DIGITAL AUDIO EMOTIONS - AN OVERVIEW OF COMPUTER ANALYSIS AND SYNTHESIS OF EMOTIONAL EXPRESSION IN MUSIC

Anders Friberg

Speech, Music and Hearing,
CSC, KTH
Stockholm, Sweden
afriberg@kth.se

ABSTRACT

The research in emotions and music has increased substantially recently. Emotional expression is one of the most important aspects of music and has been shown to be reliably communicated to the listener given a restricted set of emotion categories. From the results it is evident that automatic analysis and synthesis systems can be constructed. In this paper general aspects are discussed with respect to analysis and synthesis of emotional expression and prototype applications are described.

1. INTRODUCTION

When you ask people what they think is the most important aspect of music the answer is often its ability to express and invoke emotions (e.g. [1]). At first thought it may be surprising that we are so sensitive to sound sequences in form of music. We even attribute one tone played on a piano to different emotional expressions [2]. This is in fact similar to how we attribute meaning to simple visual moving objects [3]. Sound is a major carrier of information in speech as well as for environmental motion, such as moving objects, animals or people. Therefore, it is plausible that the same kind of processing applies also to the more orderly organized music. Still, the current research has not yet solved the most fundamental question – why is music so interesting? The currently dominating theory of music perception is that we learn common sound patterns by statistical learning. In fact, David Huron recently proposed a theory explaining how emotional reactions can be triggered by violations from the expected sound sequences [4].

The research on the analysis of emotional expressions in music has a long history. The first empirical studies even started in the 19th century. For a comprehensive overview see Gabrielsson and Lindström [5]. Kate Hevner made in the 1930s a series of experiments in which systematically varied compositions were performed for subjects which rated the perceived emotional expression. In this way she could relate the features to the emotional expression. The description of emotional expressions in terms of musical features has been one of the major goals in the subsequent research.

Juslin and Laukka [6] made a meta-analysis of 41 articles studying emotional expression in music performance and ca 104 articles studying emotions in speech. An attempt was made to summarize the musical performance and vocal features according to five different emotions. Thus even though emotional communication might be a difficult research area a large number of studies points in the same direction. If we try to summarize we see that

1. Emotional expression can be reliably communicated from performer to listener

2. Up to 80-90% of the listeners' answers can be predicted using models based on musical features.

3. Despite different semantic sets, the four emotions sadness, happiness, anger, and love/tenderness (including synonyms) seem to be the ones that are especially easy to differentiate, describe and model.

It is important to note that these results mostly concern the *perceived emotion*, that is, what the listener perceives is expressed in the music. The *induced emotion*, that is, what the listener feel, is a more complex and difficult research challenge that only recently has been approached.

Given this impressive research tradition in music and emotions it is surprising to see that very few attempts has been made to make computational models, in particular starting from audio recordings. Similar tasks, for example, predicting musical genre, has a long tradition in the Music Information Retrieval (MIR) research area. However, emotional expression has only very recently been approached in the MIR community; searching in 653 papers from the ISMIR proceedings two includes "emotion" and eight papers include "mood" in the title, most of them from the last two conferences.

We will in the following first discuss general aspects of modeling analysis/synthesis of emotional expression and conclude with application prototypes.

2. WHICH EMOTIONS ARE RELEVANT IN MUSIC?

One possibility is to adopt the more general research about emotions to the musical domain. This is non-trivial since there are many different theories and approaches in emotion research. A common approach is to use a limited set of discrete emotions. A common set is the so called basic emotions. There is not one set of basic emotions but *Happiness, Anger, Sadness, Fear, and Tenderness* has been used in a number of studies. In the summary by Juslin and Laukka [6] the 145 articles were summarized using these five general emotion categories. Although they have been criticized for oversimplifying the musical experience, it has been successfully shown that these emotions can be distinguished both by performers and listeners. Possibly, they are better suited for describing perceived rather than induced emotions.

Another approach is to express emotions in a two dimensional space with *activity* as one dimension and *valence* as the other dimension. Activity is the associated energy and valence is the positive or negative connotation of the emotion. Russell [7] showed that most discrete emotions will be positioned at specific points in the space forming a circle. An interesting coupling between the activity-valence space and the discrete emotions can be made. Happiness, anger, sadness and tenderness can be used for representing each quadrant in the activity-valence space. A possible extension to the dimensional approach is to use three dimensions. For example, Leman et al. started with a large number of emotion labels and applied multidimensional scaling. It resulted in three distinct major categories that they interpreted as *valence*, *activity* and *interest.*

Are musical emotions special? The large number of successful studies indicate that the basic emotions as well as the two-dimensional space seems to work well for describing perceived emotions. For induced emotions the picture is less clear. In general, induced emotions are positive – even if a "sad" piece is played and you start to cry often your experience is positive. However, of the five basic emotions above three are negative and half of the activity-valence space is negative.

Hevner [8], [9] presented a set of eight emotion categories specifically chosen for describing music. The most important adjective in each group were *dignified, sad, dreamy, serene, graceful, happy, exciting,* and *vigorous*. One might assume that this set was developed primarily for classical music. However, there are different kinds of genres possibly each with their own palette of expression and communicative purpose. Recently, in a free-labeling study concerning scratch music Hansen and Friberg [10] found that one of the most common labels were *cool*, a rather unlikely description of classical music.

3. ANALYSIS MODELS

Here we will concentrate on automatic analysis of audio or MIDI data thus not considering treatment of meta-data. The common basic paradigm is rather simple. The purpose of analysis models is usually to predict emotional expression from the musical surface being either symbolic data (MIDI) or audio data. This is done in two steps. First, a number of features (or cues) are extracted from the musical surface and secondly, these features are combined for predicting the emotion.

3.1. Mapping features to emotions

The analysis has until recently mainly been carried out by psychologists. The methods have been the traditional statistical methods such as multiple regression analysis (MRA) and analysis of variance (ANOVA). A typical method is to have listeners rate the emotional expression in a set of performances, extract some relevant features, and then apply multiple regression to predict the ratings. MRA is essentially a simple linear combination of features with weights for each feature. The advantage is that its statistical properties are thoroughly investigated [11]. Thus, a relevance measure for each feature can be obtained (e.g. beta weights) and there are various methods for feature selection and feature interaction. An interesting extension using this method is the lens model by Juslin [12], see Figure 1. It is modeling both how the performers are combining the features for expressing different emotions and how the listeners combine the features in decoding the emotion. MRA is used twice for quantifying these relations. In addition, general measures of communication from performer to listener are defined.

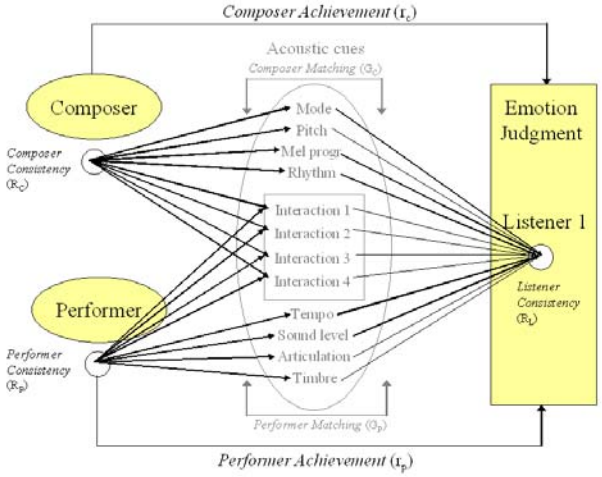

Figure 1: The extended lens model by Juslin in which the communication from composer/performer is modeled using MRA in both directions from the features (cues in the middle) (from [13]).

One limitation of MRA is its linear behavior. It implies that a feature will have a significant effect (or prediction power) only if the feature values are relatively high or low for a certain emotion in comparison with the other emotions. A typical case is tempo. There is some evidence that the tempo in a happy expression should be in an intermediate range (see [14]). If we assume that the tempo should be fast for anger and slow for sadness, the tempo feature will not be significant in a MRA that is predicting a happy rating. To overcome this, we can first transform the features by, for example, using fuzzy regions [15] or by fitting gaussians [16] and then apply a multiple regression.

Obviously, there are a multitude of more advanced prediction methods available from the field of data-mining. Predicting emotional expression from musical features is a priori not different from any other prediction of high-level perceptual/cognitive musical concepts from musical features. Thus, one can use any of the methods, such as Neural Networks, Hidden Markov Models, Bayesian modeling, or Support Vector Machines [17]. These methods are typically used within the field of music information retrieval (MIR) for detecting e.g. musical genre.

Common for these methods (including MRA) is that they usually are data-driven, that is, it is necessary to assemble databases with human annotated emotional labels and to test and optimize the model using this "ground-truth" data. An alternate approach is to directly use the quantitative data provided in the numerous previous studies. A simple real-time model for predicting anger, happiness and sadness in either audio or gestures was developed using fuzzy functions in with each feature was divided into three regions; low, medium, and high [15]. A selection of these regions was then combined for each emotion. For example, sadness was predicted by low sound level, low tempo, and legato articulation, see Figure 2.

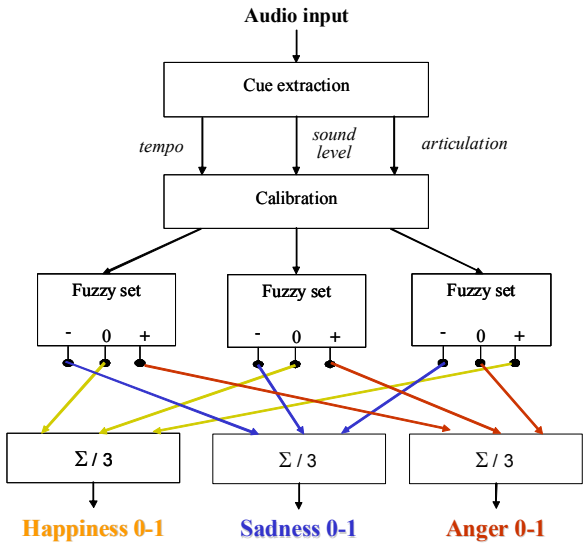

Figure 2: *Fuzzy mapper of emotional expression in music (from [15]).*

Is emotion recognition a classification task? As shown in previous research (e.g. [12]) the emotional expression can be of different strength and different emotions can exist at the same time. On the other hand, perception is often categorical. Therefore, either a classification or a gradual prediction of emotion response (such as MRA) can be appropriate depending on the practical use of the model.

3.2. Which features?

In score-based music there are two independent sources of the final emotional expression, namely the composer and the performer. Therefore it is convenient to divide the features into *performance features* and *score features*. The performance features are relatively easier to summarize and has been thoroughly investigated in many studies. The following are the most important performance features:

Timing - Tempo, tempo variation, duration contrast
Dynamics: overall level, crescendo/decrescendo, accents
Articulation: overall (staccato/legato), variability
Timbre: Spectral richness, onset velocity

The score features are more complex and harder to describe. This is not surprising given the endless possibilities of combining notes and that we extract complex perceptual concepts and patterns, such as harmony, key, and meter, out of the musical surface. The traditional music-theoretic measures, such as harmonic function, seem to be less important for emotional expression. From the summary by Gabrielsson and Lindström [5] we obtain the following list of the most important score features (omitting the performance features listed above):

Pitch (high/low)
Interval (small/large)
Melody: range (small/large), direction (up/down)
Harmony (consonant/complex-dissonant)
Tonality (chromatic-atonal/key-oriented)
Rhythm (regular-smooth/firm/flowing-fluent/irregular-rough)
Timbre (harmonic richness)

These are rather general and imprecise score features that often have been rated by experts in previous experiments. Lately, several additions have been suggested such as number of note onsets, as well as many different spectral features. The good news with these features is that we don't need to transcribe the audio recording into notes and then predict and classify voices, harmony and meter. If we take the example of harmony, we see that a measure of harmonic complexity would possibly be better than the exact harmonic analysis of the piece. Since these features already have been shown to be important for emotion communication, one approach is to develop automatic feature extractions that predict these qualitative measures according to human experts.

Most of the existing studies have used a subset of these features often starting with features developed for other purposes, such as genre classification. Leman et al. [18] used a large number of low-level features developed for auditory hearing models. Lu et al. [19] partly developed their own features trying to approximate some of the features above and obtained a relatively good accuracy. Rather than exploring advanced mapping models it appears that the most important improvement can be obtained by a further development of the relevant features. In particular, these features need to be evaluated individually so that they correspond to the perceptual counterpart. Such work has recently started with the feature extraction methods developed within the MIRToolbox[1] by the University of Jyväskylä [20].

Possibly the most complete analysis of emotional expression from audio files was done by Lu et al. [19]. They recognized the need for specific features, they used the simple and common four emotions categorizing each quadrant in the Activity-Valence space, and in addition, developed a boundary detection for determining when the emotional expression changes. The obtained average emotion detection accuracy was about 86% using a set of classical recordings.

4. SYNTHESIS MODELS

Most analysis experiments have used music examples played by musicians. Musicians are highly trained to perform music in a learned way partly using internalized subconscious knowledge. For example, even when a musician is asked to play a piece "deadpan" that is without any performance variations, still typical phrasing patters will occur, although of much lower amount. This makes it impossible to fully isolate the impact of each feature on the emotional expression using musicians. In order to do this, the best method is to synthesize the music with independent control of all features [21]. Manipulation of performance features as listed above is rather simple task if MIDI scores are used. The resulting performances can be rather convincing in terms of emotional character. However, the resulting musical quality is often

[1] www.jyu.fi/music/coe/materials/mirtoolbox

low since typical performance principles such as phrasing will be missing. Thus, one possibility is to use the KTH rule system that contains a number of principles musicians use for conveying the musical structure [22]. Bresin and Friberg [23] showed that six different emotions and a neutral expression could be successfully communicated using general features such as tempo but also using a set of six different rules such as Phrase arch and Duration contrast. Juslin et al. [24] manipulated systematically four different music performance aspects including the emotion dimension using the rule system with additional performance principles. Currently, an extension to the KTH rule system is in progress. The goal is to use the rule system and directly manipulate a recorded audio file regarding tempo, sound level and articulation [25]. A suggestion of qualitative values of general performance features and performance rules are shown in Table 1.

Table 1: *Suggested qualitative values for changing the emotional expression in synthesizing music performance (from [22]).*

	Happy	**Sad**	**Angry**	**Tender**
Overall changes				
Tempo	Somewhat fast	slow	fast	slow
Sound level	medium	low	high	low
Articulation	staccato	legato	Somewhat staccato	legato
Rules				
Phrase arch	small	large	negative	small
Final ritardando	small	-	-	small
Punctuation	large	small	medium	small
Duration contrast	large	negative	large	-

The emotion synthesis described above only manipulates performance features. A challenging task is to also vary the score features while at the same time keep the musical quality at a decent level. Using a precomposed piece, still a few of the score features such as timbre and pitch can be manipulated without altering the composition.

4.1. Applications

An obvious application of an emotion analyzer would be to include it in a music browsing system. There are a few public systems running already, like Musicovery[2] that let the user select music according to position in the Activity-Valence space. These systems rely currently on meta-data entered by experts or users. However, commercial systems including automatic feature analysis are likely to be released in the near future.

The lens model by Juslin (see above), was applied in the Feel-ME system for teaching emotional expression [26]. During a session, the student is asked to perform the same melody a number of times with different emotional expressions, the program is analyzing the used performance features in relation to a fictive listening panel, and finally the programs gives explicit feedback for each feature how to improve the communication of the emotional expression. It was shown that the program was more effective at teaching emotional expression than a regular music teacher.

The fuzzy mapper in Figure 2 has been used in several experimental applications at KTH. The Expressiball, developed by Roberto Bresin [27] is a visual feedback of a number of performance parameters including the emotion expression. A virtual ball on the screen moves and changes color and shape in real time according to the audio input. In a similar application, the visual output was instead a virtual head that changed facial expression according to the input expression [28]. The fuzzy mapper was also used in the collaborative game Ghost in the Cave [29]. One task in the game was to express different emotions either with the body or the voice.

One possible commercial application of synthesizing emotions is within computer games. A main function of computer games is obviously that the whole visual scenario changes interactively according to user actions. However, often the music still consists of prerecorded sequences. This has been recognized for some time in the game community but still there are few commercial alternatives. As music is often used to manipulate the mood of the audience in both film and computer games, an interactive mood control of the music would fit perfectly into most computer games.

As mentioned above the KTH rule system can be used for manipulating the emotional expression of a performance. Within the pDM program [30] the rules can be controlled in real time. Different 2-dimensional spaces, such as the Activity-Valence space can be used for meta-control of the rules. As an extension a "Home conducting" system was suggested that used expressive gestures analyzed by a video camera for controlling the emotional expression of the music [31]. There is first an analysis of the gesture going from low-level video features to emotion features and then the process is reversed for synthesizing the performance, see Figure 3.

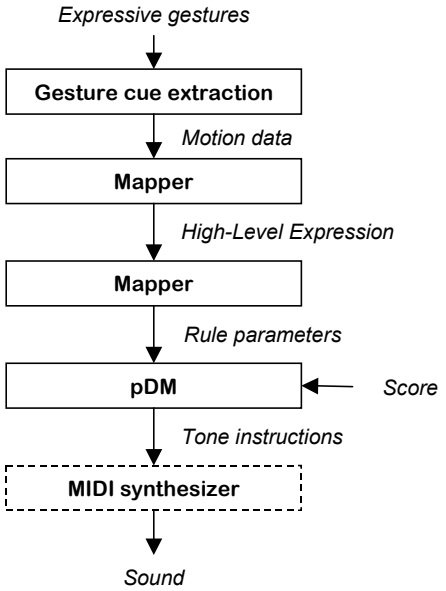

Figure 3: *An overview of the analysis/synthesis step in the Home conducting system (from [31].*

[2] www.musicovery.com

An alternative rule system for emotional expression was recently developed by Livingstone [32]. It also uses the four quadrants of the Activity-Valence space for control of the emotional expression.

These systems only use performance parameters to convey different emotions. For a more effective system it is necessary to also manipulate score features. This is particularly relevant if something else than classical music is used. In a commercial system for modifying ringtones, we used the KTH rule system together with changes in timbre and pitch (octave transpositions) to enhance the effect in popular songs[3]. Winter [33] used the pDM rules system and added timbre, transposition, harmonic complexity and rhythmic complexity in pre-composed pop songs.

5. CONCLUSION AND FUTURE POSSIBILITIES

Numerous experiments have showed that it is relatively easy to convey different emotions from performer to listener. There is not a general agreement on which emotion categories/dimensions that best describe the space of musical expression and communication. However, it seems that it is particularly easy and straightforward to use the four categories happiness, anger, sadness, and love/tenderness. They also happen to characterize each quadrant in the Activity-Valence space, thus, unifying the discrete and dimensional approach.

For developing an emotion analysis system working on ordinary audio files the most important aspect seems to be to develop mid-level/high-level features corresponding to relevant perceptual concepts. This would also be useful for analyzing other aspects such a musical style.

Emotional expression might be a particularly rewarding applied research field in the near future. The reasons are that (1) a substantial bulk of basic research has already been carried out with promising results, and (2) that emotional expression is a simple and natural way to describe and perceive the musical character even for inexperienced listeners. There is a strong commercial potential both for analysis and synthesis of emotional expression within the music and computer game industry.

6. REFERENCES

[1] P.N. Juslin and J. Laukka, “Expression, Perception, and Induction of Musical Emotions: A Review and a Questionnaire Study of Everyday Listening,” *Journal of New Music Research,* vol. 33, no. 3, pp. 217-38, 2004.

[2] F. Bonini Baraldi, G. De Poli and A. Roda`, "Communicating Expressive Intentions with a Single Piano Note", *Journal of New Music Research*, vol. 35, n. 3, pp. 197 - 210, 2006

[3] S. Coren, L.M. Ward, and J.T. Enns, *Sensation and perception*, Wileys, 2003.

[4] D. Huron, S*weet Anticipation: Music and the Psychology of Expectation*, Cambridge, Massachusetts: MIT Press, 2006.

[5] A. Gabrielsson and E. Lindström, ”The influence of musical structure on emotional expression. In P. N. Juslin, & J. A. Sloboda (Eds.), *Music and emotion: Theory and Research* New York: Oxford University Press, 2001, pp. 223-248.

[6] P.N. Juslin and J. Laukka, “Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?” *Psychological Bulletin*, vol. 129, no. 5, pp. 770-814, 2003.

[7] J. A. Russell, “A circumplex model of affect,” *Journal of personality and social psychology*, vol. 39, pp. 1161 – 1178 1980.

[8] K. Hevner, “Experimental studies of the elements of expression in music,” *American Journal of Psychology*, vol. 89, pp. 246-68, 1936.

[9] K. Hevner, The affective value of pitch and tempo in music. *American Journal of Psychology*, vol. 49, pp. 621-30, 1937.

[10] K.F. Hansen and A. Friberg, “Verbal descriptions and emotional labels for expressive DJ performances,” manuscript submitted for publication, 2008.

[11] J. Cohen, P. Cohen, S.G. West and L.S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edition, London: LEA, 2003.

[12] P.N. Juslin, “Cue utilization in communication of emotion in music performance: Relating performance to perception,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 26, pp. 1797-1813, 2000.

[13] P.N. Juslin, “From mimesis to catharsis: expression, perception, and induction of emotion in music,” In D. Miell, R. MacDonald, & D. J. Hargreaves (Eds.), *Musical communication* New York: Oxford University Press, 2005, pp. 85-115.

[14] P.N. Juslin and J. Laukka, “Communication of Emotions in Vocal Expression and Music Performance: Different Channels, Same Code?” *Psychological Bulletin*, vol. 129, no. 5, pp. 770-814, 2003.

[15] A. Friberg, “A fuzzy analyzer of emotional expression in music performance and body motion,” In J. Sundberg & B. Brunson (Eds.) *Proceedings of Music and Music Science, Stockholm, October 28-30, 2004, Royal College of Music in Stockholm*, 2005.

[16] A: Friberg, and S. Ahlbäck, “Recognition of the main melody in a polyphonic symbolic score using perceptual knowledge,” Manuscript in preparation.

[17] T. Li and M. Ogihara, “Detecting emotion in music,” In *Proceedings of the Fifth International Symposium on Music Information Retrieval*, pp. 239-240, 2003.

[18] M. Leman, V. Vermeulen, L. De Voogdt, and D. Moelants, “Prediction of Musical Affect Attribution Using a Combination of Structural Cues Extracted From Musical Audio,” *Journal of New Music Research*, vol. 34 no. 1, 2005.

[19] L. Lu, D. Liu, and H. Zhang, “Automatic Mood Detection and Tracking of Music Audio Signals”, *IEEE Transaction on Audio, Speech, and Language Processing*, vol. 14, no. 1, 2006.

[20] O. Lartillot and P. Toiviainen, "A Matlab Toolbox for Musical Feature Extraction From Audio," *International Conference on Digital Audio Effects*, Bordeaux, 2007.

[21] P.N. Juslin, “Perceived emotional expression in synthesized performances,” *Musicae Scientiae*, vol. 1, no 2, pp. 225-256, 1997.

[22] A: Friberg, R. Bresin and J. Sundberg, “Overview of the KTH rule system for musical performance,” *Advances in Cognitive Psychology, Special Issue on Music Performance*, vol. 2, no. 2-3, pp. 145-161, 2006.

[3] www.notesenses.com

[23] R. Bresin, and A. Friberg, "Emotional Coloring of Computer-Controlled Music Performances," *Computer Music Journal*, vol. 24, no. 4, pp. 44-63, 2000.

[24] P.N. Juslin, A. Friberg, and R. Bresin, "Toward a computational model of expression in performance: The GERM model." *Musicae Scientiae special issue 2001-2002*. pp. 63-122, 2002.

[25] M. Fabiani, and A. Friberg, "A prototype system for rule-based expressive modifications of audio recordings," In *Proc. of the Int. Symp. on Performance Science 2007* Porto, Portugal: AEC (European Conservatories Association), 2007, pp. 355-360.

[26] P.N. Juslin, J. Karlsson, E. Lindström, A. Friberg, and E. Schoonderwaldt, "Play it again with a feeling: Feedback-learning of musical expressivity," *Journal of Experimental Psychology: Applied*, Vol. 12, no. 2, pp. 79-95, 2006.

[27] A. Friberg, E. Schoonderwaldt, P.N. Juslin and R. Bresin, "Automatic Real-Time Extraction of Musical Expression," in *Proceedings of the International Computer Music Conference 2002* , San Francisco: International Computer Music Association, 2002, pp. 365-367.

[28] M. Mancini, R. Bresin, and C. Pelachaud, "A virtual head driven by music expressivity," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 6, pp. 1833-1841, 2007.

[29] M.-L. Rinman, A. Friberg, B. Bendiksen, D. Cirotteau, S. Dahl, I. Kjellmo, B. Mazzarino and A. Camurri, "Ghost in the Cave - an interactive collaborative game using non-verbal communication," in A. Camurri, G. Volpe (Eds.), *Gesture-based Communication in Human-Computer Interaction*, LNAI 2915, Berlin: Springer Verlag, 2004, pp. 549–556.

[30] A. Friberg, "pDM: an expressive sequencer with real-time control of the KTH music performance rules," *Computer Music Journal*, vol. 30, no. 1, pp. 37-48, 2006.

[31] A. Friberg, "Home conducting – Control the overall musical expression with gestures," *Proceedings of the 2005 International Computer Music Conference*, San Francisco: International Computer Music Association, 2005, pp. 479-482.

[32] S.R. Livingstone, *Changing Musical Emotion through Score and Performance with a Computational Rule System*, Doctoral dissertation, 2008.

[33] R. Winter, *Interactive Music: Compositional Techniques for Communicating Different Emotional Qualities*, Master thesis at Speech, Music and Hearing, KTH, 2006.