

# The Affective Remixer: Personalized Music Arranging

Jae-woo Chung, G. Scott Vercoe  
MIT Media Lab  
20 Ames St.  
Cambridge, MA 02139  
{jaewoo, scottyv}@media.mit.edu

## ABSTRACT

This paper describes a real-time music-arranging system that reacts to immediate affective cues from a listener. Data was collected on the potential of certain musical dimensions to elicit change in a listener's affective state using sound files created explicitly for the experiment through composition/production, segmentation, and re-assembly of music along these dimensions. Based on listener data, a probabilistic state transition model was developed to infer the listener's current affective state. A second model was made that would select music segments and re-arrange ('remix') them to induce a target affective state. We propose that this approach provides a new perspective for characterizing musical preference.

## Author Keywords

Music, Affective computing, Personal soundtracks, Automated music arranging, Automated remixing, Predictive listening, Affective-state-transition model

## ACM Classification Keywords

H.5.1 Multimedia Information Systems; H5.5 [Sound and music computing]: Methodologies and techniques; J.5 [Arts and Humanities]: Music

## INTRODUCTION

For a relatively passive activity, listening to music can be remarkably active. Particularly enthused listeners will mentally process the music and actively re-synthesize the signal in various forms of personal expression. Music may inspire physical interpretations such as head-nodding, hand-tapping, foot-tapping, hand-clapping, mock-performing ('air guitar'), mock-conducting, dancing and other gestures. Near-subconscious creation of these physical signals may exaggerate the physiological changes that already accompany affective change. If music affect is viewed from a cognition-arousal perspective [12], a listener's automatic gesturing would be considered part of the physiological cues that must be interpreted.

The affective response of a listener has been shown to contain seemingly conflicting data. Intuitively, we would

expect a listener physically involved in the music to respond with greater affective intensity. Studies have demonstrated a disparity between the affective response that is reported by the listener, and the response that is actually aroused. [13] The study of music-inspired listener gestures may provide some insight into the apparent gap between reported and aroused affects.

## BUILDING A STIMULUS MATRIX

A variety of affective music recommendation, retrieval and composition systems exist to serve many different functions. The macroscopic approach will typically return music in the form of song lists (*playlists*) [4], while the more microscopic approach of algorithmic music-writing systems often return compositions in the form of note and timing data [7].

Rather than high-level *playlists* or low-level algorithmic compositions, we instead focus on the re-arrangement of musical segments. Collecting listener reactions to selected music segments provides us with data that represents the affective potential of certain musical parameters. A system can then be created to take real-time affective cues from a listener and automatically re-arrange musical works. In this way, physical and physiological signals can be used as a means of directing a musical arrangement.

The goal of the project is to correlate selected musical parameters with changes in affective state. Using physical, physiological and self-reported data gathered from several participating listeners, responses were related to gradual changes in particular parameters of the current music. Given a musical context and a set of music parameters, a concrete picture of musical aesthetic can be drawn according to trends in listeners' responses.

To build a navigable stimulus matrix, sound segments are organized by parameter. The stimulus matrix allows a musical arrangement to be created simply by traversing the grid and appending segments to the arrangement. Musical arrangements can thereby be expressed as vectors in the parameter matrix. We now have a state transition model that selects a path through the music parameter matrix to move from the current to a target affective state.

## MUSIC AND AFFECT: ATTRIBUTION SPACE

Our goal is to observe how certain musical parameters elicit a change in affective response. To acquire useful

information, systems for the organization of *musical-parameter* data and *affective-state* data are needed.

### Affective States

We place a listener’s affective state into a two-dimensional space defined by the parameters *arousal* and *valence* [11].

- *Arousal*: affective parameter relating the potential of the music to elicit a reaction (positive or negative) in the listener
- *Valence*: affective parameter relating the listener’s affinity or aversion to the music

Music affect can be categorized by arousal-valence [3], allowing a vector to be used for describing the change in listener affect. Affective states can correspond to arousal-valence using a set of musically descriptive adjectives (Figure 1).

	Negative valence	Positive valence
High arousal	Annoying (S4)	Engaging (S1)
Low arousal	Boring (S3)	Soothing (S2)

Figure 1. Affective-state quadrants

### Musical Parameters

We place audio segments in a two-dimensional space defined by the independent musical parameters *vertical layering* and *horizontal complexity*:

- *Vertical layering*: musical parameter describing the thickness of layering within the piece
- *Horizontal complexity*: musical parameter describing the amount of rhythmic complexity

As musicians we know to use these dimensions carefully, observing that each can influence the other. However, as representative of two unique aspects of thematic development throughout a musical work, the parameters *layering* and *complexity* will move independently of one another. We produced the music for this experiment with these domains in mind.

The parameters *layering* and *complexity* were selected because they provide a general framework in which to compose and arrange music in a variety of styles. Musical arrangement can safely be regarded as a simpler cognitive operation for humans to perform. Naturally, designing an arrangement-only system enables the use of higher-level musical structures than would be necessary in a system designed for algorithmic composition.

By shaping a pre-conceived musical composition, thematic material need not be generated by a machine. The flexibility of the parameters *layering* and *complexity* allow music to be composed, organized and arranged in a musically intuitive way.

### Preparation of Music Clips

Pieces with constant tempi were made in five musical styles representing jazz, jazz-funk, rock, electronic and dance music. Original tracks were composed for the purposes of this experiment, so listeners had never heard the musical examples. Music was produced in various multi-track audio environments, allowing for horizontal and vertical layout of each piece.

#### Short Music Clips

Short looping audio segments (5 to 15 seconds each) representing a complete musical phrase, were extracted from each piece to reflect incremental navigation through the musical parameters *layering* and *complexity*. In the production of the short segments, care was taken to exaggerate the distinction between harmonic layering and rhythmic complexity as much as possible within the piece.

#### Music Parameter Matrix: Layering/complexity

Sixteen of the short music clips were arranged into a 4x4 table, creating a navigable map of sounds organized by relative layering/complexity strength. Audio segments in each matrix consisted exclusively of segments from a single composition.

#### Long Music Clips

Four longer audio clips (30 to 45 seconds) were assembled by incrementally traversing the matrix:

- Increasing horizontal complexity
- Increasing vertical layering
- Decreasing horizontal complexity
- Decreasing vertical layering

Each resulting sound file represents a one-dimensional vector starting at one side of the matrix and ending at the opposite side. For each of four directions, we listened to each of the four possible starting points of the vector, and chose the one that most clearly exposed the changing parameter. The selected matrix vectors were rendered as sound files. Four clips were assembled for each of the five musical styles, providing 20 clips total, each lasting between 30 and 45 seconds.

### DATA COLLECTION

Eight subjects (students, 5 male/3 female, ages 25-40) listened to 20 audio clips (30 to 45 seconds each). The experiment took approximately 35-45 minutes to complete, including 5-10 minutes for setup and instruction.

### Affect Data

Listener’s Galvanic Skin Response (GSR), foot-tapping and subjective evaluation data were recorded, representing physiological, physical, and self-report data, respectively. Three unique sets of data were selected to provide diverse measures of affective response. While a subject’s GSR has been shown to be a relatively reliable indicator of arousal level [10], foot-tapping does not yet have any such

empirical evidence. While foot-tapping should rightly be considered a physical signal, one could argue that the unconscious foot-tapping of an enthusiastic listener approaches being a physiological signal.

**Physiological Data (skin arousal)**

A Galvactivator [9] reported the GSR of each subject, which was then sampled and recorded. After setup and instruction, they were asked to relax for a minute to normalize their GSR before listening.

**Physical Data (foot-tapping)**

To gather foot-tapping data, a microphone was placed to capture the sound of the listener’s feet. The audio signal was routed to a beat-tracking algorithm programmed in MAX/MSP. The tempo and velocity of the listener’s foot-tapping were recorded. Participants were told they could tap their feet if this seemed a natural response.

**Self-reported Data (subjective evaluation)**

Subjects completed a quick subjective evaluation twice during each clip, recording separate *initial* and *final* reactions to the music. Each musical example corresponded to a set of questions on the form provided to listeners. Subjects were asked to provide their initial reaction as soon as the music began, and their reaction again at the conclusion of the musical example.

```
CLIP 5:
Initial Reaction:
  dislike [ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]
  like: engaging [ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]
  dislike:annoying [ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]
Final Reaction:
  dislike [ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]
  like: engaging [ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]
  dislike:annoying [ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]--[ ]
```

Figure 2. Self-report data example

**Affect Data Mapping**

Music affect was self-reported by the listener using a pair of 7-point scales. Results of each rating in the survey were mapped to integers from -3 to +3, where a value of 0 corresponds to no reaction. Listeners first rated how much they liked or disliked the music (*valence*). Depending on the valence score, they reported their arousal score (Figure 2). For positive reactions, the clip was scored on an engaging-soothing scale, while negative reactions prompted an annoying-boring score. Listeners reporting a valence score of 0 could proceed to either of the two following arousal scales. The valence (*like/dislike*) score was mapped to the x-axis, while arousal (*engaging/soothing* or *annoying/boring*) ratings were mapped to the y-axis. The distance between the coordinates related to initial and final reaction were stored as vectors. The resulting data reflects correlations between music parameter changes and affective response.

**DATA ANALYSIS**

The experiment provided two levels of data interpretation, regarding affective states and musical parameters.

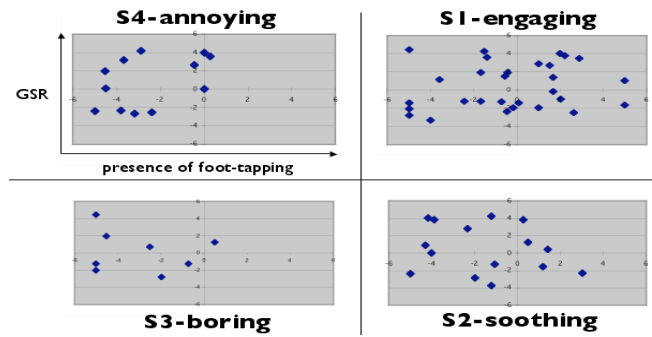


Figure 3. Affective distribution of foot-tapping and GSR

Foot-tapping presence was sampled every two seconds as a 1-bit representation of the similarity between the listener’s foot-tapping tempo and the music tempo. Each music example was split into two, the first half as initial affective state, and the second half as induced (final) state. For each section, foot-tapping presence and GSR averages were calculated, and normalized to the range [5, -5]. This data was plotted for all four affective states (Figure 3).

As shown in Figure 3, if the listener reports a boring/annoying mood, they tap less than their overall average. This suggests that valence can be disambiguated based on a listener’s body movements (Figure 4).

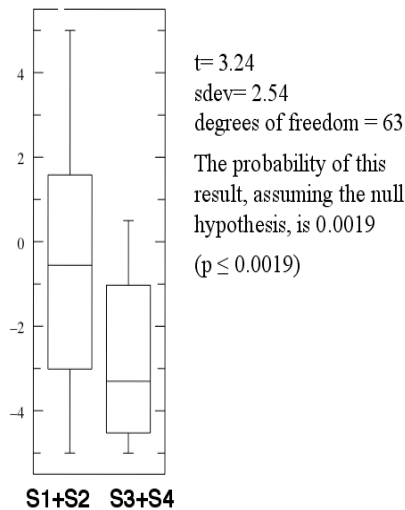
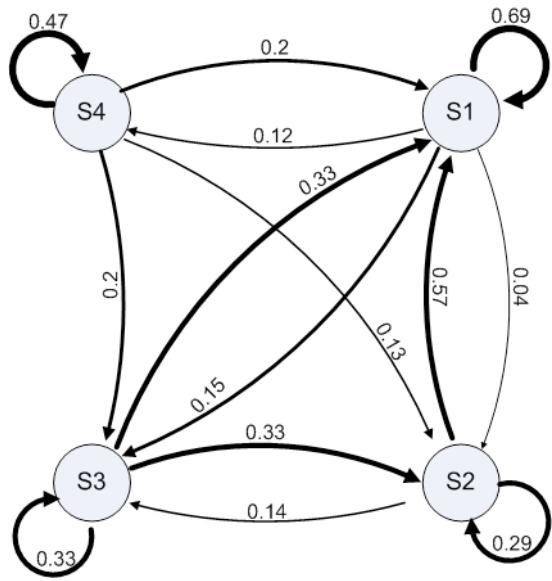
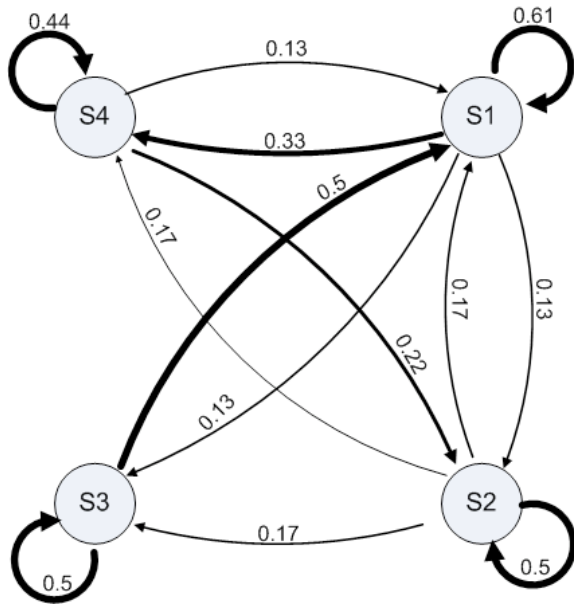


Figure 4. Statistical significance of foot-tapping between positive and negative valence

Initial and final GSR means were compared and related to self-reported state transitions. The relationship between transition probabilities of rising/falling GSR and user’s self-reported affective states appeared inconclusive (Figure 5).



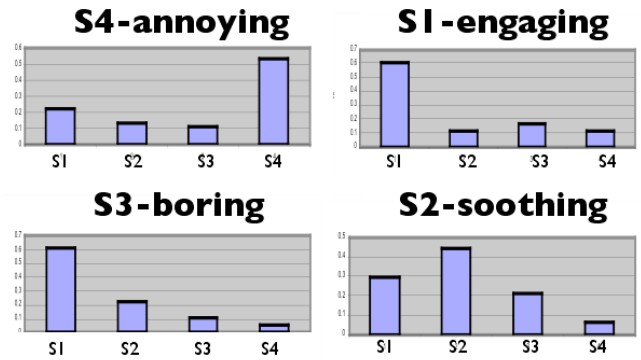
(a) GSR falling



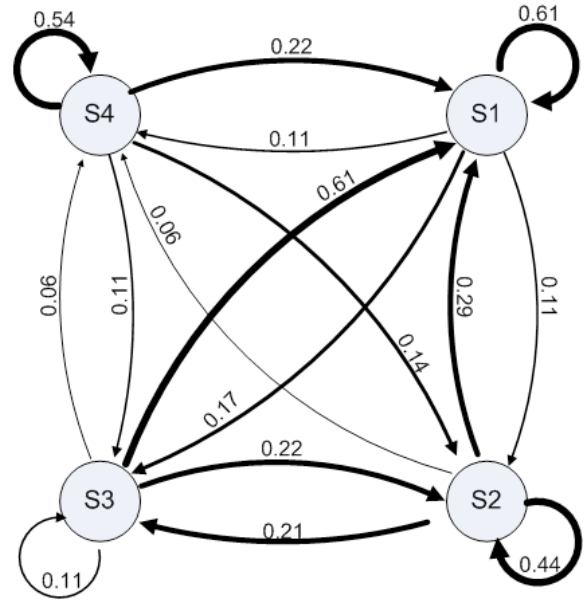
(b) GSR rising

**Figure 5. State transition chain of GSR falling and rising: S1-Engaging, S2-Soothing, S3-Boring, S4-Annoying**

Through the survey, we tried to determine which of the four actions (increase/decrease layering, increase/decrease complexity) would be most likely to induce a goal state based on the current affective state. For each action, we derived a state transition probability distribution table based on the corresponding listener data.



(a) State-transition probabilities



(b) State-transition Markov chain:  
S1-Engaging, S2-Soothing, S3-Boring, S4-Annoying

**Figure 6. Summary of state-transition probabilities**

Figure 6 shows overall trends in listener state transitions. Intuitively, we would expect positively-valenced music to continue to elicit a positive reaction as the music develops. Similarly, we expect the inverse to be true. The current data shows an exception in the case of music classified as boring. Listeners who found the music boring had a greater chance of becoming engaged than of staying bored.

The state transition summary suggests the existence of boundaries between affective states. Because of the differences in boundary thicknesses between affective states, the best path to a goal state may involve transitional states.

### STATE TRANSITION MODELS

The Affective Remixer uses a state transition model in order to move towards a target state. The real-time arranger navigates through a pre-arranged matrix of music organized by the parameters *layering* and *complexity*. Data gathered from the user (GSR, foot-tapping, and self-report) is applied

as input. To arrive at a target state, the Affective Remixer chooses a path through the music matrix that will most likely achieve the goal.

The action is appended to the current musical arrangement, new music is played for the listener. The listener's affective response to the new music is again fed back to the system.

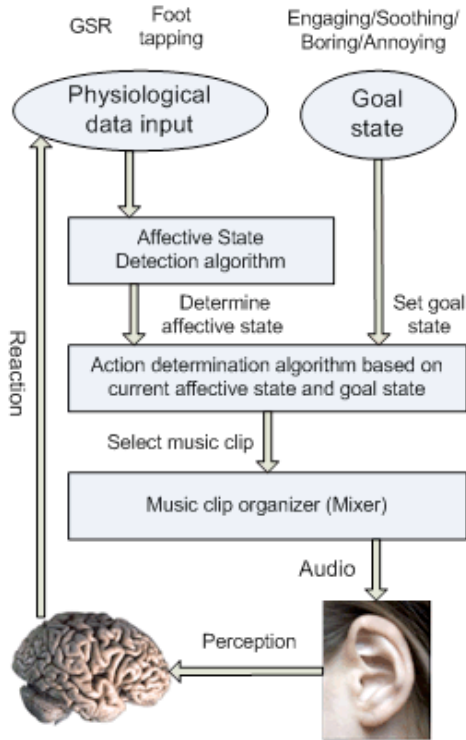


Figure 7. System diagram

### BUILDING THE MODEL

Affective state transition model consists of two components, an affect-detection model and an action-chooser model. The models are derived from data collected from subjects in the listening experiment.

#### Affect-detection model

The affect-detection model is used to analyze input signals foot-tapping and GSR, and infer the affective state of the listener. The model consists of two sets of Markov chains:

1. Markov chains that infer valence state-changes (engaging/soothing vs. annoying/boring) based on an increasing or decreasing amount of foot-tapping.
2. Markov chains that infer arousal state-changes (engaging/annoying vs. soothing/boring) based on rising or falling GSR.

The user's valence level is determined by real-time comparison between the average level of foot-tapping during the first and second halves of the music. If the average amount of foot-tapping increases in the second half of the music clip, the system uses the Markov chain associated with increased foot-tapping to determine the

subject's current valence level. Similarly, if the average foot-tapping decreases, the corresponding Markov chain is used.

The user's arousal level is determined by real-time comparison between the average GSR level during the first and second halves of the music clip. As in the valence model, two Markov chains corresponding to GSR rising/falling are constructed. If the average GSR increases, the system uses the Markov chain associated with rising GSR to infer the listener's valence. Alternatively, if the average GSR falls, the corresponding Markov chain is used. Through the processes of detecting valence and arousal, we are able to construct the model to detect affective states of a listener.

#### Action-chooser model

The action-chooser model is used to determine the action most likely to achieve the target affective state. Possible actions include increasing or decreasing the level of complexity, and increasing or decreasing the instrumental layering of the music. The appropriate action is determined based on the listener's current affective state (as determined by the affect-detection model) and a target state provided by the listener.

Based on listener data, we derived four Markov chains, each corresponding to an action in the music matrix:

1. Increasing complexity
2. Decreasing complexity
3. Increasing layering
4. Decreasing layering

Each Markov chain describes the affective-state-transition for its corresponding action. When the listener's current state is determined from the affect-detection model, state-transition probabilities (from the current state to goal states) are compared between the four action models. The action of highest probability will be selected to determine the direction the system navigates through the music matrix. The music corresponding to the chosen action will be added to the arrangement, starting at the end of the clip currently being played.

### CONCLUSION

A predictive music-arrangement system was conceived and constructed using state transition models based on collected affective response data. Final analysis of the data posed some unique questions in musical aesthetics.

For example, the data shows that if a listener becomes bored by the music, subsequently, they are most likely to become engaged. Additionally, if a listener finds the music soothing, there is very little chance that succeeding music will annoy the listener. These particular conclusions are based on data collected from the listening experiment and

are intuitive to those accustomed to thinking about music preference and emotion.

Based on the results, authors are able to provide evidence suggesting that foot-tapping is a useful indicator of a subject's valence response to music stimuli.

Examination of data collected from the experiment reveal inconsistencies in the correlation between GSR and arousal level. One possibility for the apparent discrepancy is that of serial correlation. In context of recording emotional response to music, Emery Schubert defines serial correlation as "a dependency of one data point (such as quantified, current emotional response) to another data point occurring at a previous point in time (such as a recent emotional response)." [15] Serial correlation may influence the subjects of the current study to continue liking or disliking the music depending on their initial response. While this was partly the intention of the experiment design, serial correlation is a strong force for listeners to overcome. Listeners may become too attached to their initial reaction to evaluate the second half of the music without bias. To overcome the natural bias of the previous emotional state, the experiment could be redesigned to include silence between the first and second halves of the long clips.

#### **FUTURE WORK**

To improve prediction accuracy, additional listener data might include musical preference considerations. Access to similarly-produced music by other composers would allow for a more comprehensive study.

Further investigation is required on finding correlation between musical cognitive perception and physiological reaction.

#### **ACKNOWLEDGMENTS**

The authors wish to thank all the listeners who participated, Roz Picard, Barry Vercoe, Walter Bender and Tod Machover for their valuable comments and ideas, and ITRI for their support.

#### **REFERENCES**

1. Berlyne, D. E. (1971) *Aesthetics and psychobiology*. Appleton-Century-Cofts, New York.
2. Bresin, R., and Friberg, A. (2000) Emotional Coloring of Computer-Controlled Music Performances, *Computer Music Journal*, 24:4, pp 44-63.
3. Gabrielsson, A. and Lindstrom, E. (2001) The influence of musical structure on emotional expression, in P. N

Juslin and J. A. Sloboda (eds) *Music and Emotion: Theory and Research*, pp. 223-248. Oxford University Press.

4. Healey, J., Dabek, F. and Picard, R. W. (1998) A New Affect-Perceiving Interface and Its Application to Personalized Music Selection, *Proc. from the 1998 Workshop on Perceptual User Interfaces*.
5. Herrera, P., et al. (2005) SIMAC: Semantic Interaction with Music Audio Contents, *Proceedings of the 2<sup>nd</sup> European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies*.
6. Juslin, P. N. (1997) Perceived emotional expression in synthesized performances of a short melody: Capturing the listener's judgement policy, *Musicae Scientiae*, 1:225-256.
7. Kim, S., and André, E. (2004) A Generate and Sense Approach to Automated Music Composition, *Proc. of the 9<sup>th</sup> International Conference on Intelligent User Interfaces*, pp. 268-270.
8. North, A. C. and Hargreaves, D. J. (1997) Liking, arousal potential and the emotions expressed by music, *Scandinavian Journal of Psychology* 38:47.
9. Picard, R. W. and Scheirer, J. (2001) The Galvactivator: A glove that senses and communicates skin conductivity, *Proc. from the 9<sup>th</sup> International Conference on Human-Computer Interaction*.
10. Picard, R. W. (1997) *Affective Computing*. MIT Press.
11. Russell, J. A. (1980) A Circumplex Model of Affect, *Journal of Personality and Social Psychology*, 39:6, pp. 1161-1178, American Psychological Association.
12. Schachter, S. and Singer, J. E. (1962) Cognitive, Social and Physiological Determinants of Emotional State, *Psychological Review* 69(5): 379-99.
13. Scherer, K. R. and Zentner, M. R. (2003) Music production rules, in P. N Juslin and J. A. Sloboda (eds) *Music and Emotion: Theory and Research*, pp. 361-392. Oxford University Press.
14. Schubert, E. (2001) Continuous measurement of self-report emotional response to music, in P. N Juslin and J.A. Sloboda (eds) *Music and Emotion: Theory and Research*, pp. 77-104. Oxford University Press.
15. Schubert, E. (2004) Research in expressing continuous emotional response to music as a function of its psychoacoustic parameters: Current and future directions. *Proceedings of the 18<sup>th</sup> International Congress on Acoustics*.