EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Music-aided affective interaction between human and service robot

Jeong-Sik Park[1], Gil-Jin Jang[2] and Yong-Ho Seo[3*]

Abstract

This study proposes a music-aided framework for affective interaction of service robots with humans. The framework consists of three systems, respectively, for perception, memory, and expression on the basis of the human brain mechanism. We propose a novel approach to identify human emotions in the perception system. The conventional approaches use speech and facial expressions as representative bimodal indicators for emotion recognition. But, our approach uses the mood of music as a supplementary indicator to more correctly determine emotions along with speech and facial expressions. For multimodal emotion recognition, we propose an effective decision criterion using records of bimodal recognition results relevant to the musical mood. The memory and expression systems also utilize musical data to provide natural and affective reactions to human emotions. For evaluation of our approach, we simulated the proposed human-robot interaction with a service robot, iRobiQ. Our perception system exhibited superior performance over the conventional approach, and most human participants noted favorable reactions toward the music-aided affective interaction.

1. Introduction

Service robots operate autonomously to provide useful services for humans. Unlike industrial robots, service robots interact with a large number of users in a variety of places from hospitals to home. As design and implementation breakthroughs in the field of service, robotics follow one another rapidly, people are beginning to take a great interest in these robots. An immense variety of service robots are being developed to perform human tasks such as educating children and assisting elderly people. In order to coexist in humans' daily life and offer services in accordance with a user's intention, service robots should be able to affectively interact and communicate with humans.

Affective interaction provides robots with human-like capabilities for comprehending the emotional states of users and interacting with them accordingly. For example, if a robot detects a negative user emotion, it might encourage or console the user by playing digital music or synthesized speech and by performing controlled movements. Accordingly, the primary task for affective interaction is to provide the robot with the capacity to automatically recognize emotional states from human emotional information and produce affective reactions relevant to user emotions.

Human emotional information can be obtained from various indicators: speech, facial expressions, gestures, pulse rate, and so forth. Although many researchers have tried to create an exact definition of emotions, the general conclusion that has been drawn is that emotions are difficult to define and understand [1,2]. Because of this uncertainty in defining emotions, identifying human emotional states via a single indicator is not an easy task even for humans [3]. For this reason, researchers began to investigate multimodal information processing, which uses two or more indicators simultaneously to identify emotional states.

In the conventional approaches, speech and facial expression have successfully been combined for multimodality, since they both directly convey human emotions [4,5]. Nevertheless, these indicators have several disadvantages for service robots. First, users need to remain in front of the robots while expressing emotions through either a microphone or a camera. Once a user moves out of sight, the robot may fail to monitor the emotional states. Second, the great variability in characteristics of speech or facial expression with which humans express their emotions might deteriorate the

* Correspondence: yhseo@mokwon.ac.kr
[3]Department of Intelligent Robot Engineering, Mokwon University, Daejeon, South Korea
Full list of author information is available at the end of the article

recognition accuracy. In general, different humans rarely express their emotional states in the same way. Thus, some people who express emotions with unusual characteristics may fail to achieve satisfactory performance on standard emotion recognition systems [6].

To overcome these disadvantages of the conventional approaches, this study proposes a music-aided affective interaction technique. Music is oftentimes referred to as a language of emotion [7]. People commonly enjoy listening to music that presents certain moods in accordance with their emotions. In previous studies, researchers confirmed that music greatly influences the affective and cognitive states of users [8-10]. For this reason, we utilize the mood of music that a user is listening to, as a supplementary indicator for affective interaction. Although the musical mood conveys the emotional information of humans in an indirect manner, the variability of emotional states that humans experience while listening to music is relatively low, as compared with that of speech or facial expression. Furthermore, the music-based approach has a smaller limitation with respect to the distance between a user and a robot.

The remainder of this article is organized as follows. Section 2 reviews previous studies that are relevant to this study. Section 3 proposes a framework for affective interaction between humans and robots. Section 4 provides specific procedures of music-aided affective interaction. Section 5 explains the experimental setup and results. Finally, Section 6 presents our conclusions.

2. Previous studies on affective interaction between humans and robots

An increasing awareness of the importance of emotions leads the researchers to attempt to integrate affective computing into a variety of products such as electronic games, toys, and software agents [11]. Many researchers in robotics also have been exploring affective interaction between humans and robots in order to accomplish the intended goal of human-robot interaction.

For example, a sociable robot, ‘Kismet’, understands human intention through facial expressions and engages in infant-like interactions with human caregivers [12]. ‘AIBO’, an entertainment robot, behaves like a friendly and life-like dog that responds to either the touch or sound of humans [13]. A conversational robot called ‘Mel’ introduced a new paradigm of service robots that leads human-robot interaction by demonstrating practical knowledge [14]. A cat robot was designed to simulate emotional behavior arising from physical interactions between a human and a cat [15]. Tosa and Nakatsu [16,17] have concentrated on the technology of speech emotion recognition to develop speech-based robot interaction. Their early studies, ‘MUSE’ and ‘MIC’, were capable of recognizing human emotions from speech and expressing emotional states through computer graphics on a screen. They have consistently advanced their research directions and developed more applications.

3. Framework for affective interaction

In efforts to satisfy the requirements for affective interaction, researchers have explored and advanced various types of software functions. Accordingly, it is necessary to integrate those functions and efficiently manage systematic operations according to human intentions. The best approach for this is to organize a control architecture or a framework for affective interaction between a human and a robot.

Our research target is humanoid service robots that perform human-like operations and behaviors. Thus, we propose a new framework based on a model of the human brain structure developed by the cognitive scientist Ledoux [18]. This framework consists of three individual systems associated with one another, as demonstrated in Figure 1.

The primary function of the perception system is to obtain human emotional information from the outside world through useful indicators such as facial expression and speech. The memory system records the emotional memories of users and corresponding information in order to utilize them during the interaction with humans. Finally, the expression system executes the behavior accordingly and expresses emotions of the robot.

4. Music-aided affective interaction

In the conventional approaches to achieve affective interaction, both speech and facial expression have mostly been used as representative indicators to obtain human emotional information. Those indicators, however, have several disadvantages when operated in robots, as addressed in Section 1. In addition, most of the conventional approaches convey the robot’s emotional states in monotonous ways, using a limited number of figures or synthesized speech. Thus, users easily predict the robot’s reactions and can lose interest in affective interaction with the robot. To overcome these drawbacks, we adopt music information in the framework of affective interaction.

Music is an ideal cue for identifying the internal emotions of humans and also has strong influences on the change of human emotion. Hence, we strongly believe that music will enable robots to more naturally and emotionally interact with humans. For the music-aided affective interaction, the mood of the music is recognized in the perception system and is utilized in the determination of the user’s emotional state.

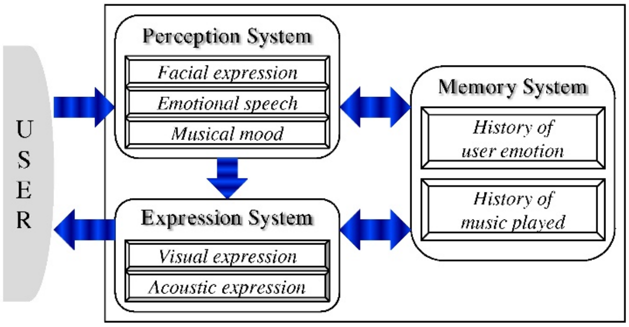

Figure 1 Framework for affective interaction.

Furthermore, our expression system produces affective reactions to the user emotions in more natural ways by playing music that the robot recommends or songs that the user previously listened to while exhibiting that emotion. The music-aided affective reaction is directly supported by the memory system. This system stores information on the music the user listens to with a particular emotional state. This section describes further specific features of each system in the framework of music-aided affective interaction.

4.1. Perception system

The perception system recognizes human emotional states on the basis of various indicators. For multimodal emotion recognition, the proposed system utilizes the musical mood as a supplementary indicator along with speech and facial expression as primary indicators. Consequently, the perception system comprises three recognition modules: for musical mood, facial expression, and speech emotion. Among them, modules based on face and speech are jointly handled as bimodal emotion recognition in this study. The overall process of this system is illustrated in Figure 2.

4.1.1. Musical mood recognition

One of the essential advantages of music-based emotion recognition is that monitoring of human emotion can be accomplished in the background without the user's attention. Users do not need to remain in front of the robot, since the musical sound can be loud enough to be analyzed in the perception system. For this reason, the module of the musical mood recognition is operated independently from the other modules in the perception system. Even though the musical mood provides a conjectured user emotion, the recognition result sufficiently enables the robot to naturally proceed with affective and friendly interaction with the user as long as the user plays music. For instance, if a user is listening to sad music, the robot can express concern, using a display or sound.

Compared to other tasks for musical information retrieval, such as genre identification, research on musical mood recognition is still in an early stage. General

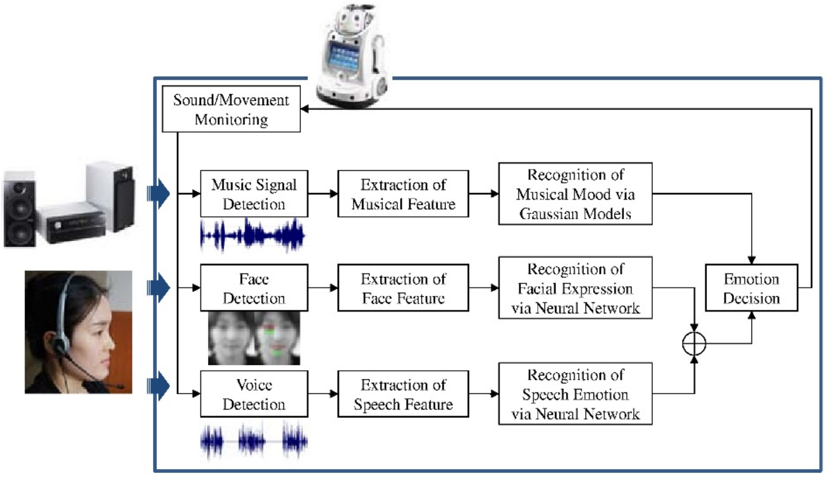
Sound/Movement Monitoring
Music Signal Detection
Extraction of Musical Feature
Recognition of Musical Mood via Gaussian Models
Face Detection
Extraction of Face Feature
Recognition of Facial Expression via Neural Network
Voice Detection
Extraction of Speech Feature
Recognition of Speech Emotion via Neural Network
Emotion Decision

Figure 2 Music-aided multimodal perception system.

approaches have concentrated on acoustic features representing the musical mood and criteria for the classification of moods [19-21]. A recent study focused on a context-based approach that uses contextual information such as websites, tags, and lyrics [22]. In this study, we attempt to identify the musical mood without consideration of contextual information to extend the range of music to instrumental music such as sound-tracks of films. Thus, we follow the general procedure of non-linguistic information retrieval from speech or sound [23,24].

The mood recognition module is activated when the perception system detects musical signals. Audio signals transmitted through a microphone of a robot can be either musical signals or human voice signals. Thus, the audio signals need to be classified into music and voice, since the system is programmed to process voice signals in the speech emotion recognition module. For the classification of audio signals, we employ the standard method of voice activity detection based on the zero crossing rate (ZCR) and energy [25]. When the audio signals indicate relatively high values in both ZCR and energy, the signals are regarded as musical signals. Otherwise, the signals are categorized as voice signals and submitted to the speech processing module.

The first step of the musical mood recognition is to extract acoustic features representing the musical mood. Several studies have reported that Mel-frequency cepstral coefficients (MFCC) provide reliable performance on musical mood recognition, as this feature reflects the nonlinear frequency sensitivity of the human auditory system [19,20]. Linear prediction coefficients (LPC) are also known as a useful feature that describes musical characteristics well [23]. These two features are commonly used as short-term acoustic features, non-linguistic characteristics of which are effectively defined with probability density functions such as a Gaussian distribution [26,27]. For this reason, we use these features as primary features. After extracting these features from each frame of 10-40 ms in music streams, their first and second derivatives are added to the feature set of the corresponding frame in order to consider temporal characteristics between consecutive frames.

The next step is to estimate the log-likelihood of the features on respective acoustic models constructed for each type of musical mood. Acoustic models should hence be trained in advance of this step. In this study, the distribution of acoustic features extracted from music data corresponding to each mood is modeled by a Gaussian density function. Thus, a Gaussian mixture model (GMM) is constructed for each musical mood in accordance with model training procedures. The log-likelihood of feature vectors extracted from given music signals is computed on each GMM, as follows:

$$\log P(X|\lambda_i) = \sum_{t=1}^{T} \log P(\vec{x}_t|\lambda_i), \tag{1}$$

where $X(=\{\vec{x}_1, ..., \vec{x}_T\})$ refers to a vector sequence of acoustic features that are extracted from the music stream, and GMM λ_i (i = 1,...,M if there are M musical moods) indicates the mood model. M log-likelihood results are then submitted to the emotion decision process.

4.1.2. Bimodal emotion recognition from facial expression and speech

Facial expression and speech are the representative indicators that directly convey human emotional information. Because those indicators provide emotional information that is supplementary and/or complementary to each other, they have successfully been combined in terms of bimodal indicators. The bimodal emotion recognition approach integrates the recognition results, respectively, obtained from face and speech.

In facial expression recognition, accurate detection of the face has an important influence on the recognition performance. A bottom-up, feature-based approach is widely used for the robust face detection. This approach searches an image through a set of facial features indicating color and shape, and then groups them into face candidates based on the geometric relationship of the facial features. Finally, a candidate region is decided as a face by locating eyes in the eye region of a candidate's face. The detected facial image is submitted to the module for facial expression recognition.

The first step of facial expression recognition is to normalize the captured image. Two kinds of features are then extracted on the basis of Ekman's facial expression features [28]. The first feature is a facial image consisting of three facial regions: the lips, eyebrows, and forehead. By applying histogram equalization and the threshold of the standard distribution of the brightness of the normalized facial image, each of the facial regions is extracted from the entire image. The second feature is an edge image of those three facial regions. The edges around the regions are extracted by using histogram equalization.

Next, the facial features are trained according to a specific classifier in order to determine explicitly distinctive boundaries between emotions. The boundary is used as a criterion to decide an emotional state for a given facial image. Various techniques already in use for conventional pattern classification problems are likewise used for such emotion classifiers. Among them, neural network (NN)-based approaches have widely been adopted for facial emotion recognition, and have provided reliable performance [29-31]. A recent study on NN-based emotion recognition [32] reported the efficiency of the back-propagation (BP) algorithm proposed by Rumelhart and McClelland in 1986 [33]. In this study, we follow a training procedure introduced in [31] that uses an advanced BP algorithm called error BP.

Each of the extracted features is trained by using two neural networks for each type of emotion. Each neural network is composed of 1610 input nodes, 6 hidden nodes, and M output nodes. The 1610 input nodes receive 1610 pixels from the input image, and the output nodes, respectively, correspond to each of M emotions. The number of hidden nodes was determined by an experiment. Finally, the decision logic determines the final emotion from the two neural network results. The face-emotion decision logic utilizes the weighted sum of the two results and a voting method of the result transitions over the time domain. The overall process of emotion recognition through facial expression is shown in Figure 3.

Once audio signals transmitted through a robot microphone are determined to be human voice signals, the speech emotion recognition module is activated. In the first step, several acoustic features representing emotional characteristics are estimated from the voice signals. Two types of acoustic features are extracted: a phonetic feature and a prosodic feature. MFCC and LPC pertaining to musical mood recognition are also employed for speech emotion recognition in terms of phonetic features, while spectral energy and pitch are used as prosodic features. As in musical mood recognition, the first and second derivatives of all features are added to the feature set.

Next, the acoustic features are recognized through a pattern classifier. Even though various classifiers such as HMM and SVM have been fed into speech emotion recognition tasks, we employ the neural network-based classifier used in the facial expression recognition module in order to efficiently handle the fusion process in which the recognition results of two indicators are integrated. We organize a sub-neural network for each emotion. The construction of each sub-network has basically the same architecture. A sub-network comprises input nodes corresponding to the dimension of the acoustic features, hidden nodes, and an output node. The number of hidden nodes varies according to the distinctness of respective emotions. When there are M emotions, acoustic features extracted from the voice signals are simultaneously fed into M sub-networks, and thus an M-dimensional vector is obtained for the recognition result. The configuration of the neural network is similar to that adopted in [17], but we adjust internal learning weights of each sub-network and the normalization algorithm in consideration of the characteristics of the acoustic features.

Figure 4 describes a simplified architecture for the proposed bimodal recognition when the number of emotions is M. As a recognition result, an M-dimensional vector is, respectively, obtained from facial expression and speech. Let us denote $R_{\text{face}}(t)$ and $R_{\text{speech}}(t)$ as the vectors obtained from the two indicators at time t. The final procedure of the bimodal emotion recognition is to perform a fusion process in which the results, $R_{\text{face}}(t)$ and $R_{\text{speech}}(t)$, are integrated. We calculate the vector $R_{\text{bimodal}}(t)$ referred to as a fusion vector, as follows:

$$R_{\text{bimodal}}(t) = W_{\text{f}}R_{\text{face}}(t) + W_{\text{s}}R_{\text{speech}}(t) + W_{\text{f}}R_{\text{face}}(t-1) + W_{\text{s}}R_{\text{speech}}(t-1) \tag{2}$$

where W_{f} and W_{s} are the weights for the respective indicators.

The weights are appropriately determined by reference to the recognition results for each indicator.

In general, the performance of standard emotion recognition systems substantially depends on the user characteristics in expressing emotional states [6]. Thus, systems occasionally demonstrate the common error of a rapid transition of human emotional states. To address this problem, we consider the general tendency that human emotional states rarely change quickly back and forth. Hence, the proposed fusion process in (2) uses two recognition results obtained just before the current

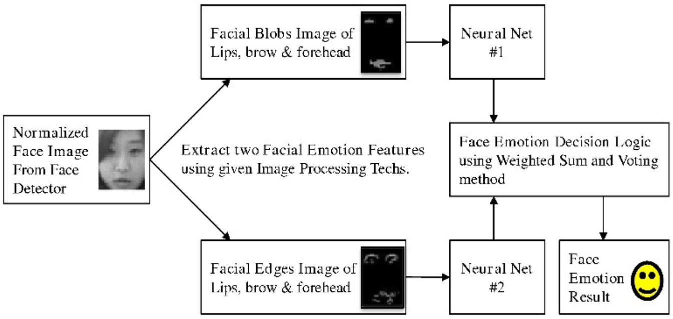
Facial Blobs Image of Lips, brow & forehead
Neural Net #1
Normalized Face Image From Face Detector
Extract two Facial Emotion Features using given Image Processing Techs.
Face Emotion Decision Logic using Weighted Sum and Voting method
Facial Edges Image of Lips, brow & forehead
Neural Net #2
Face Emotion Result

Figure 3 Process of emotion recognition through facial expression.

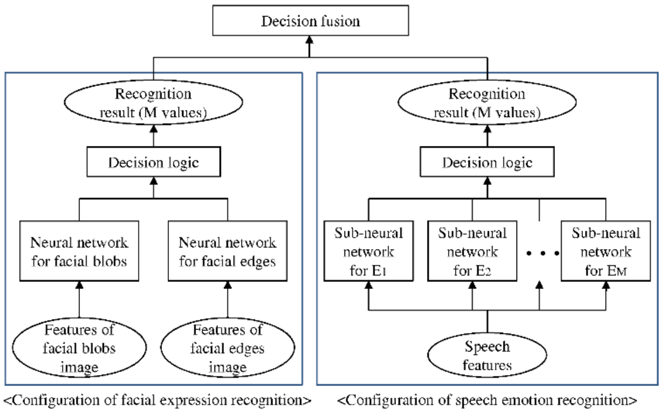
Decision fusion
Recognition result (M values)
Decision logic
Neural network for facial blobs
Neural network for facial edges
Features of facial blobs image
Features of facial edges image
Recognition result (M values)
Decision logic
Sub-neural network for E1
Sub-neural network for E2
Sub-neural network for EM
Speech features
<Configuration of facial expression recognition>
<Configuration of speech emotion recognition>

Figure 4 Architecture for bimodal emotion recognition.

time t in order to reflect the emotional state demonstrated during the previous time.

4.1.3. Emotion decision

The final procedure in the perception system is to determine an emotion on the basis of the bimodal fusion vector calculated in (2) and the mood recognition result estimated in (1). These two results indicate different scales about the values but have the same dimension corresponding to the number of emotions and moods. Let us denote $R_{\mathrm{bimodal}}(e)$ and $R_{\mathrm{music}}(e)$ as the value of the eth emotion in the fusion vector and that of the eth mood in the mood recognition result, respectively.

In addition to these two results, our decision mechanism utilizes supplementary information. This research originated from the general idea of the relationship between music and human emotion. We strongly believe that the music that a person listens to directly correlates with the emotion that the person feels. Thus, if a robot detects a musical mood similar to the mood that a user has enjoyed in the past, the user would be in an emotional state similar to the emotion the robot determined at that time. To consider this aspect, we steadily make a record of bimodal recognition results in accordance with the musical mood whenever the three recognition modules are simultaneously activated. The average values of the bimodal results on each type of musical mood are then recorded in the memory system, which will be described in the following section. Let us denote $\widehat{R}^{m}_{\mathrm{bimodal}}(e)$ as the value of the eth emotion in the averaged fusion vector on the mth musical mood. $\widehat{R}^{m}_{\mathrm{bimodal}}(e)$ demonstrates the extent of an emotion the user feels while a certain type of musical mood is being played. To utilize this value in the decision process, the mood of the music being played should be determined in advance, as follows:

$$\widehat{m} = \arg\max_{m} \mathrm{R}_{\mathrm{music}}(m), \quad m = 1, 2, \ldots, M \tag{3}$$

Finally, we determine an emotion based on three kinds of results, all of which are M-dimensional vectors, as follows:

$$\widehat{e} = \arg\max_{e} \left(W_{\mathrm{b}} R_{\mathrm{bimodal}}(e) + W_{\mathrm{m}} R_{\mathrm{music}}(e) + \widehat{W}_{\mathrm{b}} \widehat{R}^{\widehat{m}}_{\mathrm{bimodal}}(e) \right), \tag{4}$$

where W_{b}, W_{m}, and $\widehat{W}_{\mathrm{b}}$ refer to the weights or scaling factors for the corresponding results. This decision criterion is available only if either the bimodal indicators or musical indicator is activated. When only the musical signals are detected, $R_{\mathrm{bimodal}}(e)$ is automatically set to zero. If musical signals are not detected, the music-based results, $R_{\mathrm{music}}(e)$ and $\widehat{R}^{\widehat{m}}_{\mathrm{bimodal}}(e)$, are set to zero.

4.2. Memory system

The memory system consecutively records several specific types of emotions such as happy, sad, or angry from among the emotions that the perception system detects from three kinds of indicators. The system creates emotion records including the emotion type and time. Such emotion records can naturally be utilized for affective interaction with the user. For example, the robot can express concern the day after the user has been angered by something. When a negative user emotion is sustained for a long time, the memory system may attempt to change the user's negative feeling, forcing the expression system to control the degree of expression.

In addition to emotional information, the memory system records the information of music detected by the perception system. The system obtains and accumulates musical information such as the genre, title, and musician of the detected music, supported by an online music information retrieval system. The accumulated musical information is used to organize a music library directly oriented to the user, which provides explicit information of the user's favorite genres and musicians as well as the musical mood.

Although music is non-verbal information, the music library enables the robot to have more advanced and intelligent interaction skills. On the basis of this library, the robot may offer several songs befitting the user's emotion or recommend other songs similar to the music that the user is listening to. While a recommended song is played, the perception system monitors the user's response through the bimodal indicators.

The feedback, either negative or positive, on the song is then recorded in the memory system to be utilized in future interactions. The music library is continuously updated whenever the user plays a new song or provides feedback through emotional expression.

As addressed in the previous section, if both the bimodal and the musical indicators are activated, the bimodal recognition results and the musical mood are recorded in a table form in the memory system, as shown in Table 1. This table records the average values of the bimodal recognition results corresponding to M emotions for each type of musical mood. The first row and the first column are an index of emotions and moods, respectively. This information demonstrates past emotional experience of the user for each type of musical mood.

Figure 5 summarizes the functions and procedures of the memory system.

4.3. Expression system

The expression system is an intermediate interface between the memory system and robot hardware devices. This system executes behavior operations and/or emotion expressions in order to react to the user emotions. Both operations basically depend upon robot hardware devices, since every service robot has different hardware capacity to process expression operations.

However, the general types of operations are eventually concluded in visual and acoustic expressions. A straightforward method of visual expression is to use facial expression such as eye expression, the color of cheeks, and the shape of lips. In addition, operational behaviors such as movements and hand-shake or displaying graphical images on the screen are also effective ways to visually demonstrate the robot reaction. For instance, if a user feels happy, the robot could express the same emotion, by raising its hands or exhibiting a smile on either its face or the screen. Figure 6 shows five types of facial and graphical emotion expressions of the home service robot, iRobiQ. Figure 7 presents behavior-based interactions with a user.

Table 1 History of user emotions for each type of musical mood

	*E*1	*E*2	...	*EM*
***M*1**	0.23	0.78	...	0.13
***M*2**	0.87	0.12	...	0.32
:	:	:	...	:
MM	0.02	0.82	...	0.34

The second type of expression operations utilizes acoustic properties. The expression system can naturally produce emotional reactions through synthesized speech or music. Whenever the expression system determines either the context of the synthesized speech or music to play, the historical records of user emotion and music information provided by the memory system are utilized. If the perception system detects a certain type of emotion from a user, the expression system can recommend several songs that the user has previously listened to while experiencing that emotion. Since the memory device of robots can store a great number of music data, users hardly predict which song will be played. Thus, the music-aided expression system provides a more interesting and natural way of interaction between users and robots.

5. Experimental setup and results

This article proposed a framework for music-aided affective interaction between humans and robots. For evaluation of our approach, we implemented the proposed framework on a service robot, iRobiQ, and simulated the human-robot interaction. We attempted to evaluate the efficiency of two fundamental systems that play the most important roles in the framework: the perception and the expression systems. We first introduce the technical specifications of iRobiQ used as the robot platform in our research, and experimental results are subsequently presented.

5.1. iRobiQ: a home service robot

iRobiQ has been developed by Yujin Robot company under the support of the Korean government [34]. This robot is a general model of an intelligent service robot aimed at providing fun and convenience to users through various useful services such as educating children and monitoring home safety. Figure 8 summarizes several hardware and functions. A 7-inch LCD touch screen and a 1.3-megapixel camera as well as general sound devices such as a speaker and a microphone are equipped. Five kinds of sensor enable movements and reaction to human touch. iRobiQ is able to move around while avoiding obstacles.

This robot has its own hardware system for facial expression, and five types of facial expressions can be displayed: shy, disappointed, neutral, happy, and

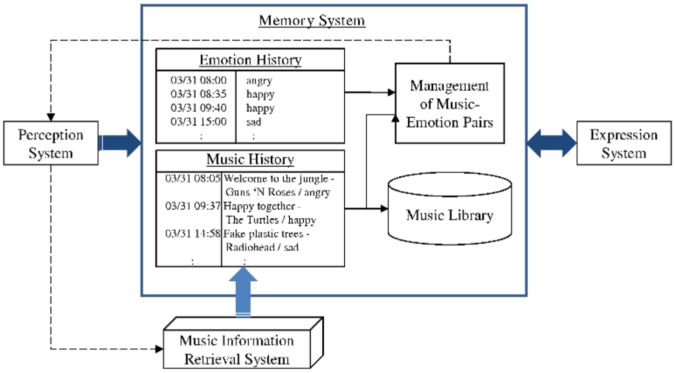
Memory System
Emotion History
03/31 08:00 angry
03/31 08:35 happy
03/31 09:40 happy
03/31 15:00 sad
Music History
03/31 08:05 Welcome to the jungle - Guns 'N Roses / angry
03/31 09:37 Happy together - The Turtles / happy
03/31 11:58 Fake plastic trees - Radiohead / sad
Management of Music-Emotion Pairs
Music Library
Perception System
Expression System
Music Information Retrieval System

Figure 5 Architecture of the memory system.

surprised. In addition, iRobiQ has two eyes designed on a segmented LCD displaying eye expressions. It also has LED dot matrices on its cheeks and mouth with which various emotions are expressed.

On the LCD screen located on the robot's chest region, a variety of graphical images can be displayed. For this study, we implemented several graphical face images and used them to represent the robot emotions more directly while interacting with a user. When compared to existing mechanical face robots, which require very complex motor-driven mechanisms and artificial skin, this kind of facial expression can deliver robot emotions in a more intimate manner [35].

5.2. Evaluation of the perception system

For the evaluation of the proposed perception system, we first conducted three kinds of emotion-recognition experiments independently: facial expression recognition, speech emotion recognition, and musical mood recognition. We then investigated the performance improvement in bimodal emotion recognition based on the proposed fusion process. Finally, music-aided multimodal emotion recognition was evaluated.

5.2.1. Experimental setups

To fairly verify each recognition module and to simulate bimodal and multimodal emotion recognition, we used four kinds of emotions or musical moods in each experiment: neutral, happy, angry, and sad. We chose 'angry' and 'sad' as the most representative negative emotions, whereas 'neutral' and 'happy' were chosen as non-negative emotions.

A typical difficulty in a standard multimodal emotion recognition task is data collection. In general, people of different countries have their own characteristic ways of expressing emotions facially and vocally. Thus, there are few standard multimodal databases collected from people around the world. Instead, most research studies depend on facial images and speech data obtained from nationals of a country [36,37]. We prepared training and evaluation data from ten Korean participants (five men and five women) who were asked to express emotions while making an emotional face and speaking short phrases of ordinary dialogue. Each participant generated five facial images and spoken data, respectively, for each emotion, while changing the contents of the dialogue. Consequently, 200 facial images and 200 spoken data were collected. All data were recorded in a quiet environment without any background noise.

All experiments were conducted by *k*-fold cross validation to fairly assess the recognition performance for respective persons. In *k*-fold cross validation, the original sample is partitioned into *k* subsamples and each subsample is retained in turn as evaluation data for testing while the remaining *k*-1 subsamples are used as

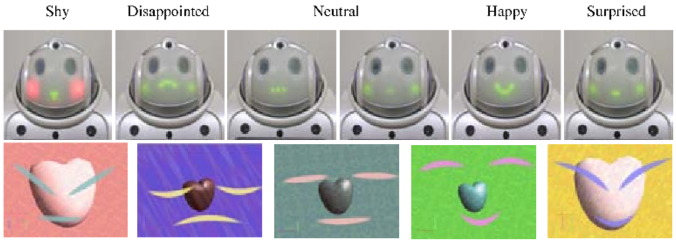

Figure 6 Facial expression (upper) and graphical expression (below) of iRobiQ representing five types of emotions.

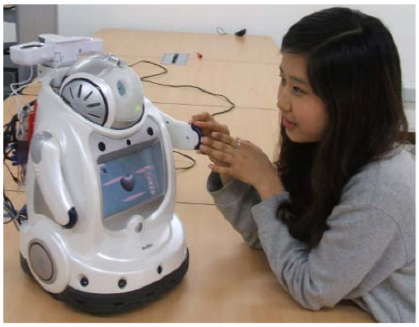

Figure 7 Behavior-based emotion expression of iRobiQ.

training data. The cross validation is thus repeated *k* times, with each of the *k* subsamples used exactly once as validation data. Hence, we repeated the evaluations ten times in accordance with a tenfold cross validation.

For the musical mood recognition, we collected several music clips of 30-s duration from a website, AMG, which provides a variety of music clips categorized by musical mood [22]. However, this website provides a limited number of clips in each category and does not include a category of neutral mood. For this reason, we added several categories to each of our four fundamental types of moods, as shown in Table 2. This mood categorization was reasonably verified by a human listening test that we conducted. For each mood type, 30 clips and 10 clips consisting of different songs were used for model training and evaluation, respectively.

In the human listening test, 30 participants (native speakers of Korean) listened to music clips chosen randomly from the website. The participants then classified each clip into one type from among the four types of musical moods provided in Table 2. Because all the participants are native Koreans, they could concentrate on the musical mood, ignoring the lyrics of the clips that were in English.

Table 2 Musical mood categories in which similar types of moods were combined and used for the clip selection from AMG

Neutral types	Happy types	Angry types	Sad types
Romantic	Happy	Angry	Sad
Gentle	Joyous	Aggressive	Melancholy
Sweet	Fun	Tense/Anxious	Gloomy

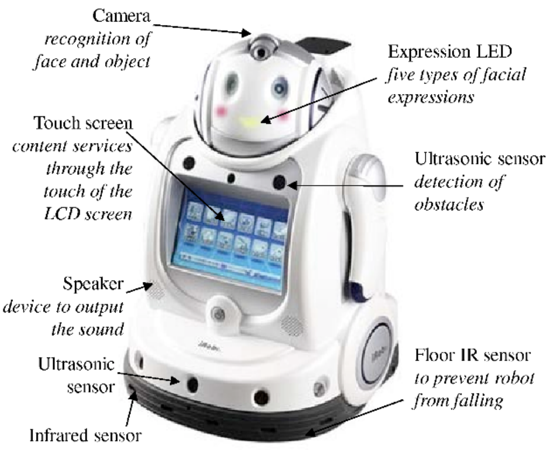
Camera
recognition of face and object
Expression LED
five types of facial expressions
Touch screen
content services through the touch of the LCD screen
Ultrasonic sensor
detection of obstacles
Speaker
device to output the sound
Ultrasonic sensor
Floor IR sensor
to prevent robot from falling
Infrared sensor

Figure 8 Specifications of iRobiQ.

5.2.2. Experimental results of unimodal and bimodal emotion recognition

First, we evaluated the performance of the facial recognition module. In this experiment, facial images were categorized according to emotional states. We investigated the performances of each of two neural networks trained for facial blobs images and facial edges images, respectively. The recognition results are presented in Table 3. Our facial expression recognition module achieved average recognition accuracy of 76.5 and 77.1% on each neural network. However, when the best result was selected from the two networks by a decision process using the weighted sum, the module performance was slightly improved to 78.4%.

Next, we evaluated the performance of the speech emotion recognition module. Spoken data were divided into two groups of men and women based on an assumption that men and women tend to express their emotional states with different speaking characteristics. Table 4 demonstrates the recognition performance. Our speech emotion recognition module showed average accuracy of 78.0 and 76.6% for the male and female tasks, respectively.

These two kinds of experimental results indicate that the two indicators (face and speech) do not operate in the same way. For example, the facial expression module best categorized the happy expression images, whereas the speech module determined the speech of sad emotion with the best accuracy. Meanwhile, the angry emotion was better detected in the speech than the facial expression. Such results emphasize the general necessity of bimodal emotion recognition.

Table 3 Performance (%) of neural network-based facial expression recognition module

Neural network #1		Neural network #2	
Neutral	77.4	Neutral	64.1
Happy	84.1	Happy	83.1
Angry	66.1	Angry	81.4
Sad	78.4	Sad	79.9
Average	76.5	Average	77.1

Table 4 Performance (%) of neural network-based speech emotion recognition module

Neural network for men		Neural network for women	
Neutral	74.6	Neutral	71.1
Happy	73.1	Happy	76.4
Angry	82.9	Angry	74.6
Sad	81.4	Sad	84.4
Average	78.0	Average	76.6

To simulate bimodal recognition experiments, we asked each participant to make an emotional face and simultaneously to emotionally vocalize several given sentences for respective emotions. An emotion was then determined for each trial in real time based on the bimodal fusion process. We investigated the efficiency of the proposed fusion process described in (2), which considers the general tendency that human emotional states rarely change quickly back and forth. For this evaluation, emotions that the participants were requested to express were given sequentially without rapid changes. For the purpose of comparison, we also investigated the results on a simple fusion process that uses the sum of two unimodal results without consideration of the previous emotion.

Table 5 represents the recognition results for the two fusion approaches. The bimodality improved the performances of the two unimodal indicators. It is of interest that the differences in the recognition accuracy over emotions were significantly reduced when compared to the results in Tables 3 and 4. This indicates that the bimodal approach provides more sophisticated determination of emotions for a person's bimodal emotion expression, as compared with a single indicator-based emotion recognition. In particular, the results verify the efficiency of the proposed fusion process, confirming that it is valid to use previous emotional information in the determination of a current emotion.

Finally, we evaluated the musical mood recognition module. This experiment was directly conducted with iRobiQ, which determined a musical mood while listening to a music clip played at a slight distance from the robot. Table 6 shows a confusion matrix of the mood recognition. The average recognition accuracy was reported as 81.3%, which represents comparable performance with that of the bimodal indicators. This demonstrates that music-based emotion recognition is quite suitable to our affective system and is expected to complement several disadvantages of the bimodal indicators.

Table 5 Performance (%) of bimodal emotion recognition

	With a simple fusion process	With the proposed fusion process
Neutral	78.4	79.2
Happy	82.1	82.9
Angry	79.0	80.8
Sad	81.3	82.5
Average	80.2	81.4

Table 6 Confusion matrix of the musical mood recognition

	Neutral	Happy	Angry	Sad
Neutral	0.76	0.08	0.05	0.11
Happy	0.05	0.83	0.10	0.02
Angry	0.04	0.09	0.85	0.02
Sad	0.10	0.05	0.04	0.81

5.2.3. Experimental results of music-aided multimodal emotion recognition

In order to investigate whether the results of musical mood recognition can complement the emotion results from the bimodal indicators, two types of multimodal experiments were conducted. In the first experiment, we virtually simulated multimodal recognition in iRobiQ, directly utilizing the evaluation data prepared for the unimodal experiments. We assumed that the evaluation data, that is, the audio-visual data, and music clips, are entered into the affective system of the robot. Each recognition module in the perception system computed the results from the data independently and emotions were determined by the decision criterion described in Section 4.1.3. In this section, we introduced the use of previously recorded bimodal recognition results. When three recognition modules in the perception system are activated at the same time, bimodal recognition results are recorded in relation to the musical mood. The records are later utilized in an emotion-decision process along with the results of bimodal and musical mood recognition. We attempted to verify the efficiency of this approach, but could not naturally perform the evaluation since the evaluation data are obtained from participants who should imitate the emotional expression and the music clip is selected without consideration of personal musical preference. Thus, in the first experiment, the previous record of bimodal results, $\widehat{W}_{\mathrm{b}}$, was set to zero so as to be ignored.

The second experiment was conducted in a more natural way, supported by human participants. Each participant was asked to carry out the same behavior as in the bimodal recognition experiments, looking at iRobiQ and making an emotional face and speech for respective emotions. At the same time, we played several songs categorized into a mood similar to the corresponding emotion, at a slight distance from the robot. At that instance, the robot receives three kinds of emotional data and proceeds to compute the recognition results via respective recognition modules in the perception system. It should be noted that if both musical and human voice signals are entered into a single microphone, two types of signals act as noise signals for each other, deteriorating the recognition accuracy. The ideal solution for this problem is to operate a process of blind source separation that divides audio signals into music and voice [38]. However, the correctness of the separation task would naturally affect the performance of the two audio recognition modules. For this reason, this study does not consider the problem caused by a single microphone in order to concentrate on the performance evaluation of respective recognition modules and their multimodality. Thus, iRobiQ receives two different types of audio signals, respectively, from two different microphones, one of which is an ear microphone the participant wears for the voice input and the other is a general microphone equipped on the robot for the music input. The musical signals that the robot-equipped microphone receives while the ear microphone is activated are regarded as music-mixed voice signals and are excluded from the musical mood recognition task.

In this experiment, we attempted to use the previous bimodal results in the emotion decision. Once the results of bimodal and musical mood recognition were computed in respective modules, the bimodal recognition results were used to update the average value of the bimodal results on the determined musical mood. The average value was then utilized in the emotion decision process along with the bimodal and musical mood recognition results, on the basis of (4).

Figure 9 compares the performances of two multimodal experiments with bimodal results. The multimodal recognition achieved superior accuracy compared to bimodal results over all emotions. The results in this figure confirm that the proposed music-aided multimodal approach notably advanced the standard bimodal approach. The results of the first multimodal experiment (called ‘Multimodal1’) and the second experiment (called ‘Multimodal2’), respectively, represented relative improvements of 15 and 46% (2.7 and 8.5% in absolute improvement) over the bimodal result. The performance improvement was more significant for emotions such as happy and angry, where the musical mood recognition indicated higher accuracy. In particular, Multimodal2 showed relative improvement of 36% over Multimodal1, which only uses the results of bimodal and musical mood recognition. This supports our idea that the record of bimodal results in relation with the musical mood provides supplementary information in the emotion decision during the current time when the relevant musical mood is entered.

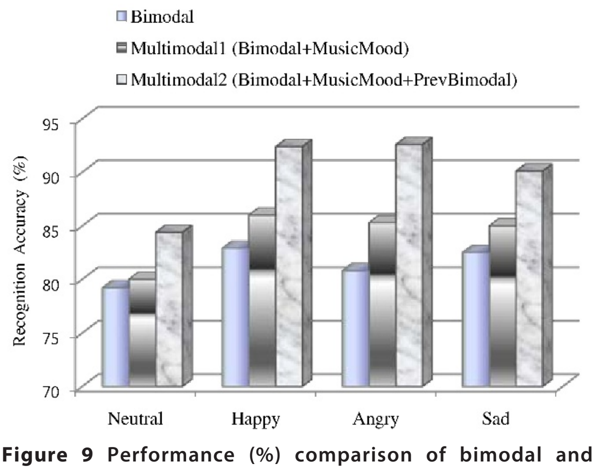

Figure 9 Performance (%) comparison of bimodal and multimodal emotion recognition.

From these experiments, we conclude that musical mood information can effectively be utilized as a supplementary and complementary indicator in standard emotion recognition tasks based on speech and facial expression. Nevertheless, we need to further consider that different users might enjoy different types of musical moods while being in a certain emotional state. In the human listening test conducted for the verification of mood categorization, at least 70% of the participants determined an identical mood for each clip. This result indicates that humans tend to feel similar emotions while listening to music. Even when people feel different emotions prior to listening to music, they will experience the same emotion due to the certain mood of the music. This conclusion is closely associated with the general knowledge addressed in Section 1 that music greatly influences the affective states of humans. Consequently, musical mood recognition has strong possibility of improving the reliability of affective interaction between humans and robots.

5.3. Evaluation of the expression system

In the proposed affective interaction, the expression system provides a natural and intermediate interface between humans and service robots. As addressed in Section 4.3, the system enables service robots to appropriately react to the user emotion through visual and acoustic expressions. In particular, the proposed expression system produces an affective reaction in more natural ways by playing music that the robot recommends or several songs that the user listened to while being in that emotional state in the past.

We evaluated the music-aided expression system in a subjective manner. Each participant was asked to attempt to emotionally interact with the robot during a period of a few minutes. When a participant makes speech and facial expression corresponding to a certain emotion type, iRobiQ determines the emotional state and plays several music clips relevant to the emotion, representing audio-visual expression such as eye expression, cheek color, and synthesized speech. After finishing the interaction test, participants recorded a score from 1 to 5 regarding naturalness, affectiveness, and interest for each robot property. Figure 10 illustrates the results. Every property indicated a different score in each item, while the music recommendation property of the robot showed better results for the overall items. Compared to synthesized speech, music-driven interface was reported to provide acoustically more affective and natural interaction for human. Although these results were obtained by a subjective test, most of the participants reported a favorable reaction toward the music-aided affective interaction.

6. Conclusions

This study proposed an efficient framework for affective interaction between humans and robots. The framework comprises three systems: perception, memory, and expression. In each system, musical moods are utilized as important information. The perception system recognizes the mood of the music that the user listens to and uses it to determine an emotional state of the user along with facial expression and speech. The memory system records musical moods corresponding to emotional states of the user and submits the information to the expression system, which enable the robot to produce more natural reaction by playing music relevant to the user emotion. On emotion recognition experiments conducted with a service robot, iRobiQ, the music-aided multimodal approach demonstrated superior performance over unimodal and bimodal approaches. Moreover, human participants reported favorable reactions toward the music-aided interaction with a robot.

In future study, we will evaluate our approach on more amounts of emotional data. In addition, we will investigate an ideal combination of emotional features and classifiers including SVM and HMM in the proposed approach.

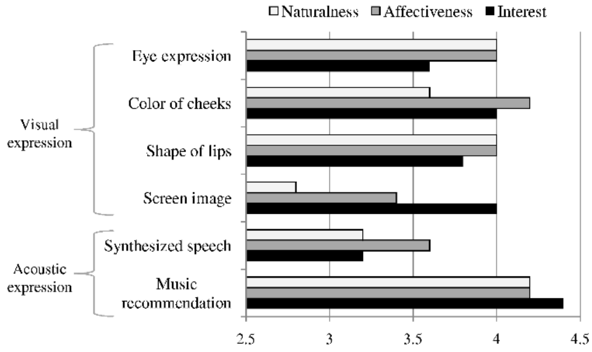

Figure 10 Results of interaction test for expression properties of robot.

Acknowledgements
This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (MEST) (2011-0013776 and 2010-0025642).

Author details
[1]Department of Intelligent Robot Engineering, Mokwon University, Daejeon, South Korea [2]School of Electrical and Computer Engineering, Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea
[3]Department of Intelligent Robot Engineering, Mokwon University, Daejeon, South Korea

Competing interests
The authors declare that they have no competing interests.

Received: 2 April 2011 Accepted: 19 January 2012
Published: 19 January 2012

References

1. M Richins, Measuring emotions in the consumption experience. J Consum Res. **24**, 127–146 (1997). doi:10.1086/209499
2. R Cowie, E Cowie, N Tsapatsoulis, G Votsis, S Kollias, W Fellenz, J Taylor, Emotion recognition in human-computer interaction. IEEE Signal Process Mag. **18**, 32–80 (2001). doi:10.1109/79.911197
3. T Nwe, S Foo, L Silva, Speech emotion recognition using hidden Markov models. Speech Commun. **41**(4), 603–623 (2003). doi:10.1016/S0167-6393(03)00099-2
4. M Paleari, B Huet, R Chellali, Towards multimodal emotion recognition: a new approach. in *Proc Conf Image Video Retrieval*, Xi'an, China 174–181 (2010)
5. LC De Silva, PC Ng, Bimodal emotion recognition, in *Proc of Fourth IEEE Int Conf Automatic Face Gesture Recog*, Grenoble, France, pp. 332–335 (2000)
6. LM Ignacio, OR Carlos, GR Joaquin, R Daniel, Speaker dependent emotion recognition using prosodic supervectors, in *Proc of Interspeech*, Brighton, UK, pp. 1971–1974 (2009)
7. CC Pratt, in *Music as the Language of Emotion: a Lecture delivered in the Whittall Pavilion of the Library of Congress* (US Govt. Print. Off., Washington, 1952)
8. KR Scherer, MR Zentner, A Schacht, Emotional states generated by music: an exploratory study of music experts. Musicae Scientiae. **2001-2002**, 149–171 (2002)
9. A Makiko, N Toshie, K Satoshi, N Chika, K Tomotsugu, Psychological research on emotions in strong experiences with music. Human Interface. **2003**, 477–480 (2003)
10. A Gabrielsson, Some reflections on links between music psychology and music education. Res Higher Music Educ. **2002**(2), 77–86 (2002)
11. C Bartneck, M Okada, Robotic user interfaces, in *Proc Human Comp Conf* (Aizu-Wakamatsu, Japan, 2001), pp. 130–140
12. C Breazeal, B Scassellati, A context-dependent attention system for a Social Robot, in *Proc Sixteenth Int Joint Conf Art Intel* (Stockholm, Sweden, 1999), pp. 1146–1151
13. RC Arkin, M Fujita, T Takagi, R Hasegawa, An ethological and emotional basis for human-robot interaction. Robot Autonomous Syst. **42**, 191–201 (2003). doi:10.1016/S0921-8890(02)00375-5
14. CL Sidner, C Lee, CD Kidds, N Lesh, C Rich, Explorations in engagement for humans and robots. Artif Intell. **166**, 140–164 (2005). doi:10.1016/j.artint.2005.03.005
15. T Shibata, T Tashima, K Tanie, Emergence of emotional behavior through physical interaction between human and artificial emotional creatures, in *Proc Int Conf Robotics Automation* (San Francisco, USA, 2000), pp. 2868–2873
16. N Tosa, R Nakatsu, Life-like communication agent-emotion sensing character "MIC" & feeling session character "MUSE", in *Proc Int Conf Multi Comp Syst* (Hiroshima, Japan, 1996), pp. 12–19
17. R Nakatsu, J Nicholson, N Tosa, Emotion recognition and its application to computer agents with spontaneous interactive capabilities, in *Proc IEEE Int Workshop Multi Signal Process* (Copenhagen, Denmark, 1999), pp. 439–444
18. J Ledoux, *The Emotional Brain: The Mysterious Underpinning of Emotional Life* (Simon & Schuster, New York, 1996)

19. EM Schmidt, D Turnbull, YE Kim, Feature selection for content-based, time-varying musical emotion regression, in *Proc ACM SIGMM Int Conf Multimedia Info Retrieval* (Philadelphia, USA, 2010), pp. 267–274
20. EM Schmidt, YE Kim, Prediction of time-varying musical mood distributions from audio, in *Proc Int Soc Music Inform Retrieval Conf* (Utrecht, Netherlands, 2010), pp. 465–470
21. YH Yang, YC Lin, YF Su, HH Chen, A regression approach to music emotion recognition. IEEE Trans Audio Speech Lang Process. **16**(2), 448–457 (2008)
22. YE Kim, EM Schmidt, R Migneco, BG Morton, P Richardson, J Scott, JA Speck, D Turnbull, Music emotion recognition: a state of the art review, in *Proc Int Soc Music Inform Retrieval Conf* (Utrecht, Netherlands, 2010), pp. 255–266
23. P Ahrendt, Music genre classification systems–a computational approach. Ph.D. dissertation, Technical University, Denmark (2006)
24. JS Park, JH Kim, YH Oh, Feature vector classification based speech emotion recognition for service robots. IEEE Trans Consum Electron V. **55**(3), 1590–1596 (2009)
25. X Yang, B Tan, J Ding, J Zhang, J Gong, Comparative study on voice activity detection algorithm, in *Proc Int Conf Elect Control Eng* (Wuhan, China, 2010), pp. 599–602
26. O Kwon, K Chan, J Hao, T Lee, Emotion recognition by speech signals, in *Proc Eurospeech* (Geneva, Switzerland, 2003), pp. 125–128
27. R Huang, C Ma, Toward a speaker-independent real time affect detection system, in *Proc Int Conf Pattern Recog* (Hong Kong, China, 2006), pp. 1204–1207
28. P Ekman, WV Friesen, in *Facial Action Coding System: Investigator's Guide* (Consulting Psychologists Press, Palo Alto, 1978)
29. S Giripunje, P Bajaj, A Abraham, Emotion recognition system using connectionist models, in *Proc Int Conf Cog Neural Syst* (Boston, USA, 2009), pp. 1–2
30. L Franco, A Treves, A neural network facial expression recognition system using unsupervised local processing, in *Proc Int Symposium Image Signal Process. Anal* (Pula, Croatia, 2001), pp. 628–632
31. HA Rowley, S Baluja, T Kanade, Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**(2), 23–38 (1998). doi:10.1007/BF03025294
32. X Zhu, Emotion recognition of EMG based on BP neural network, in *Proc Int Symposium Network. Network Security* (Jinggangshan, China, 2010), pp. 227–229
33. DE Rumelhart, JL McClelland, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (MIT Press, Cambridge, 1986)
34. J Han, S Lee, E Hyun, B Kang, K Shin, The birth story of robot, IROBIQ for children's tolerance, in *18th IEEE Int Symposium Robot Human Inter Comm* (Toyama, Japan, 2009), p. 318
35. HG Lee, MH Baeg, DW Lee, TG Lee, HS Park, Development of an android for emotional communication between human and machine: EveR-2, in *Proc Int Symposium Adv Robotics Machine Intell* (Beijing, China, 2006), pp. 41–47
36. D Ververidis, C Kotropoulos, Emotional speech recognition: resources, features, and methods. Speech Commun. **48**(9), 1162–1181 (2006). doi:10.1016/j.specom.2006.04.003
37. E Cowie, N Campbell, R Cowie, Roach P, Emotional speech: towards a new generation of databases. Speech Commun. **40**(1), 33–60 (2003). doi:10.1016/S0167-6393(02)00070-5
38. P Vanroose, Blind source separation of speech and background music for improved speech recognition, in *Proc of the 24th Symposium on Information Theory* (Yokohama, Japan, 2003), pp. 103–108

doi:10.1186/1687-4722-2012-5
Cite this article as: Park *et al.*: Music-aided affective interaction between human and service robot. *EURASIP Journal on Audio, Speech, and Music Processing* 2012 **2012**:5.