RAPSCOM - A FRAMEWORK FOR RAPID PROTOTYPING OF SEMANTICALLY ENHANCED SCORE MUSIC

Julian Rubisch
Institute for Media Production
University of Applied Sciences
St. Pölten
jrubisch@fhstp.ac.at

Jakob Doppler
Institute for Media Production
University of Applied Sciences
St. Pölten
jdoppler@fhstp.ac.at

Hannes Raffaseder
Institute for Media Production
University of Applied Sciences
St. Pölten
hraffaseder@fhstp.ac.at

ABSTRACT

In film and video production, the selection or production of suitable music often turns out to be an expensive and time-consuming task. Directors or video producers frequently do not possess enough expert musical knowledge to express their musical ideas to a composer, which is why the usage of temp tracks is a widely accepted practice. To improve this situation, we aim at devising a generative music prototyping tool capable of supporting media producers by exposing a set of high-level parameters tailored to the vocabulary of films (such as mood descriptors, semantic parameters, film and music genre etc.). The tool is meant to semi-automate the process of producing and/or selecting temp tracks by using algorithmic composition strategies to either generate new musical material, or process exemplary material, such as audio or MIDI files. Eventually, the tool will be able to provide suitable raw material for composers to start their work. We will also publish parts of the prototype as an open source framework (the RaPScoM framework) to foster further development in this area.

1. INTRODUCTION

1.1 Context

In contemporary film or video productions, score music composition or selection is widely regarded as vital for the movie's reception and the conveying of moods, metaphors and meanings. It is, however, also sometimes treated as an orphan because of its expensiveness and time-consuming qualities. Moreover, movie directors or video producers frequently lack the musical expertise to communicate their wishes and ideas to a film composer. Therefore, in the majority of cases temp tracks are used as a fallback.

Within the community of film composers, however, temp tracks are being disapproved of, as they often confine the composer's imagination. In many cases, directors also cling to their temp tracks' musical features (themes, harmonies, rhythmic features etc.) very tightly, which makes film music production a complex and inefficient process for both sides.

1.2 Objectives

The research project GeMMA (Generative Music for Media Applications [1]) tries to ameliorate this situation by

- semi-automating the generation of temp tracks,
- enhancing the communication between director and composer (not necessarily by providing the optimal output, but with a focus on optimal description/selection of the desired output),
- providing an intuitive user interface, usable for both musical experts and laypersons and
- processing initial musical structures (audio or MIDI files) so as to facilitate rapid prototyping by example.

It is also a goal to release parts of the employed algorithms as an open source project, the RaPScoM framework.

1.3 Related Work

There have been numerous initiatives to exploit algorithmic composition in an affective or semantic way, of which we will cite a few ones:

[1] describes a state based sequencer for automatic sound track generation that consists of pluggable agents for tempo, key, chord, instrumentation and rhythm. The generation process uses mainly genetic algorithms and markov models for chord progressions but requires a thorough screenplay, character, location and event annotation and does not analyze existing material.

[2] designed an Algorithmic Music Evolution Engine (AMEE) which uses an emotional mapper to link a perceived mood to a set of musical parameters, such as consonance, pitch, mode, articulation, tempo etc. Their model, however, relies on a discrete, taxative set of moods which cannot be altered by the user.

[3] use Russell's circumplex model of affect [4] as a basis to alter a set of musical parameters. Notably, they have divided their approach into a timing (groove), a harmonic and a voicing module to combine several features.

A common weakness heretofore, at least to our knowledge, is the general neglection of the necessity to find novel ways of description regarding audiovisual source material other than plain technical video or audio descriptors. In

[1] http://gemma.fhstp.ac.at

the TRECVID evaluation [5], semantic indexing tasks have been rather poorly performing to the present day.

2. METHODOLOGY

The field of algorithmic composition and generative music is structured by the influence of various disciplines, such as computational intelligence, music psychology and philosophy, semiotics and, of course, musicology. Therefore, a tripartite approach towards the problem has to be taken.

2.1 Socio-Cultural Approach

When analyzing music for film or video, it is essential to consider

- the functions it is able to incorporate,
- the levels of impact it triggers as well as
- what semiotic structures it possesses.

[6] [7] [8] and [9] provide an overview of the structured analysis of the field of film music and sound semiotics, while [10] analyzed what classes of functions film music incorporates and utilizes to achieve a certain impact.

These classes include ([10], p.2)

the emotive class: mood induction (emotions experienced by the audience) and communication of emotion (only identified by the audience)

the informative class: communication of meaning, communication of values and establishing recognition

the descriptive class: describing setting or physical activity

the guiding class: indicative (guide the audience's attention) or masking (disguise other noises or narrative elements)

the temporal class: providing continuity (disguising of cuts from scene to scene, usage of leitmotifs etc.), defining structure and form (forming the perception of time and speed)

the rhetorical class: commenting the narrative

The thorough understanding of film music semiotics is crucial here, because music in general uses symbolic gestures to fulfill the mentioned functions. In order to maintain a manageable scope of this vast field of research, we decided to focus on two central aspects:

- representation of affects and moods, and
- analysis of semiosis by abstract musical symbols.

2.1.1 Representation of Affects and Moods

Widely used in music psychology, Russell's Circumplex Model of Affect [4] presents a solid basis for both a computational representation of moods and emotions as well as an intuitive user interface. Briefly, the model assumes that every human affect is a linear combination of two neurophysiological systems called valence (ranging from unpleasant to pleasant) and activation (ranging from passive to active). While this is indisputably an oversimplification, the model serves well in many music-psychological studies as well as music information retrieval (MIR) tasks [11].

2.1.2 Semantics

By many semioticians, music is seen as a semiotic system without semantic density [12], i.e. musical signs (e.g. melodies, motifs, rhythmic patterns etc.) have syntactic relations, as defined by music theory and harmonics, but no inherent meaning (as compared to linguistics, where words are assumed to carry a certain meaning). On the other hand, certain musical segments do carry clear denotative (e.g. hunting horn signals) or connotative (e.g. pastoral, sacral, etc. music) significances, and film music makes use of these connotative meanings quite excessively.

In fact, it seems advisable in the special case of film music to not only regard syntactic and semantic features, but to see the sounding material in the context of the plot - i.e. to consider the pragmatic aspects of a movie. After all, photography, editing, acting, sound design, music, lighting, and many more aspects of a movie are welded together to form a certain narrative. It is thus possible to charge a simple musical gesture (e.g. a simple chord or tune) with a clearly defined meaning by setting it in an appropriate, coded context. The relation of the music to the (mostly visually defined) context can be either

paraphrasing: duplicating what is seen on the screen (e.g. a romantic tune to a love scene),

polarizing: charging a neutral context with meaning, or

contrapuntal: contradicting the visual narrative, thus introducing another semantic layer (cf. [7] [8] [13]).

To establish a language system, it is necessary to introduce conventional codes, i.e. stereotypes, which render film music a communicative art and form styles and traditions. However, it is necessary to differentiate stereotypes from cliches, which are stereotypes reduced to a certain, isolated meaning. The use of cliches is affirmative, it merely uses fixed assignments of signifieds without questioning underlying socio-cultural developments, i.e. the progress of tradition (cf. [13], p. 83f).

To analyze the usage of musical signs in various contexts (i.e. their use as stereotypical gestures), we divided our analysis in

- events (dominant, temporally confined narrative elements of a scene) and
- symbols (higher-level dramaturgical motifs of a scene) (cf. [6] [8])

2.2 Aesthetic Approach

It is of course crucial to consider whether and how aesthetically interesting output can be produced by an algorithmic engine. While it is also clear that this field entails many related questions (ethical, philosophical, cognitive ones), it is impossible to approach it in a quantitative way. The aesthetic content of a piece of music to a great degree relies on the listeners' anticipations, associations from their personal history as well as cultural backgrounds.

To obtain musically interesting results, artificial intelligence (AI) and/or life (AL) methods are experimented on, including pattern recognition and supervised learning methods, as well as artificial neural networks (e.g. echo state networks) and genetic algorithms.

In order to monitor the music's aesthetic impact on listners, we are performing qualitative user and listener reviews as an accompanying measure (including interviews with composers and directors concerning style and impact of the generated music).

2.3 Technical Approach

The technical realization of the project yields yet another number of problems:

2.3.1 High-Level Architecture

As mentioned in the introduction, the tool is meant to be usable for both experts and laypersons. Therefore, it has to be clarified which set of parameters should be exposed to the users as well as how audio or MIDI input can be analyzed, including the evaluation of salient harmonic, melodic, and rhythmic features. Furthermore, models and algorithms for the generation of musical content have to be reviewed and tested.

2.3.2 User Interface

A central question that has to be addressed is in what way (non-)expert users should be enabled to interact with an intelligent music-generating engine.

2.3.3 Low-Level Building Blocks

Finally, the pivotal issue in this project is the question how the findings from the above mentioned approaches can be broken down into independent components which serve as building blocks for a generative music-making automaton.

3. RESULTS

3.1 Requirements and Constraints for Temp Tracks

Common score music production workflow is a loose triangular communication between editor, director and composer. First, in the spotting sessions editor and director use temp tracks from similar productions to produce a rough cut and expose ideas on the music theme and semantics of a scene. The composer then is required to transform these ideas into unique sounding score music. In an iterative process involving all persons the score music is merged with the simultaneous evolving rough cut to form the final product [13] [14]. The RaPScoM framework aims at improving

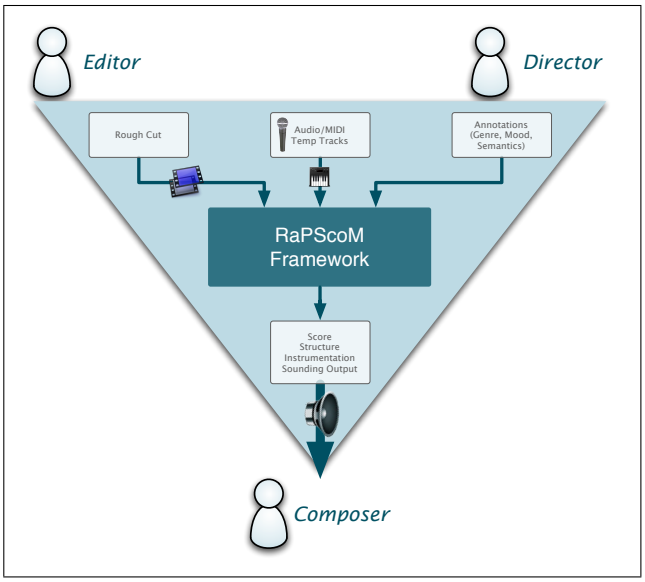

Figure 1. Triangular commucation between editor, director and composer, enhanced by the RaPScoM framework.

the film scoring workflow by generating musically meaningful raw material such as musical structures and instrumentation for temp tracks that can directly be processed by the composer (see figure 1).

3.2 Semantic Analysis

As indicated above, a thorough understanding of film (music) semiotics is pivotal to the project of providing accurate music according to the director's raw ideas.

To gain insight concerning affective and semantic de- and connotations used in the language of film music, we conducted the following investigations:

3.2.1 Representation of Affects and Moods

A listener test was conducted, using the above mentioned Circumplex Model of Affect, to discover possible correlations between library score music and musical/timbral parameters by playing a set of music clips to a group of listeners (8 male, 6 female, between 21 and 50 years old) and asking them to

- place the respective clip in the Circumplex, and
- fill out a semantic differential of these musical/timbral parameters (*brightness, harmonicity, loudness, consonance, tempo, rhythm, pulse, dynamics, quietude and hardness*).

The subsequent statistical analysis indicated some strong correlations between loudness, hardness, quietude, tempo and activation, and somewhat weaker correlations between harmonicity, rhythm, pulse and activation, as well as brightness, consonance, tempo and valence. To a great extent, these findings correspond to phenomenons described in relevant literature (e.g. [15]).

3.2.2 Symbolic Content of Movie Scenes

In order to find possible similarities regarding musical symbols and their significances, we analyzed a corpus of approx. 400 short movie clips (randomly selected) according to symbol, event and the employed musical instruments. To facilitate the analysis, symbols and events were aggregated into clusters by agglomerative hierarchical clustering [16].

In order to obtain a meaningful distance measure for these nominally scaled sets of data, we decided to construct a binary vector to represent the occurences of symbols/events per clip (where *1* means *clip is tagged with a symbol* while *0* means the opposite). Thus, the record sets could be compared to each other by use of a hamming distance and clustered accordingly [17].

It turned out that agglomerative clustering with single linkage tends to quickly merge small clusters into larger ones, leaving single clips unclustered, which is why finally complete linkage clustering was employed. Agglomerative hierarchical clustering features the advantage of being able to determine the amount of clusters after the clustering process has completed. Therefore, the following *8 symbol clusters* and *8 event clusters* were selected and labelled by hand:

Symbol	Event
Action/Violence	Movement
Fear/Tension	Drama
Freedom	Accident
Joy/Comedy	Shock
War	Violence
Tragedy	Surprise
Romance	Death
Desolation	Celebration

Table 1. Symbol and Event super categories retrieved by machine clustering

Further listening tests showed that 59 % of test persons (N=87) identified a strong relationship between symbol and instrumentation of a scored movie scene. Only within the symbol groups of *freedom* and *war*, melody is awarded a higher degree of correspondence with the intended symbolic meaning. Rhythmic features are almost never associated with semantic content of a movie scene.

Currently, a correspondence analysis of symbols and used instrumentation/solo intruments and melodic parameters (e. g. melodic contour, mode, ambitus) is conducted. Plausibly, as can be argued from music history, instrumental (or in a smaller degree melodic) stereotypes are used to convey a certain scene setting, e.g. horns for a war or hunting scene, flutes for a pastoral scene etc. Ideas for this approach were taken from the german Handbook of Film Music [13] and van Leeuwens Speech, music, sound [18]. The goal here is to provide a probability matrix for the instrumentation and melodic composition of scenes according to their semantic features.

3.3 Implementation Prototypes

In order to test the validity of the above mentioned affective and semantic models, we decided to implement different generator algorithms in a bottom-up approach first, before designing the entire framework. These include:

3.3.1 MotifFactory

This building block is planned to operate on a low level of the framework, and comprises methods to model a melody (and variations thereof) as well as a tune's consonance and rhythm (and variations). It is able to analyze and process initial MIDI or audio input and provide appropriate variations according to a predefined set of parameters. To accomplish this, the melody is broken down into a first- or second-order Markov chain and reassembled randomly.

Moreover, it is possible to pick a motif according to its melodic envelope (e.g. a falling slope, or first rise and then fall, etc.). On a higher level, the most appropriate variation will be selected and formed into a complete musical segment by an intelligent algorithm (e.g. an artificial neural network or an agent-based artificial life algorithm).

The major aim of this experimental implementation was to gain experience about how the variation of very short musical segments can already influence the perceived mood or affect of a tune. First results of this evaluation, which was conducted on short known melodies (e.g. Beethoven's *Für Elise*) sound very promising in terms of musical originality, while maintaining a clear similarity to the original and providing affectively biased variations.

3.3.2 SemanticChordProgressionGenerator

The purpose of this demo implementation was to investigate how larger-scale musical segments, spanning over a wider range of e.g. 8 bars, can be used to create, sustain and release musical tension. For the realization of this task, statistical chord progression data from [15], as well as algorithms from [19] and [20] were used. Currently, we are reviewing and implementing composition rule frameworks (e.g. prohibit the use of parallel fifths, encourage the use of a certain register, use close or open harmony, divide in antecedent/consequent etc.) to be included in this model. We are strongly convinced that in the use of small-scale variation of certain parameters (such as arpeggio style, dynamics, direction, rhythmic complexity and others) in a larger-scale context of harmony lies one of the pivotal foundations of semantically enhanced music generation.

In a first attempt to include symbolic information (e.g. *war* or *romance*), we decided on including an orchestral sample library here, and have the algorithm lock a certain instrumental arrangment before generating the chord progression. Another task we are focusing on is the investigation and evaluation of harmonies frequently used in film music.

4. THE RAPSCOM FRAMEWORK

4.1 Requirements

Our approach to rapid prototyping for score music is based on the semantic annotation of the rough cut. A movie de-

scriptor contains a set of global properties (musical film style/genre) and timeline parameters (movie semantics, emotions) which are tailored to the vocabulary of films. In the above mentioned study we found that music inherent movie semantics are best described as a set of 6 scene symbols (e.g. action, fear, romance) which are sparsely intermitted by 8 events (e.g violence, surprise, shock). For the representation of emotions the mentioned Circumplex Model of Affect is used. Various input modalitites such as tagging tools and 2D panels for example on touch tablets are currently under review for generating and editing annotations.

4.2 Environment

The framework's generation algorithms are geared to producing MIDI raw material; it is thus necessary for the user to install a MIDI-based host-application (such as any contemporary digital audio workstation, or sampler). To ensure a certain channel-instrument mapping (e.g. violins on MIDI channel #1, violas on channel #2 etc.), templates for a certain set of audio workstations and samplers will be provided. In order to be compatible with the General MIDI standard, the appropriate program change messages will be sent on each channel, too. It is however also possible to alter the default channel mappings in an external configuration file.

4.3 Structure

4.3.1 Models

Following [19], we decided to devise a hierarchical framework of musical structures to manage the analysis and generation of music on several symbolic levels. Thus, the backbone of the framework consists of models of *Note, Chord, Motif, Theme, and Piece*, where the latter can always contain multiple instances of the former (a *Piece* can contain many *Themes* etc., see also figure 2).

Every class in the hierarchy (except for *Note*) extends a common base class holding general properties of *NoteContainers*, e.g. a key, time signature, and handles for traversing the hierarchy (getParent, getChildren). They furthermore implement two interfaces, *IAnalyzeable* (making it compatible to several analyzer methods, see below), and *IVariable* (making it possible to create intelligent variations of a certain element).

For rhythmic purposes, every note has a *NotePosition* in upper (beat-), mid (tactus-) and lower (sub-tactus-)level format, as Temperley proposed in [20]. This note position can be moved by note values, such as (triolic, dotted) quarter notes etc.

4.3.2 Processors

To accomplish the task of generating, analyzing and varying musical content, it is advisable to conceive a set of processor classes, divided in analyzers, such as

- KeyAnalyzer
- RhythmAnalyzer

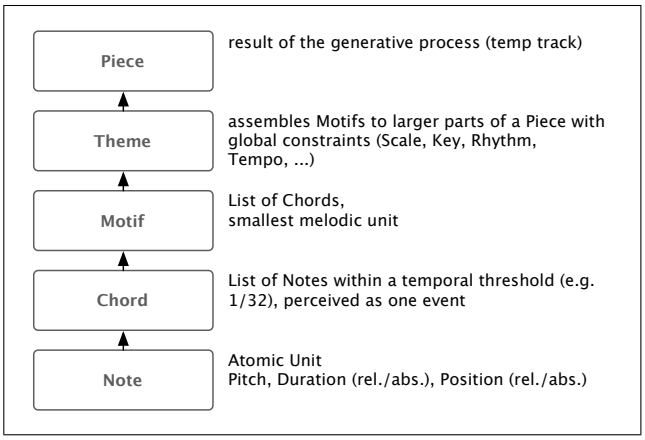

Figure 2. Model hierarchy

- ConsonanceAnalyzer

and generators, such as

- ChordProgressionGenerator
- ArpeggioGenerator
- MonophonicMotifFactory

These modules rely on *KnowledgeBases* to provide essential rulesets for the tasks of music analysis and generation (see also figure 3).

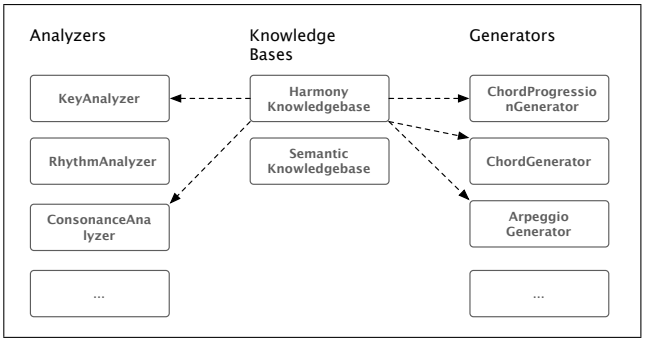

Figure 3. Processor modules

4.3.3 IO Modules

Every instance of the above mentioned hierarchy can be sent to a *Renderer* (e.g. a MIDI file renderer) to produce the respective output. A central part of the IO framework is the *ChannelsDescriptor*, where relevant information on the instrument-channel-mappings (such as name of the mapped instrument, pitch range, etc.) are stored.

4.4 Session

A RaPScoM project with inputs, outputs, requirements and defined workflow is called a *session*. A session is taken to mean a complete reference implementation of the framework's independent modules, allowing to semi-automatically generate musical output, as outlined in the introduction. It comprises (also see figure 4)

Semantic Movie Descriptor: film material annotated with valence/arousal, symbol, event, style and genre

Instrumentation: we pursue an *instrumentation first* policy here - before generating the actual musical elements, the orchestral arrangement is locked, based on the symbolic annotation of the scene. However, we also implement a feedback loop to be able to try out different instrumentations of the same score after it has been generated.

Sequence: the music sequence (a hierarchy starting with a *Piece*, down to the single *Notes*, is generated in accordance to instrument mappings, affective, stylistic and semantic annotations, as well as composition rules and harmonic guidelines. Every decision is logged, in order to trace back every step and from there start another generation of variations.

Channel Descriptor: as indicated above, information on the instrument-channel-mapping, voicing etc. is stored here

Host Application: determines the way the MIDI data is transformed to sounding material on the user's computer. It will be a future task to develop a description format (e.g. XML-based) of sound generators, in order to ensure that the generated audio material fits the listener's expectations

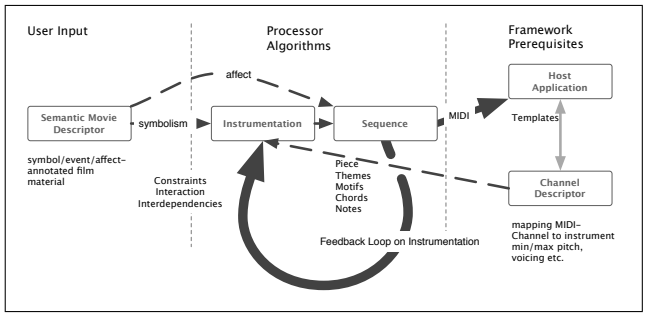

Figure 4. Flowchart of a RaPScoM session.

5. CONCLUSION

In this paper, we have introduced first steps toward a software framework to achieve semi-automated production of film music temp tracks in order to enhance the video and film production process. The main difficulty in such a project derives from the various different methodologies that have to be included, as well as the requirements put to the prototype by professional users, such as directors or video producers.

We proposed different ways of approaching these independent problems and identified salient factors for describing and evaluating score music according to semiotics and aesthetics. We are currently in the process of reviewing and experimenting on generative music-making algorithms in order to produce meaningful content which is capable of replacing or augmenting the wide practice of using temp tracks in film and video production.

We have further outlined the structure of the RaPScoM (Rapid Prototyping of Semantically Enhanced Score Music) framework, its requirements and constraints as well as in what type of environment it is meant to be used. We hope that we will be able to complete the core of the framework along with a prototypical reference implementation by mid 2012, which will fuel further discussions about music production in the creative industry.

Acknowledgments

GeMMA is funded by the Austrian Research Promotion Agency (FFG [2]) under the COIN (Cooperation & Innovation) programme on behalf of the Austrian Federal Ministry for Transport, Innovation and Technology [3], as well as the Austrian Federal Ministry of Economy, Family and Youth [4].

6. REFERENCES

[1] M. O. Jewell, "Motivated music: Automatic soundtrack generation for film," Thesis, 2007. [Online]. Available: http://eprints.ecs.soton.ac.uk/13924/

[2] M. Hoeberechts and J. Shantz, "Real-Time emotional adaptation in automated composition," in *Proceedings of Audio Mostly 2009 - a conference on interaction with sound*, Glasgow, UK, 2009.

[3] I. Wallis, T. Ingalls, and E. Campana, "Computer-Generating emotional music: The design of an affective music algorithm," in *Proceedings of the 11th Int. Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[4] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[5] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. New York, NY, USA: ACM Press, 2006, pp. 321–330.

[6] B. Flückiger, *Sound Design : die virtuelle Klangwelt des Films*. Marburg: Schüren, 2001.

[7] M. Chion, *Audio-vision : sound on screen*. New York: Columbia University Press, 1994.

[8] H. Raffaseder, *Audiodesign*, 2nd ed. München [u.a.]: Fachbuchverl. Leipzig im Carl-Hanser-Verl., 2010.

[9] T. V. Leeuwen, *Speech, music, sound*. Houndmills Basingstoke Hampshire ;New York: Macmillan Press St. Martin's Press, 1999.

[10] J. Wingstedt, "Narrative functions of film music in a relational perspective," in *Proceedings of ISME - Sound Worlds to Discover*, Santa Cruz, Tenerife, Spain.

[2] http://www.ffg.at
[3] http://www.bmvit.gv.at
[4] http://www.bmwfj.gv.at

[11] L. Lu, D. Liu, and H. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 5–18, 2006.

[12] U. Eco, *Einführung in die Semiotik*, 9th ed. UTB, Stuttgart, Feb. 2002.

[13] E. Schneider, *Handbuch Filmmusik*, [Verschiedene aufl., nachdruck] ed. Konstanz: UVK Verlagsgesellschaft, 2006.

[14] R. Davis, *Complete guide to film scoring : the art and business of writing music for movies and TV*. Boston MA Milwaukee Wis.: Berklee Press Distributed by Hal Leonard, 1999.

[15] D. Huron, *Sweet anticipation : music and the psychology of expectation*. Cambridge Mass. London: MIT, 2008.

[16] R. Duda, P. Hart, and D. Stork, *Pattern classification*, 2nd ed. New York: Wiley, 2001.

[17] X. Hu, M. Bay, and J. Stephen, "Creating a simplified music mood classification Ground-Truth set," in *Proceedings of the 8th International Conference on Music Information Retrieval*, Wien, Sep. 2007, pp. 309–310.

[18] T. V. Leeuwen, *Speech, music, sound*. Houndmills Basingstoke Hampshire ;New York: Macmillan Press St. Martin's Press, 1999.

[19] R. Rowe, *Machine musicianship*. Cambridge Mass. London: MIT, 2004.

[20] D. Temperley, *Music and probability*. Cambridge Mass. London: MIT Press, 2010.