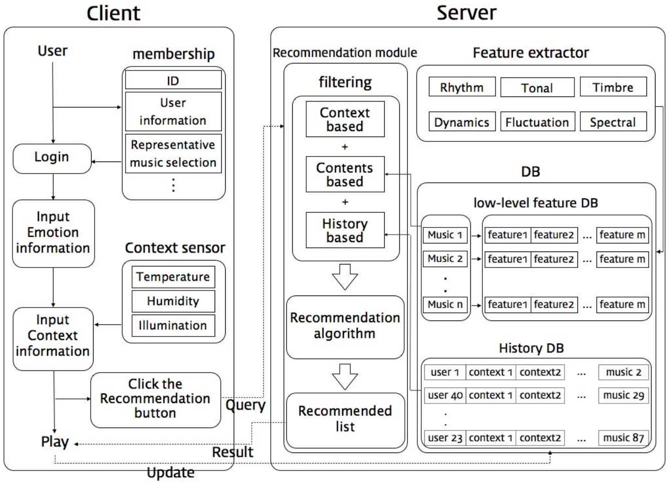


Fig. 3. Architecture of personalized music recommendation system

A. Client

As shown in the box named client in Fig. 3, users first register himself/herself on the recommendation system by providing couple of personal information and a representative song for each mood class to avoid the cold-start problem. Four representative mood classes of angry, happy, sad, and peaceful are given from the extended Thayer's mood model [10]. The extended Thayer's mood model is a frequently referenced mood model, which is enhanced from the original Thayer's mood model to simplify the representative moods, in music retrieval researches as shown in Fig. 4.

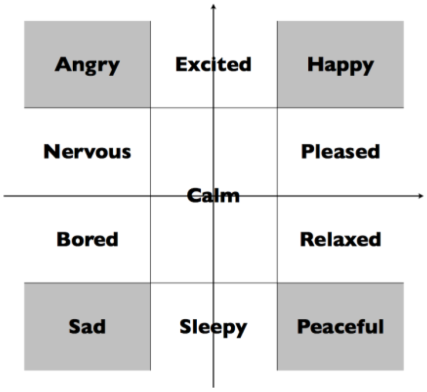
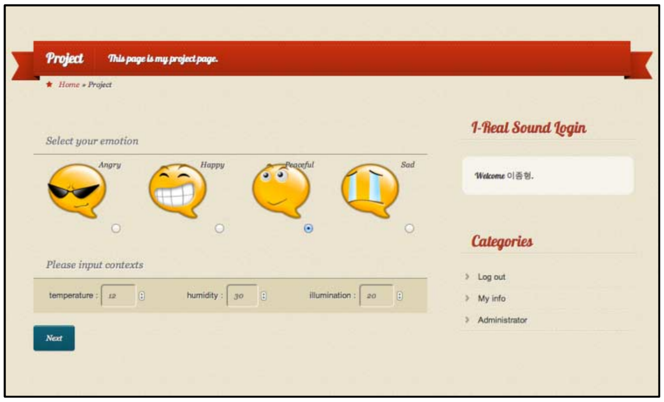


Fig. 4. Extended Thayer's mood model.

Once the user is registered, he/she provides emotion information by selecting one of those four moods through a user interface, such as the one shown in Fig. 5, whenever he/she logs in. After providing user's emotion information, the context information of temperature, humidity, and illumination is provided from the sensor. Then the system is ready to recommend and play. The user can either select a song from the playlist and listen, or get a recommendation list from the server by clicking on the recommendation button. When a song is selected and played, the emotion and context information with the played song information are sent to the server and saved in the history database.



Select your emotion

Angry ○ Happy ○ Peaceful ● Sad ○

Welcome 이종형.

Please input contexts

temperature : 12 humidity : 30 illumination : 20

Fig. 5. Client UI of personalized music recommendation system

B. Server

The server of the personalized music recommendation system has three main modules of recommendation, feature extraction, and database as shown in Fig. 3.

The database is composed of a music information database, low-level feature database, and history database. The music information database stores basic metadata for individual music such as identifier, title, singer, album title, and file location. The low-level feature database stores low-level features from each song, extracted through the method proposed in section 2. The history database stores individual user's listening history of songs with corresponding emotion and context information, every time the user selects and plays a song.

The recommendation module creates recommendation list when it is requested. Fig. 6 shows the flow of the recommendation module.

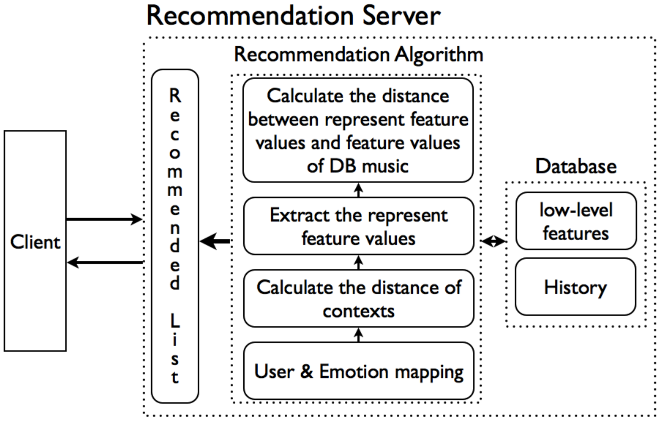

Recommendation Server
Recommendation Algorithm
Client
Recommended List
Calculate the distance between represent feature values and feature values of DB music
Extract the represent feature values
Calculate the distance of contexts
User & Emotion mapping
Database
low-level features
History

Fig. 6. Flow diagram of the recommendation module

The recommendation module retrieves listening history, low-level features, and the context information from the database and creates recommended list of songs to be sent to the client. In detail, the recommendation module retrieves the low-level feature values of the first ranked song which is played at the most similar situation based on the Euclidian distance from the current emotion, temperature, humidity, and illumination as the representative low-level feature values. If there are more than one song with the same nearest distance, averaged low-level feature values are selected as the representative low-level feature values. Therefore, the recommendation module tries to find songs with emotional influences similar to the songs that the user usually had selected when he/she was in the similar situation in the past, regardless of whether the selected songs reinforces or changes the current emotional state, under the assumption that the combination of the selected low-level features can reflect the characteristics of emotional influences of the selected song.

With the representative low-level feature values, the recommendation list of the songs are created in the increasing order of Euclidian distance between the representative low-level feature values and the corresponding feature values of the songs in the database. Once again, these feature values are the values of the selected low-level features selected as the emotion triggering features in section 2.

The recommendation system is implemented using Java language with 400 songs of mixture of K-pops, western popular songs and classic music. The user interface is provided by HTML5 for convenient access through web browsers.

IV. Experimental Results

The low-level features that are selected by the proposed method depend on the training set of songs and evaluation results. To check if there is any peculiar competition in the 16 competitions of training set, in which the evaluation of the competition may have been influenced by the specific set of songs performed or by the specific audience panel with special preference, we also constructed 16 sets of training sets by selecting 15 competitions out of 16 candidate competitions, in addition to the training set with all of the 16 candidate competitions. From the 16 training sets composed of 15 competitions, 10 to 19 features are selected from individual training set with a total of 40 distinct features. From the training set with all the 16 competitions, 17 features are selected. Out of the 17 features extracted from the training set of 16 competitions, all the 17 features appear at least in one of the 16 training sets of 15 competitions, 16 features appear in more than four of 16 sets of 15 competitions, and 11 features appear in more than eight of 16 sets of 15 competitions. Out of 40 features extracted from the 16 training sets of 15 competitions, 20 features appeared in more than 4 training sets, and 13 features appeared in more than 8 training sets, as shown in Table I.

Based on this fact, we composed four modules as follows: Module 1 use the 17 features extracted from the 16 competition training set; Module 2 use the union of 40 features extracted from any one of the 15 competition training sets; Module 3 use the 20 features from the 40 feature union set, which appear in at least four training sets of 15 competition training sets; Module 4 use the 13 features from the 40 feature union set, which appear in at least eight training sets of 15 competition training sets. Each module is used in two types of experiments, which are prediction of competition results based on audience panel evaluation and recommendation of user's preferring songs.

TABLE I
Four Different Modules Depending on The Selected Feature Set

Module	Description	Number of Features
1	Feature set from the training set containing all the 16 competitions	17
2	Union of all the features that appear in any one of the 16 training sets with 15 competitions.	40
3	Union of features that appear in at least four of the 16 training sets with 15 competitions.	20
4	Union of features that appear in at least eight of the 16 training sets with 15 competitions.	13

Finally, for the best performance module, we performed an additional experiment of selecting the best-performing p-value by changing p-value from 0.01 to 0.1 with difference of 0.01 for the prediction module.

A. *Prediction of Competition Results*

In first type of experiments, we tried to predict the competition results of 8 competitions not used in the training using four different sets of features. In this prediction system, weight of each feature is given by the correlation value acquired by the regression analysis.

Regression analysis is a statistical model of predicting dependent variable based on the combinations of given independent variables. Multiple linear regression analysis model is given in (2), where Y_i represents the dependent variable, which is the predicted ranking of a song, of *i*th observation, X_{ki} represents *k*th feature value of *i*th observation, and β_k represents weight of the feature X_{ki}.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i, i = 1,2,\cdots,n \quad (2)$$

Fig. 7 shows the result of the prediction. Along the x-axis is the competition number of 8 competitions and along the y-axis is the average difference between the predicted ranking and the actual ranking given by the audience panel.

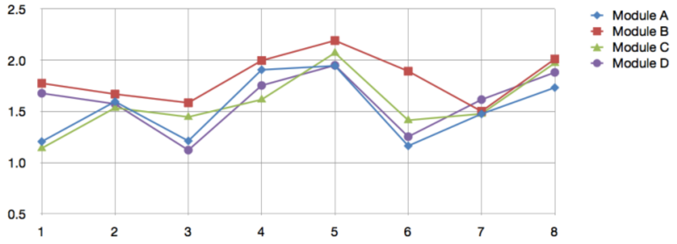


Fig. 7 Results of the prediction

Table II shows the average difference between the prediction and the actual ranking given by the audience panel of the eight competitions, depending on the features selected as given in Table I.

TABLE II
MEAN ERROR OF THE FOUR DIFFERENT MODULES

Module	Mean Error
1 (17 Features)	1.52
2 (40 Features)	1.82
3 (20 Features)	1.58
4 (13 Features)	1.60

As seen in this experimental result, the prediction module using selected features from the training set of 16 competitions gives better results than the modules using features selected by combining features from the 16 training sets of 15 competitions.

B. *User evaluation of the proposed recommendation systems*

We performed the user evaluation on the recommendation systems based on the four different modules used in the previous experiments. For this evaluation experiments, 30 users used the given recommendation system for a certain period of time and gave their evaluation on the four different modules of recommendation system. There are 400 songs in the database of the recommendation system and each user listened average of 96 songs using the recommendation system.

The users are requested to give evaluation of each module by selecting one of four criteria of very good, good, not bad, and bad. The result is given in Fig. 8.

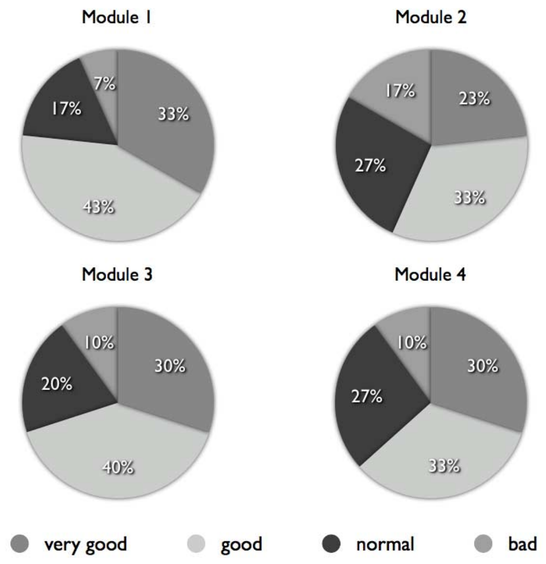


Fig. 8 User evaluation result of each module

As shown in Fig. 8, positive results of very good and good are given in the order of module 1, module 3, module 2, and module 4. The highest negative evaluation score is given to module 2, and the lowest negative evaluation score is given to module 1. In short, the recommendation system based on the features extracted from the training set of all the 16 competitions performs the best based on the subjective user evaluation.

C. *Finding the optimum p-value*

Through the previous two experiments, we have shown that using the features extracted from the larger training set containing all the available competition results performs better than using union set of features extracted from multiple smaller training sets.

This last experiment is to find the optimal p-value for the usage of module 1, which has shown the best performance. Changing the p-value starting from 0.01 to 0.1 by 0.01, we have performed extraction of features and prediction of ranking based on the extracted features ten times, as the number of selected features changes as the p-value changes.

Fig. 9 shows the results of this experiment. Along the x-axis are p-values and along the y-axis are the cumulative errors of prediction, which is the difference between the predicted ranking and the actual ranking given by the audience panel for eight competitions used in the test set.

As we can see from Fig. 9, the best prediction is done with p-value of 0.05. As the p-value gets closer to 0.05, the performance of prediction gets better. Actually, p-value of 0.05 is the value recommended most in statistical experiments [11].

V. Conclusion

In this paper, we extracted low-level features using 112 songs performed by 23 singers with evaluation results of 12,000 audience panel from a weekly Korean TV music program. Using these extracted features, prediction of the evaluation results by audience panel is performed to check the validity and the effectiveness of the selected features. Using these extracted features, we also build a personal recommendation system.

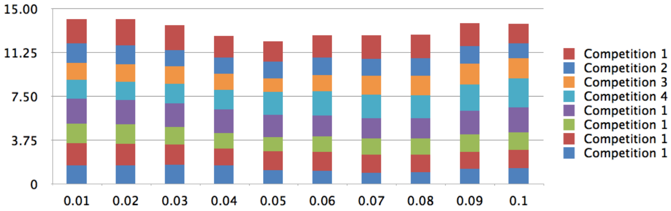


Fig. 9 Prediction results with features extracted using various p-values

As the actual ranking of each round of competition is given by 500 people of various age groups and sex, we can say that the ranking is not affected by any personal preference on either genre or singer of the songs. We cannot say that the evaluation of the audience panel members is not affected by the visual aspect of the performance. However, it is also a fact that large part of the evaluation is affected by the aural aspect of the performance, and this experiment is an effort to find correlations between the low-level features representing the aural aspect of the performance and the scoring of the evaluation. It is also very hard to clearly show the relationship between the scoring and the level of emotion triggering. What kind of emotion each song triggers is an issue specific to personal experience and culture, but it is also very well known fact that good music triggers emotions, even though the type of emotion may differ from person to person, Also, it is known that an evaluation of experience is affected by the mood and the music affects the mood of the evaluator [12]. Therefore, we believe that the ranking or the evaluation results given by the audience panel is very likely to be a measure of how touching the song is. From the experiment of predicting the evaluation results of audience panel, the average difference of 1.5 between the predicted rankings and the actual rankings from the audience panel shows that the selected features can universally represent the emotion-triggering characteristics of the songs.

The user study on the personal recommendation system also implies that the system using these selected features, extracted from the training set of all 16 competitions, shows satisfactory recommendation results. A prototype of intelligent music player employing the proposed personal recommendation system is being built as shown in Fig. 10.

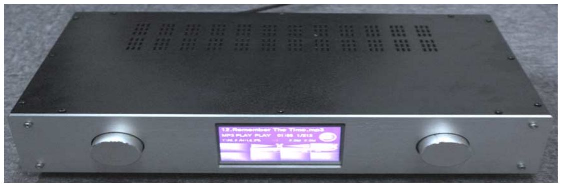


Fig. 10. Prototype of intelligent music player

Finally, we have shown through the experiments that the best prediction and recommendation results are given when the features selected with the p-value of 0.05 are used.

It is very well known that the mechanism of human emotion is very complex. Just like there are various kinds of music, people have various preferences on music. Even within a single genre of music, for example within rock and roll or classical music, the preference on music can differ from person to person. Sometimes, a personal preference of a same person may also change depending on situation or emotional states.

We notice that the features extracted by the proposed method may not be the best or complete features for representing aural features triggering human emotions. However, by increasing the size of the training set and by getting help from various fields of studies such as psychology, music and emotions, we believe that more accurate and general set of features shall be extracted.

As the experimental results shows that the features extracted from the training set of 16 competitions performs better than any of the training set of 15 competitions, we also plan to increase the training set hoping to find a saturation point of feature selection.

References

[1] M. Zentner, D. Grandjean, and K. Scherer, “Emotions evoked by the sound of music: Characterization, classification, and measurement,” *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[2] X. Hu and J. S. Downie, "Exploring mood metadata: relationships with genre, artist and usage metadata," *Proc. ISMIR*, pp. 67-72, Sept. 2007.

[3] D. J. Levitin, *This is your brain on music: the science of a human obsession*, Plume/Penguin: New York, 2007, pp. 35-38

[4] X. Zhu, Y.-Y. Shi, H.-G. Kim and K.-W. Eom, “An integrated music recommendation system,*” IEEE Trans. on Consumer Electronics,* vol. 52, no. 3, pp. 917-925, Nov. 2006.

[5] F. Kuo, M. Chiang, M. Shan, and S. Lee, “Emotion-based music recommendation by association discovery from film music,” *Proc. ACM Multimedia*, pp. 507–510, Nov. 2005.

[6] Jaehyung Ahn, *Anyone can do statistical analysis using the R*, Hannalae: Seoul Korea, 2011 (in Korean).

[7] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, “Multi-feature modeling of pulse clarity: design, validation, and optimization,” *Proc. ISMIR*, pp. 521–526, Sept. 2008.

[8] H. Lilliefors, “On the kolmogorov-smirnov test for normality with mean and variance unknown,” *Journal of the American Statistical Association*, vol. 62, pp. 399–402, 1967.

[9] O. Lartillot, *MIRtoolbox User’s Manual*.

[10] R. E. Thayer, *The Biopsychology of Mood and Arousal*, Oxford University Press: USA, 1989.

[11] D. H. Johnson, “The insignificance of statistical significance testing,” *Journal of Wildlife Management*, vol. 63, pp. 763-772, 1999.

[12] M.A. Cameron, J. Baker, M. Peterson, and K. Braunsberger, "The effect of music, wait-length evaluation, and mood on a low-cost wait experience," *Journal of Business Research*, vol. 56, issue 6, pp. 421-430, June, 2003.

BIOGRAPHIES

Kyoungro Yoon (M'00-SM'11) received a B.S. degree in Computer and Electronic Engineering from Yonsei University, Seoul, Korea in 1987, a M.S.E. degree in Electrical Engineering/Systems from University of Michigan, Ann Arbor in 1989, and a Ph.D. degree in Computer and Information Science from Syracuse University in 1999. He was a principal researcher and a group leader in Mobile Multimedia Research Lab, LG Electronics Institute of Technology from 1999 to 2003. He joined the school of Computer Science and Engineering in 2003 as an assistant professor and is an associate professor now. He served as a chair of Ad Hoc Groups on User Preferences and MPEG Query Format of ISO/IEC JTC1 SC29 WG11 (a.k.a. MPEG) and is currently serving as a chair of Metadata Subgroup and JPSearch Ad Hoc Group of ISO/IEC JTC1 SC29 WG1 (a.k.a. JPEG). He is also an editor of various international standards such as ISO IS 15938-12, 23005-2, 23005-5, 23005-6, 24800-3, 24800-5, and 24800-6.

Jonghyung Lee received M.S degree from Konkuk University, Seoul, Korea. His research interests are music information retrieval and music signal processing and system implementation.

Min-Uk Kim received a B.S. and M.S. in computer science and engineering from Konkuk University in 2007 and 2009, respectively. Since 2010, he is a Ph.D. candidate of Konkuk University. His research interests include multimedia retrieval, high-dimensional indexing, and their applications.