THEORETICAL FRAMEWORK OF A COMPUTATIONAL MODEL OF AUDITORY MEMORY FOR MUSIC EMOTION RECOGNITION

Marcelo Caetano
Sound and Music Computing Group
INESC TEC, Porto, Portugal
mcaetano@inesctec.pt

Frans Wiering
Dep. Information and Computing Sciences
Utrecht University, The Netherlands
f.wiering@uu.nl

ABSTRACT

The bag of frames (BOF) approach commonly used in music emotion recognition (MER) has several limitations. The *semantic gap* is believed to be responsible for the *glass ceiling* on the performance of BOF MER systems. However, there are hardly any alternative proposals to address it. In this article, we introduce the theoretical framework of a computational model of auditory memory that incorporates temporal information into MER systems. We advocate that the organization of auditory memory places time at the core of the link between musical meaning and musical emotions. The main goal is to motivate MER researchers to develop an improved class of systems capable of overcoming the limitations of the BOF approach and coping with the inherent complexity of musical emotions.

1. INTRODUCTION

In the literature, the aim of music emotion recognition (MER) is commonly said to be the development of systems to automatically estimate listeners' emotional response to music [2, 7, 8, 11, 18, 19, 33] or simply to organize or classify music in terms of emotional content [14, 17]. Applications of MER range from managing music libraries and music recommendation systems to movies, musicals, advertising, games, and even music therapy, music education, and music composition [11]. Possibly inspired by automatic music genre classification [28, 29], a typical approach to MER categorizes emotions into a number of classes and applies machine learning techniques to train a classifier and compare the results against human annotations, considered the "ground truth" [14, 19, 28, 32]. Kim *et. al* [14] presented a thorough state-of-the-art review, exploring a wide range of research in MER systems, focusing particularly on methods that use textual information (e.g., websites, tags, and lyrics) and content-based approaches, as well as systems combining multiple feature domains (e.g., features plus text). Commonly, music features are estimated from the audio and used to represent the music. These features are calculated independently from each other and from their temporal progression, resulting in the bag of frames (BOF) [11, 14] paradigm.

The 'Audio Mood Classification' (AMC) task in MIREX [5, 10] epitomizes the BOF approach to MER, presenting systems whose performance range from 25 % to 70 % (see Table 1). Present efforts in MER typically concentrate on the machine learning algorithm that performs the map in an attempt to break the 'glass ceiling' [1] thought to limit system performance. The perceived musical information that does not seem to be contained in the audio even though listeners agree about its existence, called 'semantic gap' [3, 31], is considered to be the cause of the 'glass ceiling.' However, the current approach to MER has been the subject of criticism [2, 11, 28, 31].

Knowledge about music cognition, music psychology, and musicology is seldom explored in MER. It is widely known that musical experience involves more than mere processing of music features. Music happens essentially in the brain [31], so we need to take the cognitive mechanisms involved in processing musical information into account if we want to be able to model people's emotional response to music. Among the cognitive processes involved in listening to music, memory plays a major role [27]. Music is intrinsically temporal, and time is experienced through memory. Studies [12, 16, 25] suggest that the temporal evolution of the musical features is intrinsically linked to listeners' emotional response to music.

In this article, we speculate that the so called 'semantic gap' [3] is a mere reflection of how the BOF approach misrepresents both the listener and musical experience. Our goal is not to review MER, but rather emphasize the limitations of the BOF approach and propose an alternative model that relies on the organization of auditory memory to exploit temporal information from the succession of musical sounds. For example, BOF MER systems typically encode temporal information in delta and delta-delta coefficients [1], capturing only local instantaneous temporal variations of the feature values. In a previous work [2], we discussed different MER systems that exploit temporal information differently. Here, we take a step further and propose the theoretical framework of a computational model of auditory memory for MER . Our aim is to motivate MER research to bridge the 'semantic gap' and break the so called 'glass ceiling' [1, 3, 31].

The next section discusses the complexity of musical emotions and how this relates to the glass ceiling preventing BOF MER systems to improve their performance as a motivation for proposing a paradigm change. Then, we briefly introduce the model of auditory memory adopted, followed by the proposed framework and considerations about its implementation. Finally, we present the conclusions and discuss future directions of this theoretical work.

2. MACHINE LEARNING AND MER

It is generally agreed that music conveys and evokes emotions [9, 13]. In other words, listeners might feel happy listening to a piece or simply perceive it as happy [9]. Research on music and emotions usually investigates the musical factors involved in the process as well as listeners' response to music. There are many unanswered questions [13, 21], such as "which emotions does music express?", "in what context do musical emotions occur?", "how does music express emotions?", and "which factors in music are expressive of emotions?" Researchers need to address controversial issues to investigate these questions. On the one hand, the relevant musical factors, and on the other hand, the *definition* and *measurement* of emotion.

There is evidence [13] of emotional reactions to music in terms of various subcomponents, such as *subjective feeling, psychophysiology, brain activation, emotional expression, action tendency, emotion regulation* and these, in turn, feature different psychological mechanisms like *brain stem reflexes, evaluative conditioning, emotional contagion, visual imagery, episodic memory, rhythmic entrainment,* and *musical expectancy*. Each mechanism is responsive to its own combination of information in the music, the listener, and the situation. Among the causal factors that potentially affect listeners' emotional response to music are *personal, situational,* and *musical* [21]. Personal factors include age, gender, personality, musical training, music preference, and current mood; situational factors can be physical such as acoustic and visual conditions, time and place, or social such as type of audience, and occasion. Musical factors include genre, style, key, tuning, orchestration, among many others.

Most modern emotion theorists suggest that an emotion episode consists of coordinated changes in three major reaction components: physiological arousal, motor expression, and subjective feeling (the emotion triad). According to this *componential approach to emotion*, we would need to measure physiological changes, facial and vocal expression, as well as gestures and posture along with self reported feelings using a rich emotion vocabulary to estimate the listeners' emotional response. In MER, the emotional response to music is commonly collected as self-reported annotations for each music track, capturing "subjective feelings" associated or experienced by the listener. Some researchers [9] speculate that musical sounds can effectively cause emotional reactions (via *brain stem reflex*, for example), suggesting that certain music dimensions and qualities communicate similar affective experiences to many listeners. The literature on the emotional effects of music [9, 13] has accumulated evidence that listeners often agree about the emotions expressed (or elicited) by a particular piece, suggesting that there are aspects in music that can be associated with similar emotional responses across cultures, personal bias or preferences.

It is probably impractical to hope to develop a MER system that could account for all facets of this complex problem. There is no universally accepted model or explanation for the relationship between music and emotions. However, we point out that it is widely known and accepted that MER systems oversimplify the problem when adopting the BOF approach [11]. In this context, we propose a theoretical framework that uses the organization of auditory memory to incorporate temporal information into MER. We argue that time lies at the core of the complex relationship between music and emotions and that auditory memory mediates the processes involved. In what follows, we focus on the link between musical sounds and self-reported subjective feelings associated to them through music listening. In other words, the association between the audio features and perceived emotions.

2.1 The Glass Ceiling on System Performance

The performance of music information retrieval (MIR) systems hasn't improved satisfactorily over the years [1, 10] due to several shortcomings. Aucouturier and Pachet [1] used the term 'glass ceiling' to suggest that there is a limitation on system performance at about 65% *R*-precision when using BOF and machine learning in music similarity. Similarly, Huq *et. al* [11] examined the limitations of the BOF approach to MER. They present the results of a systematic study trying to maximize the prediction performance of an automated MER system using machine learning. They report that none of the variations they considered leads to a substantial improvement in performance, which they interpret as a limit on what is achievable with machine learning and BOF.

MIREX [10] started in 2005 with the goal of systematically evaluating state-of-the-art MIR algorithms, promoting the development of the field, and increasing system performance by competition and (possibly) cooperation. MIREX included an "Audio Mood Classification" (AMC) task for the first time in 2007 'inspired by the growing interest in classifying music by moods, and the difficulty in the evaluation of music mood classification caused by the subjective nature of mood' [10]. MIREX's AMC task uses a categorical representation of emotions divided in five classes. These five 'mood clusters' were obtained by analyzing 'mood labels' (user tags) for popular music from the All Music Guide [1].

The MIREX wiki [2] presents the "Raw Classification Accuracy Averaged Over Three Train/Test Folds" per system. Table 1 summarizes system performance over the years for the MIREX task AMC, showing the minimum, maximum, average, and standard deviation of these values across systems. Minimum performance has steadily

[1] All Music Guide http://www.allmusic.com/

[2] http://www.music-ir.org/mirex/wiki/MIREX_HOME

Table 1: MIREX AMC performance from 2007 to 2013.

	Minimum	Maximum	Average	STD
2007	25.67%	61.50%	52.65%	11.19%
2008	30.33%	63.67%	52.39%	7.72%
2009	34.67%	65.67%	57.67%	6.17%
2010	36.12%	63.78%	56.48%	6.36%
2011	39.81%	69.48%	57.98%	9.33%
2012	46.14%	67.80%	62.67%	6.17%
2013	28.83%	67.83%	59.81%	10.29%

improved, but maximum performance presents a less significant improvement. The standard deviation of performance across systems has a general trend towards decreasing (suggesting more homogeneity over the years). Most algorithms are also tested in different classification tasks (musical genre, for example), and the best in one task are often also very good at other tasks, maybe indicating there is more machine learning than musical knowledge involved. Sturm [28] discusses the validity of the current evaluation in MER. He argues that the current paradigm of classifying music according to emotions only allows us to conclude how well an MER system can reproduce "ground truth" labels of the test data, irrespective of whether these MER systems use factors irrelevant to emotion in music.

2.2 Bridging the Semantic Gap

In MIR, audio processing manipulates signals generated by musical performance, whereas music is an abstract and intangible cultural construct. The sounds *per se* do not contain the essence of music because music exists in the mind of the listener. The very notion of a 'semantic gap' is misleading [31]. The current BOF approach to MER views music simply as data (audio signals) and therefore misrepresents musical experience. Machine learning performs a rigid map from "music features" to "emotional labels", as illustrated in part a) of Fig. 1, treating music as a stimulus that causes a specific emotional response irrespective of personal and contextual factors which are known to affect listeners' emotional response [12, 16, 25] such as listeners' *previous exposure* and the impact of the *unfolding musical process*. Memory is particularly important in the recognition of patterns that are either stored in long-term memory (LTM) from previous pieces or in short-term memory (STM) from the present piece. Music seems to be one of the most powerful cues to bring emotional experiences from memory back into awareness.

Wiggins [31] suggests to look at the literature from musicology and psychology to study the cognitive mechanisms involved in human music perception as the starting point of MIR research, particularly *musical memory*, for they define music. He argues that music is not just processed by the listeners, it is defined by them. Wiggins states that "music is a cognitive model", therefore, only cognitive models are likely to succeed in processing music in a human-like way. He writes that "to treat music in a way which is not human-like is *meaningless*, because music is *defined by humans.* Finally, he concludes that the

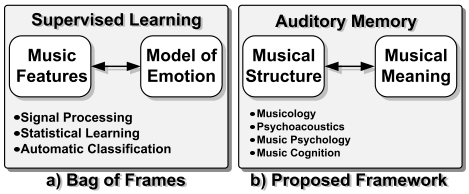

Figure 1: Approaches to MER. Part a) illustrates the BOF approach, which uses machine learning to map music features to a region of a model of emotion. Part b) illustrates the proposed approach, which relies on the organization of auditory memory to estimate musical emotions as a form of musical meaning emerging from musical structure.

human response to memory is key to understanding the *psychophysiological effect* of musical stimuli, and that this domain is often missing altogether from MIR research. In this work, we view perceived musical emotions as a particular form of musical meaning [12, 16, 25], which is intimately related to musical structure by the organization of auditory memory [27], as represented in part b) of Fig. 1.

3. AUDITORY MEMORY AND MER

Conceptually, memory can be divided into three processes [27]: sensory memory (echoic memory and early processing); short-term memory (or working memory); and long-term memory. Each of these memory processes functions on a different time scale, which can be loosely related to levels of musical experience, the "level of event fusion", the "melodic and rhythmic level", and the "formal level", respectively. Echoic memory corresponds to early processing, when the inner ear converts sounds into trains of nerve impulses that represent the frequency and amplitude of individual acoustic vibrations. During feature extraction, individual acoustic features (e.g., pitch, overtone structure) are extracted and then bound together into auditory events. The events then trigger those parts of long-term memory (LTM) activated by similar events in the past, establishing a context that takes the form of expectations, or memory of the recent past. Long-term memories that are a part of this ongoing context can persist as current "short-term memory" (STM). Short-term memories disappear from consciousness unless they are brought back into the focus of awareness repeatedly (e.g. by means of the rehearsal loop). When the information is particularly striking or novel, it may be passed back to LTM and cause modifications of similar memories already established, otherwise it is lost.

The three types of processing define three basic time scales on which musical events and patterns take place, which, in turn, affect our emotional response to music. The event fusion level of experience (echoic memory) is associated with pitch perception. The main characteristic of the melodic and rhythmic level is that separate events on this time scale are grouped together in the present as melodic grouping and rhythmic grouping, associated with STM. Units on the formal level of musical experience consist of entire sections of music and are associated with

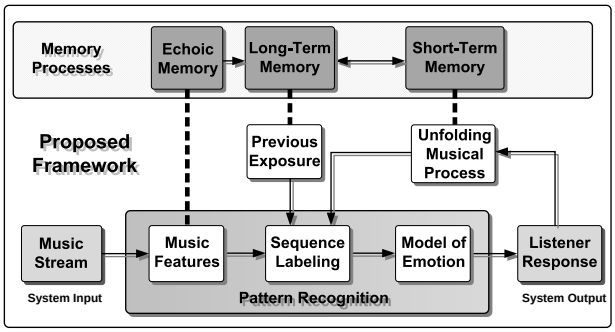

Figure 2: The proposed framework for MER. The blocks are system components and the arrows indicate the flow of information. In the shaded area is pattern recognition, and outside are the proposed processes, namely, the "unfolding musical process" and the listener's "previous exposure". The figure also illustrates how the organization of auditory memory is related to system blocks.

LTM and our previous musical exposure. Echoic memory and early processing provide our immediate experience of the present moment of music in the focus of conscious awareness, and help to segment it into manageable units; STM establishes the continuity and discontinuity of that movement with the immediate past; and LTM provides the context that gives it meaning, by relating the moment to a larger framework of ongoing experience and previous knowledge. The organization of memory and the limits of our ability to remember have a profound effect on how we perceive patterns of events and boundaries in time. Time is a key element in memory processes and should be brought to the foreground of MER [2].

4. THE PROPOSED FRAMEWORK

Fig. 2 shows the framework we propose to incorporate memory processes in MER systems to illustrate how auditory memory affects musical experience. The blocks associated with the system have a white background, while memory processes have a dark background. The arrows represent the flow of information, while the dashed line represents the relationship between memory processes and system blocks. The proposed framework can be interpreted as an extension of the traditional approach (shaded area) by including two blocks, *previous exposure* and *unfolding music process*. In the BOF approach, the music features are associated with echoic memory, related to very short temporal scales and uncorrelated with the past or predictions of future events. The framework we propose includes the "Unfolding Musical Process" and "Previous Exposure" to account for LTM and STM. The "Unfolding Musical Process" represents the listeners' perception of time (related to musical context and expectations), while "Previous Exposure" represents the personal and cultural factors that makes listeners unique.

4.1 Unfolding Musical Process

The *unfolding music process* uses temporal information from the current music stream to account for repetitions and expectations. As Fig. 2 suggests, the unfolding musical process acts as feedback loop that affects the map between the music features and the listener response. The dynamic aspect of musical emotion relates to the cognition of musical structure [12, 16, 25]. Musical emotions change over time in intensity and quality, and these emotional changes covary with changes in psycho-physiological measures [16, 25]. The human cognitive system regulates our expectations to make predictions [12]. Music (among other stimuli) influences this principle, modulating our emotions. As the music unfolds, the model is used to generate expectations, which are implicated in the experience of listening to music. Musical meaning and emotion depend on how the actual events in the music play against this background of expectations.

4.2 Previous Exposure

The framework in Fig. 2 illustrates that *previous exposure* accounts for musical events stored in LTM that affect the listeners' emotional response to music. Musical emotions may change according to musical genre [6], cultural background, musical training and exposure, mood, physiological state, personal disposition and taste [9, 12]. This information is user specific and depends on context thus it cannot be retrieved from the current music stream, rather, it has to be supplied by the listener.

5. IMPLEMENTATION ISSUES

Here we address how to treat individual components of the model, which parts need human input and which are automatic, and how the different system components communicate and what information they share. The proposed framework urges for a paradigm change in MER research rather than simply a different kind of MER systems, including representing the music stream, collecting time-stamped annotations, and system validation and evaluation [28]. Thus we propose a class of dynamic MER systems that continuously estimate how the listener's perceived emotions unfold in time from a time-varying input stream of audio features calculated from different musically related temporal levels.

5.1 Music Stream as System Input

The proposed system input is a music stream unfolding in time rather than a static (BOF) representation. To incorporate time into MER, the system should monitor the temporal evolution of the music features [25] at different time scales, the "level of event fusion", the "melodic and rhythmic level", and the "formal level". The feature vector should be calculated for every frame of the audio signal and kept as a time series (i.e., a time-varying vector of features). Time-series analysis techniques such as linear prediction and correlations (among many others) might be used to extract trends and model information at later stages.

5.2 Music Features

Eerola [6, 7] proposes to select musically relevant features that have been shown to relate to musical emotions. He presents a list of candidate features for a computational model of emotions that can be automatically estimated from the audio and that would allow meaningful annotations of the music, dividing the features into musically relevant levels related to three temporal scales. Snyder [27] describes three different temporal scales for musical events based on the limits of human perception and auditory memory. Coutinho *et. al* [4] sustain that the structure of affect elicited by music is largely dependent on dynamic temporal patterns in low-level music structural parameters. In their experiments, a significant part of the listeners' reported emotions can be predicted from a set of six psychoacoustic features, namely, loudness, pitch level, pitch contour, tempo, texture, and sharpness. Schubert [26] used loudness, tempo, melodic contour, texture, and spectral centroid as predictors in linear regression models of valence and arousal.

Fig. 1 suggests that MER systems should use the musical structure to estimate musical meaning such as emotions. Musical structure emerges from temporal patterns of music features. In other words, MER systems should include information about the rate of temporal change of music features, such as how changes in loudness correlate with the expression of emotions rather than loudness values only. These loudness variations, in turn, form patterns of repetition on a larger temporal scale related to the structure of the piece that should also be exploited. Thus the features should be hierarchically organized in a musically meaningful way according to auditory memory [27].

5.3 Listener Response and System Output

Recently, some authors started investigating how the emotional response evolves in time as the music unfolds. Krumhansl [16] proposes to collect listener's responses continuously while the music is played, recognizing that retrospective judgements are not sensitive to unfolding processes. Recording listener's emotional ratings over time as time-stamped annotations requires listeners to write down the emotional label and a time stamp as the music unfolds, a task that has received attention [20]. Emotions are dynamic and have distinctive temporal profiles that are not captured by traditional models (boredom is very different from astonishment in this respect, for example). In this case, the temporal profiles would be matched against prototypes stored in memory. Some musical websites allow listeners to 'tag' specific points of the waveform (for instance, SoundCloud [3]), a valuable source of temporal annotations for popular music.

5.4 Unfolding Musical Process

The *unfolding musical process* acts as feedback loop that exploits the temporal evolution of music features at the three different time scales. The temporal correlation of each feature must be exploited and fed back to the mapping mechanism (see 'unfolding musical process') to estimate listeners' response to the repetitions and the degree of "surprise" that certain elements might have [26]. Schubert [25] studied the relationship between music features and perceived emotion using continuous response methodology and time-series analysis. Recently, MER systems started tracking temporal changes [4, 22–24, 30]. However, modeling the *unfolding musical process* describes how the time-varying emotional trajectory varies as a function of music features. Korhonen *et al.* [15] use auto-regressive models to predict current musical emotions from present and past feature values, including information about the rate of change or dynamics of the features.

5.5 Previous Exposure

Previous exposure is responsible for system customization and could use reinforcement learning to alter system response to the *unfolding musical process*. Here, the user input tunes the long-term system behavior according to external factors (independent from temporal evolution of features) such as context, mood, genre, cultural background, etc. Eerola [6] investigated the influence of musical genre on emotional expression and reported that there is a set of music features that seem to be independent of musical genre. Yang *et al.* [33] studied the role of individuality in MER by evaluating the prediction accuracy of *group-wise* and *personalized* MER systems by simply using annotations from a single user as "ground truth" to train the MER system.

6. CONCLUSIONS

Research on music emotion recognition (MER) commonly relies on the bag of frames (BOF) approach, which uses machine learning to train a system to map music features to a region of the emotion space. In this article, we discussed why the BOF approach misrepresents musical experience, underplays the role of memory in listeners' emotional response to music, and neglects the temporal nature of music. The organization of auditory memory plays a major role in the experience of listening to music. We proposed a framework that uses the organization of auditory memory to bring time to the foreground of MER. We prompted MER researchers to represent music as a time-varying vector of features and to investigate how the emotions evolve in time as the music develops, representing the listener's emotional response as an emotional trajectory. Finally, we discussed how to exploit the *unfolding music process* and *previous exposure* to incorporate the current musical context and personal factors into MER systems.

The incorporation of time might not be enough to account for the subjective nature of musical emotions. Culture, individual differences and the present state of the listener are factors in understanding aesthetic responses to music. Thus a probabilistic or fuzzy approach could also represent a significant step forward in understanding aesthetic responses to music. We prompt MER researchers to

[3] http://soundcloud.com/

adopt a paradigm change to cope with the complexity of human emotions in one of its canonical means of expression, music.

7. ACKNOWLEDGEMENTS

This work was partially supported by the Media Arts and Technologies project (MAT), NORTE-07-0124-FEDER-000061, which is financed by the North Portugal Regional Operational Programme (ON.2 O Novo Norte), under the National Strategic Reference Framework (NSRF), through the European Regional Development Fund (ERDF), and by national funds, through the Portuguese funding agency, Fundao para a Ciência e a Tecnologia (FCT). Frans Wiering is supported by the FES project COMMIT/.

8. REFERENCES

[1] J.J. Aucouturier, F. Pachet: "Improving timbre similarity: How high is the sky?" *Journ. Neg. Res. Speech Audio Sci.*, Vol. 1, No. 1, 2004.

[2] M. Caetano, A. Mouchtaris, F. Wiering: *The role of time in music emotion recognition: Modeling musical emotions from time-varying music features*, LNCS, Springer-Verlag, 2013.

[3] O. Celma, X. Serra: "FOAFing the music: Bridging the semantic gap in music recommendation," *Journ. Web Semantics* Vol. 6, No. 4, 2008.

[4] E. Coutinho, A. Cangelosi: "Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements," *Emotion* Vol. 11, No. 4, pp. 921–937, 2011. 2004.

[5] S. Cunningham, D. Bainbridge, J. Downie: "The impact of MIREX on scholarly research (2005-2010)," *Proc. ISMIR*, 2012.

[6] T. Eerola: "Are the emotions expressed in music genre-specific? An audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journ. New Mus. Res.*, Vol. 40, No. 4, pp. 349–366, 2011.

[7] T. Eerola: "Modeling listeners' emotional response to music," *Topics Cog. Sci.*, Vol. 4, No. 4, pp. 1–18, 2012.

[8] A. Friberg: "Digital audio emotions - An overview of computer analysis and synthesis of emotional expression in music," *Proc. DAFx*, 2008.

[9] A. Gabrielsson, E. Lindström. The role of structure in the musical expression of emotions. *Handbook of Music and Emotion*, Oxford University Press, 2011.

[10] X. Hu, J. Downie, C. Laurier, M. Bay, and A. Ehmann: "The 2007 MIREX audio mood classification task: Lessons learned," *Proc. ISMIR*, 2008.

[11] A. Huq, J. Bello, R. Rowe: "Automated music emotion recognition: A systematic evaluation," *Journ. New Mus. Res.*, Vol. 39, No. 3, pp. 227–244, 2010.

[12] D. Huron: *Sweet Anticipation: Music and the Psychology of Expectation*, Bradford Books, MIT Press, 2008.

[13] P. Juslin, S. Liljeström, D. Västfjäll, L. Lundqvist: How does music evoke emotions? Exploring the underlying mechanisms. In: *Handbook of Music and Emotion*, Oxford University Press, 2011.

[14] Y. Kim, E. Schmidt, R. Migneco, B. Morton, P. Richardson, J. Scott, J. Speck, D. Turnbull: "Music emotion recognition: A state of the art review," *Proc. ISMIR*, 2010.

[15] M. Korhonen, D. Clausi, M. Jernigan: "Modeling Emotional Content of Music Using System Identification," *IEEE Trans. Syst., Man, Cybern.*, Vol. 36, No. 3, pp. 588–599, 2005.

[16] C. Krumhansl: "Music: A Link Between Cognition and Emotion," *Current Direct. Psychol. Sci.*, Vol. 11, No. 2, pp. 45–50, 2002.

[17] C. Laurier, M. Sordo, J. Serrà, P. Herrera: "Music mood representations from social tags," *Proc. ISMIR*, 2009.

[18] L. Lu, D. Liu, H. Zhang: "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, Lang. Proc.*, Vol. 14, No. 1, pp. 5–18, 2006.

[19] K. MacDorman, S. Ough, H. Chang: "Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison," *Journ. New Mus. Res.*, Vol. 36, No. 4, pp. 281–299, 2007.

[20] F. Nagel, R. Kopiez, O. Grewe, E. Altenmüller: "EMuJoy: Software for continuous measurement of perceived emotions in music," *Behavior Res. Meth.*, Vol. 39, No. 2, pp. 283–290, 2007.

[21] K. Scherer: "Which emotions can be induced by music? what are the underlying mechanisms? and how can we measure them?" *Journ. New Mus. Res.*, Vol. 33, No. 3, pp. 239–251, 2004.

[22] E. Schmidt, Y. Kim: "Modeling musical emotion dynamics with conditional random fields," *Proc. ISMIR*, 2011.

[23] E. Schmidt, Y. Kim: "Prediction of time-varying musical mood distributions from audio," *Proc. ISMIR*, 2010.

[24] E. Schmidt, Y. Kim: "Prediction of time-varying musical mood distributions using kalman filtering," *Proc. ICMLA*, 2010.

[25] E. Schubert: "Modeling perceived emotion with continuous musical features," *Music Percep.: An Interdiscipl. Journ.*, Vol. 21, No. 4, pp. 561–585, 2004.

[26] E. Schubert: "Analysis of emotional dimensions in music using time series techniques," *Context: Journ. Mus. Res.*, Vol. 31, pp. 65–80, 2006.

[27] B. Snyder: *Music and Memory: An Introduction.*, MIT Press, 2001.

[28] B. Sturm: "Evaluating music emotion recognition: Lessons from music genre recognition?," *Proc. IEEE ICMEW*, 2013.

[29] G. Tzanetakis, P. Cook: "Musical Genre Classification of Audio Signals," *IEEE Trans. Speech, Audio Proc.*, Vol. 10, No. 5, pp. 293–302, 2002.

[30] Y. Vaizman, R. Granot, G. Lanckriet: "Modeling dynamic patterns for emotional content in music," *Proc. ISMIR*, 2011.

[31] G. Wiggins: "Semantic gap?? Schemantic schmap!! Methodological considerations in the scientific study of music," *Proc. Int. Symp. Mult.*, 2009.

[32] Y. Yang, H. Chen: "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech, Lang. Proc.*, Vol. 19, No. 4, pp. 762–774, 2011.

[33] Y. Yang, Y. Su, Y. Lin, H. Chen: "Music emotion recognition: the role of individuality," *Proc. HCM*, 2007.