Emotion in Music: representation and computational modeling

Emotie in muziek: representatie en computationele modellering

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen
op maandag 19 september 2016
des ochtends te 10.30 uur

door

Anna Aljanaki

geboren op 8 augustus 1987
te Feodosia, Oekraïne

Promotor: Prof.dr. R.C. Veltkamp
Copromotor: Dr. F. Wiering

Contents

Acknowledgements

Four years ago I arrived in Utrecht on the Queen's day to start my PhD. Since then, I've met many people who shared their world view, beliefs and experiences with me and influenced me and this thesis in many ways.

First of all, I would like to thank my supervisors Frans Wiering and Remco Veltkamp for their endless support, confidence and tranquillity which was very important to me during this project. I also thank my colleagues Anja, Bas, Dimitrios, Marcelo, Vincent and Jan for being always there to discuss any matters that I needed to discuss with someone. Also, I am very grateful to my first office roommate Anna van der Zalm, who helped me to feel less lonely in the first weeks in the Netherlands and to start appreciating this country and its people. She is one of the most kind, open and cheerful people I know.

I thank my master thesis supervisor Konstantin Tretjakov, who introduced MIR field to me and got me profoundly interested in it, which is why I ended up in Utrecht.

Thanks go also to my friend and paranymph Yasi for co-working sessions with tea sessions, to Valeri for making the design of the thesis and for his support in general, to Ravi for friendly conversations in BBG, to Genia for borrowing me her macbook so I could write these acknowledgements and Denis for his WiFi, to Tom for MIR-related conversations, and also to all the people who made my stay in Utrecht memorable and happy: Mila, Misha, Claudia, Masha, Yulia. Also, Improvables, with you all the Tuesdays were so much more fun and I will miss them. USKO choir rehearsals were the never-ending soundtrack of my days in Utrecht and I am still singing Bach in my head.

I am grateful to the COMMIT/ research community for funding this research. I also thank my reading committee for carefully reading the whole thesis and providing their valuable remarks, criticism, and discussion.

Abstract

Music emotion recognition (MER) is a subfield of Music Information Retrieval (MIR) that deals with music classification and develops music similarity measures, using signal processing and machine learning techniques. In its methodology and applications MER is similar to another flagship subfield of MIR — music genre recognition. However, unlike genre ontology (which is also ambiguous), emotion ontology for music is even less well established. Musical emotion can be conceptualized through various emotional models: categorical, dimensional, or domain-specific. Emotion can be represented with a single label, multiple labels, or probability distributions. Also, the time scale to which an emotion label can be applied ranges from half a second to a complete musical piece.

Describing musical audio with emotional labels of any kind is an inherently subjective task and it is not easy for the listener, because the task of recognizing musical emotion is usually learned from exposure to music in an implicit way. MER field relies on ground truth data from the human labelers. The quality of the ground truth labels is crucial to the performance of the algorithms that are trained on these data. Lack of agreement between the annotators leads to conflicting cues and poor discriminative ability of the algorithms. Therefore, conceptualizing musical emotion in a way that is most natural and understandable for the listener is crucial both to create better quality ground truth and to build intuitive music retrieval systems.

In this thesis we mainly deal with the problem of representation of musical emotion and how it affects the computational modeling of musical emotion. The thesis consists of three parts.

Part I. In the first part, we focus on the problem of computational modeling of induced musical emotion. To solve this problem, substantial amount of labeled ground truth data is necessary, which is absent. We create a game with a purpose (Emotify) to collect emotional labels through an app released on Facebook. We use Geneva Emotional Music Scales, a new and promising domain-specific model that has not yet been put to extensive testing in real-life labeling scenario. We analyze the data from the game and find that the game is able to produce good enough agreement (and high quality ground truth) with such fine-grained labels, as nostalgia or tenderness. However, some modifications to GEMS model are suggested. Also, we find that two factors influence induced emotion most: musical taste of the listener and current mood. Next, we use the data from the game to create a computational model. We show that the

performance of the model can be improved substantially through developing better features and this step is more crucial than finding a more sophisticated learning algorithm. We suggest new features that describe the harmonic content of music. These features slightly improve the performance. However, a much bigger improvement is expected, when it will become possible to predict high-level musical concepts such as rhythmic complexity, articulation or tonalness. This requires progress in all the areas of MIR: onset detection, source separation, beat tracking, and a deep understanding of human music perception, which can be hopefully achieved in the future.

Part II. In this part (in collaboration with M. Soleymani and Y.H. Yang) we create a benchmark for MER algorithms. The benchmark is mainly focused on Music Emotion Variation Detection (MEVD) algorithms (tracking per-second change in musical emotion). We describe our evaluation metrics and list the steps taken to improve the quality of the ground truth. Then we conduct a systematic evaluation of the algorithms and the feature sets presented at the benchmark. The results suggest that the best approach is to develop separate feature sets for Valence and Arousal dimensions, and that incorporating local context either through algorithms that are capable of extracting data from the time-series (LSTM-RNN), or through smoothing, is crucial.

Part III. In this part we build on the experience obtained in benchmark organization and suggest that a better approach to MEVD is to view music as a succession of emotionally stable segments and transitional unstable segments. We proceed to list the reasons why the established MEVD approach is flawed and can not create good quality ground truth. Then we compare different approaches to emotional segmentation problem and propose an approach based on a CNN combined with MER-informed filtering.

Three public datasets, corresponding to each part of the thesis, are released.

CHAPTER 1

Introduction

A short fiction story called ‘The Fog Horn’ by Ray Bradbury is set far in the sea on a lighthouse, which emits red and white light and a sound, called the ‘Voice’, to warn the passing ships of the proximity of land in foggy weather. On one dark November night something happens. An ancient, giant, dinosaur-like sea creature, the last of its kind, arrives at the lighthouse and makes a deep cry very similar to the ‘Voice’, trying to communicate. The whole gist of the short story is in how the fog horn sounds. This sound is specified in the story by describing what the creator of the ‘Voice’ must have been thinking:

> I’ll make a voice that is like an empty bed beside you all night long, and like an empty house when you open the door, and like the trees in autumn with no leaves. A sound like the birds flying south, crying, and a sound like November wind and the sea on the hard, cold shore. I’ll make a sound that’s so alone that no one can miss it, that whoever hears it will weep in their souls, and to all who hear it in the distant towns. I’ll make me a sound and an apparatus and they’ll call it a Fog Horn and whoever hears it will know the sadness of eternity and the briefness of life.

This description of a sound is very peculiar in being metaphoric throughout. Not a single property of the sound is actually mentioned: whether it was brief or sustained, shrill or deep, noisy or melodic. Yet, probably, you could imagine this sound much more vividly than if you were provided with its amplitude, range of frequencies, and pattern of their fluctuations. We can understand how it feels to enter a cold empty house. We have heard November wind and the cries of the birds flying south. But have your experiences been similar to mine? Were the birds that flew over your house in autumn the same birds, that flew over mine? And did you, in the end, imagine the ‘Voice’ in the same way that I imagined it?

People resort to emotional terms when talking about sound or music because it is convenient and natural. Cross-cultural studies of emotional expression in music show that many of the expressive cues in music are universal. Balkwill and Thompson (1999) showed that Western listeners, unfamiliar with Hindustani ragas, could recognize joy, sadness or anger far better than chance relying on such cues as tempo, melodic complexity and timbre. Similarly, cross-cultural understanding was demonstrated by Balkwill, Thompson, and Matsunaga (2004) for Japanese music and by Fritz et al. (2009) for Western music. Music shares some of its expressive cues with speech prosody (Ilie & Thompson, 2006), which possesses universal traits across languages (Elfenbein & Ambady, 2002). Anger is expressed in most cultures by raising the loudness and tempo of speech, sadness is expressed by lowering them, and happiness is characterized by high pitch and fast tempo. It is not a coincidence, that the same traits are shared by angry, sad and happy music. But music is an abstract art that can express nuances of meaning way beyond the simplicity of the basic emotions. The relationship between acoustic cues and emotional expression is less universal and more culture-specific for refined emotions (such as tenderness, humour, solemnity or triumph). Patrick Juslin proposed to explain emotional expression in music through a version of a Brunswick's lens model: as a process of communicating emotion through a set of redundant and probabilistic cues and interactions between them (Juslin, 2000; Juslin & Lindström, 2010). The cues are encoded by composer and performer, and decoded by the listener. Figure 1.1 illustrates this process.

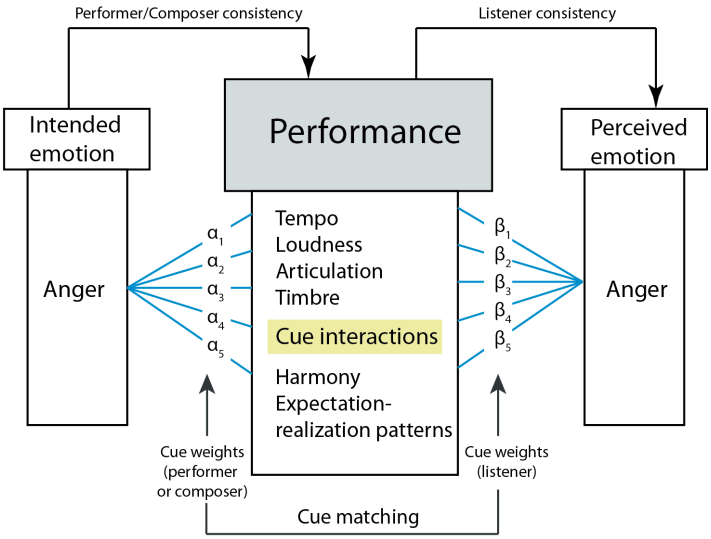


Figure 1.1: Lens model of emotional expression in music (adapted from (Juslin, 2000) and (Juslin & Lindström, 2010)).

Within a musical culture, correlational relationships are established between cues and emotions. The mapping is not deterministic, but probabilistic, and therefore several cues need to be combined for effective communication. The accuracy of the communication (r_a) is confined by the degree of matching between listeners' and performers' cues (G) and the consistency with which they are able to use them (R_l for

the listener and R_p for the performer), according to the lens model equation (Juslin, 2000):

$$r_a = GR_pR_l + C\sqrt{1-R_p^2}\sqrt{1-R_l^2}, \tag{1.1}$$

where r_a is a point-biserial correlation between the performer's intentions and listener's judgements (i.e., efficiency of the communication), R_p and R_l are multiple correlations between emotion (intended or perceived) and cues in the performer and listener models (similarly to α and β weights on Figure 1.1), and C, or unmodelled matching, represents the correlation between the residuals of the performer's model and the residuals of the listener's model. The lens model describes the sources of distortion in emotional communication for perceived (not felt) musical emotion. For felt musical emotion, there are more listener-related and situation-related factors, such as associations with a piece of music, listener's mood, evoked imagery (we will study how some of these additional factors influence induced emotion in Chapter 2). For the purposes of this thesis we are mostly interested in the right half of the lens model — the process of decoding an emotion by the listener from a set of probabilistic acoustic cues.

The cues are learned by the listener through exposure to music. Some of the cues are identical to the ones in emotional speech, as we mentioned above. Others, such as mode, rhythm and harmony related cues — are music-specific and most often learned through enculturation in the music culture of the listener.

Generally, even musicians do not receive or receive little formal training on the relationship between the acoustic cues and emotional expression. In contrast, a lot of time is spent on teaching performance technique, and other formal knowledge: labelling harmonic sequences, chords, sections and recognizing musical instruments. Though ability to play expressively is regarded as one of the most important skills for a performer, perhaps *the* most important skill, the training is often vague and mostly relies on metaphors, aural modeling, and felt emotion to teach expressive skills (Lindström, Juslin, Bresin, & Williamon, 2003). For an overview of studies analysing the content of music lessons we refer to (Zentner & Eerola, 2011a).

The problem of music emotion recognition, therefore, is very well adapted for supervised machine learning, with the task to learn a relationship between a set of acoustic cues and emotional expression. In that it is similar to other problems in Music Information Retrieval (MIR) that rely on human annotations, such as genre recognition and audio similarity tasks.

1.1 Music Information Retrieval

An early definition of MIR was given by Futrelle and Downie (2002):

> Music Information Retrieval (MIR) is an ... interdisciplinary research area encompassing computer science and information retrieval, musicology and music theory, audio engineering and digital signal processing, cognitive science, library science, publishing, and law. Its agenda, roughly, is to develop ways of managing collections of musical material for preservation, access, research, and other uses.

The MIR community is driven by the demand to tackle the massive amounts of musical data with automatic categorization and retrieval methods. MIR research is directed at enabling the creation of the tools for automatic audio and score analysis. The field is concerned with a range of problems: audio-to-score transcription, chord, key and onset detection, source separation, as well as problems that lack well-defined answers and can only be solved through supervised machine learning, such as music similarity, classification by genre or emotion. The set of problems is well represented on the annual Music Information Retrieval Evaluation eXchange[1] (MIREX) — a benchmark for MIR algorithms.

1.1.1 Music Emotion Recognition

Music industry strives to improve automatic music categorization methods for huge online music collections, and emotion is one of the key targets. Enabling keyword-based search is crucial both for personal and commercial use, especially for production music databases, when providing a more detailed query (by artist or composer) is impossible. In (Inskip, Macfarlane, & Rafferty, 2012), an analysis of written music queries from creative professionals showed that 80% of the queries for production music contain emotional terms, making them one of the most salient and important components of exploratory music search.

Music emotion recognition (MER) is one of the cornerstone audio classification tasks in MIR, which thrives on all of the MIR knowledge and expertise. For such an all-encompassing and ubiquitous in music concept as emotion, almost any MIR retrieval or recognition task, feature extraction method, sound processing technology is relevant in some way.

The MER field is only about a decade young, with the first study dating to the year 2003 (Li & Ogihara, 2003). There are still many unsolved problems in MER:

1. The subjectivity of the task leads to performance ceiling, which can hang quite low depending on the data.

2. There is a so-called "semantic gap" between the cognitive perceptual concepts and the data that we are currently able to extract from the audio signal.

3. The process of human emotion recognition is not well understood yet.

4. MER completely relies on human annotated ground truth, but due to high subjectivity and lack of understanding of human music emotion cognition assembling ground truth is a very challenging task.

We will review the different aspects of the MER field later in this thesis, for now we just refer to (Y.-H. Yang & Chen, 2012) for a survey.

[1] music-ir.org/mirex

1.2 Contribution

1.2.1 Problem statement

Music emotion recognition research depends on the ground truth data annotated by humans. As described in the previous section, there is a lot of ambiguity and subjectivity in such data. When the consistency of the annotations is low, and there is no information in the data to account for inconsistency, no learning algorithm, no matter how powerful, can achieve improvement in prediction accuracy. This problem, presumably, is now faced by the MIREX mood classification competition, where the performance has not increased for the last five years (Figure 1.2).

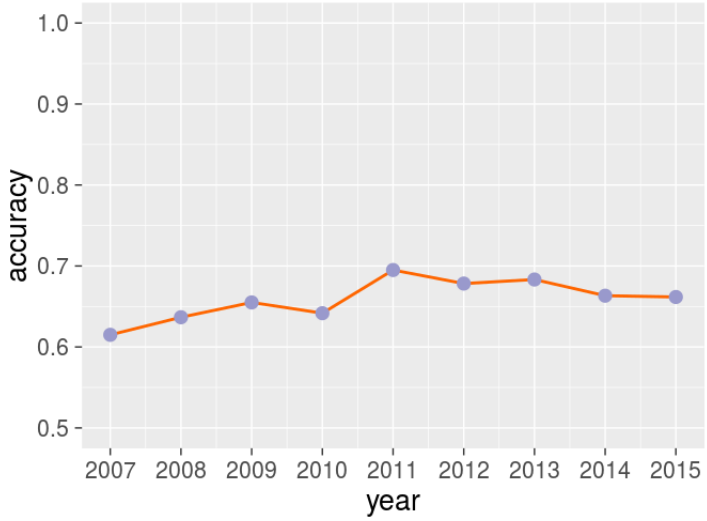


Figure 1.2: Classification accuracy in Music Mood Classification on MIREX competition, since year 2007 when the task was first introduced.

In the MIREX mood competition, the algorithms have to classify 30 second audio excerpts into five mood clusters. Some examples of the moods represented in these clusters are *passionate*, *cheerful*, *bittersweet*, *humorous* and *aggressive*. For most of the clips, only two out of three judges could agree on the assignment of the cluster (Hu, Downie, Laurier, Bay, & Ehmann, 2008), which indicates around 66% agreement. This is approximately the performance ceiling that was reached by the algorithms.

Part of the solution to the problem of subjectivity and inconsistency in musical emotion ratings lies in the realm of representation. When abstract musical meaning is translated into words, the emotional ontology is of utter importance in order not to hinder, but to help this process in a natural way. There are many ways to represent musical emotion — as static or time-varying, as a single label or multiple labels, using dimensional or categorical models. In this thesis we will show that some of these representation choices are more natural and can lead to better recognition results, and to more convenient music retrieval for the end users.

There are other sources of inconsistency as well, such as individual variation in listener perception of cue–emotion relationship (R_l in the lens model equation), listener inconsistency in cue application, and situational factors. In this thesis we will also deal with some of these variability sources.

1.2.2 List of contributions

1. Datasets of music annotated with emotion, publicly available:
   - The **Emotify** dataset: musical excerpts labelled on GEMS scale collected using a game with a purpose, and demographic information on the labellers (this is the first public dataset available on induced emotion).
   - **MediaEval Database for Emotional Analysis in Music (DEAM)** — the biggest available dataset of per-second ratings of musical emotion on the Valence and Arousal scale (1744 music excerpts and 58 complete pieces).
   - An **emotional segmentation** dataset, which offers a novel way of studying emotion variation in music through annotations of emotionally stable and unstable segments, boundaries between them, and Valence and Arousal values of the stable segments.
2. We put a promising domain-specific emotional model GEMS through a robustness test in an online game, and suggest modifications to this model. Namely, new categories *boredom*, *humour* and *impetus* are suggested, and some of the existing categories should be renamed or removed.
3. We find which perceptual features are related to induced emotion and show that the bottleneck in the performance of MER systems is extraction of these sort of cognitively motivated features. We also suggest new features to describe harmonic content of audio.
4. In a joint effort with Mohammad Soleymani and Yi-Hsuan Yang we create an evaluation framework for dynamic (per second) music emotion recognition algorithms. Using this framework, we systematically evaluate music emotion variation detection methods and feature sets.
5. Based on experience from the benchmark, we develop a more cognitively meaningful way to conceptualize continuity in musical emotion through a series of emotionally stable segments and transitions (unstable sections) between them.
6. We propose the first supervised method for emotional boundary detection based on CNN and filtering boundaries by strength.

1.3 Organization of the thesis

This thesis consists of three parts. The first part deals with induced emotion, data collection and computational modeling. The second part describes a benchmark for music emotion variation detection algorithms and its outcomes. The third part deals with a novel problem of emotion-based segmentation.

Chapter 1

In an introductory chapter we explain the focus of this thesis: representation of musical emotion. We motivate this focus by a need to improve the consistency of emotional annotations, which will result in better models.

1.3.1 Part I. Induced musical emotion

This part consists of two chapters. The first chapter describes a Game with a Purpose and a dataset collected using this game. Second chapter describes creating a model to explain these data.

Chapter 2

We introduce a game with a purpose, called Emotify, and a corpus of annotations collected using that game. For the data annotation, we choose a domain-specific emotional model GEMS, which was specifically created to describe induced musical emotion (Zentner, Grandjean, & Scherer, 2008). We test consistency and comprehensibility of the model's categories, and analyze the extra-musical parameters (mood, age, gender, musical preferences) through a series of mixed models and report which parameters influence induced emotion. These findings can be used to improve induced emotion prediction in group-based predictions.

Chapter 3

We show that a 9-item GEMS scale can be sufficiently consistent to be used as an underlying model for computational modelling, except for some of the categories. We suggest new features that describe harmonic content of audio, and show that the state-of-the-art low level spectral features seriously underperform. However, the result achieved with cognitive perceptual ratings shows that there is a lot of room for improvement for MER algorithms.

1.3.2 Part II. Benchmarking MEVD algorithms

Chapter 4

We describe the design of the benchmark for Music Emotion Variation Detection (MEVD) algorithms. The winning algorithms and feature sets over the years are analyzed, and the design, evaluation metrics and data that we used are described. We also release the largest available dataset of continuous annotations of music with emotion, and suggest some transformation and data cleaning procedures which improve the quality of these data.

1.3.3 Part III. Emotion-based segmentation

Chapter 5

Based on the experience of benchmark organization, we challenge the basic assumptions of dynamic MER approach, and argue that the problems that we encountered

originate from the unnaturally low time resolutions that dynamic MER is usually dealing with. We suggest to move to longer and more meaningful segments of music and conduct a proof-of-concept experiment. The three annotators achieve a very good agreement on emotional boundary annotations. We also show that though emotional boundaries often coincide with the structural ones, there is no full overlap.

Chapter 6

We propose a method for emotional segmentation based on a combination of CNN trained on mel-spectrograms for boundary candidate detection, and MER-informed emotional boundary strength estimation for filtering.

1.4 Related publications

Chapter 2

- Aljanaki, A., Bountouridis, D., Burgoyne, J. A., van Balen, J., Wiering, F., Honing, H., & Veltkamp, R. C. (2014). Designing games with a purpose for data collection in music research. Emotify and Hooked: Two case studies. In Lecture notes in computer science, pages 29–44, 2014.

- Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2016). Studying emotion induced by music through a crowdsourcing game. Information Processing and Management, 52(1), 115–128.

Chapter 3

- Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2014). Computational modeling of induced emotion using GEMS. In Proceedings of the 15th International Society for Music Information Retrieval Conference (pp. 373–378).

Chapter 4

- Aljanaki, A., Soleymani, M., & Yang, Y.-H. (2015). Emotion in Music Task at MediaEval 2015. In Working Notes Proceedings of the MediaEval 2015 Workshop.

- Aljanaki, A., Soleymani, M., & Yang, Y.-H. (2014). Emotion in Music Task at MediaEval 2014. In Working Notes Proceedings of the MediaEval 2014 Workshop.

- Soleymani, M., Aljanaki, A., & al., Y.-H. Y. et. (2014). Emotional Analysis of Music: a comparison of methods. In Proceedings of the ACM International Conference on Multimedia.

- Soleymani, M., Caro, M. N., Schmidt, E. M., & Yang, Y.-H. (2013). The MediaEval 2013 Brave New Task: Emotion in Music. In Working Notes Proceedings of the MediaEval 2013 Workshop.

Chapter 5

- Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2015). Emotion-based segmentation of musical audio. In Proceedings of the 16th International Society for Music Information Retrieval Conference (pp. 770–777).

Part I

Induced musical emotion

CHAPTER 2

Emotify: game with a purpose and a data corpus

Within music-related emotions, an important distinction can be made between emotions that are expressed by music (while the listener is not necessarily feeling them as well), and emotions felt by the listener in response to music. The former are referred to as *perceived* emotions, and the latter as *induced* emotions. In this chapter we describe and analyze the dataset of music annotated with induced (felt) musical emotion. We designed an online game with a purpose, Emotify (`emotify.org`), to collect induced emotion annotations from a varied sample of participants, using a nine item domain-specific emotional model GEMS (Geneva Emotional Music Scales) (Zentner et al., 2008). This corpus of annotations[1] will be used in the next chapter to create a content-based music induced emotion recognition system. In this chapter we will describe the game and analyse the corpus, i.e., agreement of participants on different categories of GEMS and influence of extra-musical factors on induced emotion (gender, mood, music preferences).

2.1 Introduction

Some of the practical applications of music emotion recognition (MER) demand predicting induced emotion — e.g., emotion-based music recommendation, generation of situational context-based playlists (for example, a workout playlist or meditation playlist), music retrieval for music therapy. Machine learning approaches to computational modeling of musical emotion require large amounts of annotated data, however large enough datasets are only available for *perceived* emotion (and not for *induced* emotion). In an absence of data, it is not even possible to determine whether non-personalized approach to induced emotion is viable (this kind of approach could be

[1] `www.projects.science.uu.nl/memotion/emotifydata`

useful in a public space, such as fitness club, restaurant, supermarket, or on a non-subscription website).

Predicting induced emotion is arguably more difficult than predicting perceived emotion, because individual differences play an even larger role. Patrick Juslin proposed BRECVEMA framework that describes eight mechanisms, through which music might arouse emotion (Juslin, 2013). Some of these mechanisms are related to the properties of the music (brain stem reflex, rhythmic entrainment, musical expectancy), but most are specific to the listener or situation (evaluative conditioning, visual imagery, episodic memory, aesthetic judgement). The predictions can be improved, if we understand the factors that contribute to individual fluctuations. For instance, group-wise MER approach suggested by Y.-H. Yang, Lin, Su, and Chen (2008) would benefit from such knowledge.

Another source of variance, which is equally important both for induced and perceived emotion, emerges when musical emotions are translated into verbal descriptions with fuzzy meaning. There is still no consensus between researchers on the most efficient model to describe musical emotions, despite numerous attempts to find one (see Section 2.2.3 for a review of different models). The choice of model is essential to the performance of MER algorithms. A model that fails to describe the phenomenon precisely will result in poor agreement between listeners and conflicting musical cues associated with different emotions, which will impede the accuracy of prediction. A model that oversimplifies the problem might result in a better agreement, but would be less useful in a retrieval system. Thus the difficulty is, on the one hand, in creating a model that reflects the complexity and subtlety of the emotions that music can demonstrate, while on the other hand providing a linguistically unambiguous framework that is convenient to use to refer to such a complex non-verbal concept as musical emotion. Currently, a wide variety of emotion ontologies can be found not only in research, but in music industry as well, from the valence–arousal model used by Musicovery[2], or ten categories ranging from *happy* and *fun* to *dramatic* and *stressful* by Aupeo[3], to no ontology at all, as in a bottom-up approach of Stereomood[4] that hosts emotional playlists non-systematically created by users. In 2008, a model designed to be domain-specific for music was developed by Zentner et al. (2008): Geneva Emotional Music Scales (GEMS). It was targeted specifically at induced musical emotion, and, as compared to other categorical models, GEMS describes refined positive responses to music in much more detail. Since 2008, GEMS has been used in several studies with promising results (J. K. Vuoskoski & Eerola, 2010; Torres-Eliard, Labbe, & Grandjean, 2011; Baltes, Avram, Miclea, & Miu, 2011; Jaimovich, 2013). The music corpora used in these studies were not large enough to serve as ground truth for a content-based MER system, and no public data have been released.

GEMS is a relatively new and somewhat less frequently employed emotional model, which still requires additional real-life verification. Some of the choices of the original study by Zentner et al. (2008) left some questions to be answered: mostly classical music was used, and the study was conducted in French (the terms were later translated to English). The data from that study is also not available.

[2] musicovery.com
[3] aupeo.com
[4] stereomood.com

Obtaining ground truth remains a challenging task for MER research, where both music copyright and costs of annotation (with music annotation being a particularly time-consuming task) pose problems. Outside the laboratory, there are two possible ways of assembling a dataset labeled with emotion annotations: through social tag mining (relying on websites such as `last.fm` or `allmusic.com`) and in a more systematic way through user surveys or data collection games. Social tag mining makes it possible to collect a huge dataset, but lacks the homogeneity and control that a preselected emotional model and a controlled experimental setting provides. In most cases it is unfeasible in tag mining to measure the level of agreement between multiple users on certain tags (or it would be necessary to apply an additional cross-verification procedure as it was done in case of the MIREX audio mood recognition task (Hu et al., 2008)). A controlled user experiment would be an ideal way of data collection. In this case, in addition to self-report, researchers can collect physiological measurements and exclude external factors that might influence the outcome. However, firstly, such a setup lacks ecological validity, and secondly, tasks involving music are very time-consuming. In the end, researchers seem to be left with a difficult choice between a small-scale or a very expensive survey. To avoid these costs, some researchers are trying to collect information through online multiplayer games with the underlying purpose of collecting scientific data. Games with a purpose are also not without their flaws as a way of data collection: it is very difficult to design a game that is attractive, fun to play, and difficult to cheat (where cheating can result in garbage data).

For our purpose (collecting a large enough dataset suitable for training a MER method on induced emotion recognition and studying the influence of extra-musical factors on emotion induction), it is essential to attract as many participants as possible. This is why in this study we used a game with a purpose to collect the data on induced musical emotion. We advertised our game, Emotify, through social networks, and it attracted a big and varied set of participants. Games with a purpose are designed in such a way that the winning strategy is to provide the most correct and precise result possible. It is not possible to exclude vandalism or errors entirely when dealing with human-provided data, but as designers of a GWAP, we tried to minimize the risks, and we will discuss the techniques that we employed for that below.

2.1.1 Organization

This chapter is organized as follows. In Section 2.2, background research related to serious games, emotional models (and, in particular, our model of interest — GEMS) is reviewed. Then we proceed with the experimental part. Section 2.3 presents the procedure (the GWAP) of the experiment. In Section 4, we describe the music we used and the questionnaire. In Section 2.4, we describe the dataset that we collected and released as an outcome of this study. In Section 2.5, we analyze the consistency of responses made using GEMS model, and the feedback and suggestions from game players. In Section 2.7, we analyze the extra-musical factors that influence emotion. In Section 2.8, we discuss the main findings and Section 2.9 concludes the chapter.

2.2 Background

The research in this chapter closely concerns two fields: music information retrieval and music psychology. In the last decade, researchers from both fields actively engaged in research of music-related emotion. We will review the work in this domain that is the most relevant for this chapter. First, we will describe the situation with the datasets for MER. Then, we will talk about serious games and music-related GWAPs. Then we will review dimensional and categorical approaches to representing emotion, and, lastly, experiments which involved the GEMS model and measured its reliability.

2.2.1 Datasets of music annotated with emotion

Data collection for MIR is hindered by two major problems: copyrighted music which can not be freely distributed among researchers, and cost of annotation (annotating a song takes much more time than annotating an image or a sentence, for instance). The biggest datasets collected for MER usually use folksonomy tags instead of being annotated in the lab. Y.-C. Lin, Yang, and Chen (2011) use tags from AMG[5] to create an impressive dataset of 7922 songs annotated with 183 emotional tags (the most popular tags are *confident*, *amiable*, *playful*, *earnest*). The same data source (AMG) was used by the MIREX Mood Classification task (Hu et al., 2008). Hu and Downie (2010a) assembled a dataset of 5296 songs labelled with 18 emotion labels from last.fm (most popular — *calm*, *sad*, *glad*, *romantic*). Though online sources allow to collect much bigger datasets, the online tags are often noisy, and we certainly can not know whether tagger meant induced or perceived emotion. Also, the songs with most labels are usually popular commercial songs, and the audio can not be redistributed. Sometimes the audio features extracted from the audio can be distributed instead of the audio recording, as in the case of the Million Song Dataset (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011), but audio feature development is a very active area of research in MIR (one of the key areas) and absence of audio is a very serious hindrance.

The largest *public* datasets for MER annotated by consistently instructed annotators (in the lab and on Amazon Mechanical Turk) are AMG1608 and DEAM. AMG1608 contains 1608 30-second fragments of music in various music genres annotated by 665 subjects on valence and arousal (Chen, Wang, Yang, & Chen, 2015). DEAM contains 1802 song excerpts and full songs annotated continuously on valence and arousal and is extensively described in Chapter 4. Another large dataset compiled by Y.-H. Yang and Chen (2011b) contains 1240 Chinese pop songs annotated with valence and arousal using rankings scheme (the annotators compared two songs on arousal and valence). The latter dataset is not public.

All the mentioned datasets have perceived emotion annotations (or it is unknown whether emotion is perceived or induced). There is however a much smaller dataset of 195 song excerpts annotated with induced emotion (valence and arousal) (Y.-H. Yang et al., 2008), but the data is not public. The only sizeable (120 one minute excerpts) and public dataset annotated with induced emotion is the DEAP dataset, created by Koelstra et al. (2012). In this dataset, each clip was annotated by 14–16 listeners (50% female), who were asked to rate the felt valence, arousal and dominance on a 9-point

[5] allmusic.com

scale for each clip. However, DEAP contains annotations for music videos and not music alone.

2.2.2 Serious gaming

Serious games, i.e., games that have non-entertainment purposes, have found multiple applications in health-care (Deponti, Maggiorini, & Palazzi, 2009), education (Gaggi, Galiazzo, Palazzi, Facoetti, & Franceschini, 2012), and professional training (Backlund, Engstrom, Hammar, Johannesson, & Lebram, 2007). Serious games are often perceived as a type of edutainment (Ratan & Ritterfeld, 2009), for which there can be a diversity of goals: acquiring new skills, theoretical knowledge application in a simulation of a real world situation, or even informing oneself about a particular political situation. Serious gaming comprises all games that pursue goals other than entertainment (Ratan & Ritterfeld, 2009). The non-educational type of serious games is called 'games with a purpose' (or GWAPs). These games normally have the purpose of gathering data from participants as a form of crowdsourcing.

The term and concept of 'gaming with a purpose' was first suggested by Luis von Ahn, a pioneer in the area of human-based computation games (Ahn & Dabbish, 2004), who introduced the ESP game in 2004. The ESP game is a competitive two-player game, whereby people provide labels for the pictures and score points by guessing the same answer. Google purchased a license to create its own version of the game in 2006. The data collected by the ESP game have been used to improve image search and image recognition algorithms.

GWAPs for music research

Games with a purpose are relatively popular in MIR. For some tasks, such as tonality or chord labeling, musical experts are needed to perform the annotation task. For other tasks, such experts are unnecessary and sometimes even undesirable. This is the case when annotating music with tags or emotions, or measuring music similarity, because for these tasks researchers are more interested in variation in listeners' perception in general, as opposed to the theoretically more consistent opinions of the experts.

Hence, there are several reasons why games with a purpose are especially suitable for data collection in the realm of music:

1. The commonsense expertise that every adult music listener possesses is usually sufficient to participate.

2. Listening to music is pleasant and self-rewarding and therefore it is easy to create engagement.

3. Sometimes it is simply infeasible to collect data in other ways.

In 2008, a game called TagATune, similar in design to ESP, was created to enable music annotation with tags (Law, Ahn, Dannenberg, & Crawford, 2007). TagATune was designed to produce tags that would be much less subjective than those one could obtain from social music websites like last.fm. In TagATune, a player is randomly paired with a partner, both of whom must label a short (thirty-second) musical excerpt with a series of tags. Based on their opponent's tags, players must guess whether

they and their opponent have listened to the same song or not. In such a setup, tags referring to personal musical taste or subjective associations with music will naturally be avoided by players, as such tags are unlikely to provide useful information to a random opponent. Some of the tags, collected by the game, were also mood-related. MajorMiner and HerdIt! (Mandel & Ellis, 2008; Barrington, O'Malley, Turnbull, & Lanckriet, 2009) are similar to TagATune in design and purpose, but HerdIt! uses Facebook as a platform and supports multiplayer games.

A specifically emotion-targeted GWAP called MoodSwings, for continuous emotional annotation of music, was created by Kim, Schmidt, and Emelle (2008). In this game, players are paired up with a partner and both of them mark the perceived musical emotion on a per second basis on the valence–arousal plane. They earn points by guessing their opponent's position on the valence–arousal plane for the same fragment of music. The GWAP we present, Emotify, is different from MoodSwings in several respects: it collects data on induced (not perceived) emotion, the measurements are discrete rather than continuous, and it uses a categorical emotional model instead of a dimensional one.

2.2.3 Models of musical emotion

In this section we will review different ways to conceptualize emotion, with a particular emphasis on emotions in music. Juslin and Västfjäll (2008) define emotions as follows: "Relatively intense affective responses that usually involve a number of subcomponents — subjective feeling, physiological arousal, expression, action tendency, and regulation — which are more or less 'synchronized'."

Several areas of science, such as psychology, musicology and neuroscience, have come up with general or domain-specific models of emotion. These models can be divided in two groups: categorical and dimensional models. Categorical models present emotions as consisting of several basic clusters. Dimensional models present emotions as changing along several orthogonal dimensions.

Categorical models

Categorical emotion models are based on the idea of existence of a certain number of basic emotions. Based on his studies of facial expression, Paul Ekman suggested that there are 6 basic emotions: anger, disgust, fear, happiness, sadness, and surprise (Ekman, 2005). Plutchik proposed eight basic emotions: anger, disgust, fear, joy, sadness, surprise, anticipation and trust (Plutchik, 1980), and arranged them in a circular diagram (Figure 2.1). Each emotion has gradations of strength (e.g., weak — *serenity*, medium — *joy*, and strong — *ecstasy*). Opposite emotions (e.g., *interest* — *distraction*) are situated against each other on the diagram. There are also eight additional derivative emotions added that are each composed of the two basic ones. For instance, *annoyance* combined with *boredom* create *contempt*.

Categorical models are problematic, because emotion words may not have exact translations in different languages. This problem does not exist in the case of dimensional models.

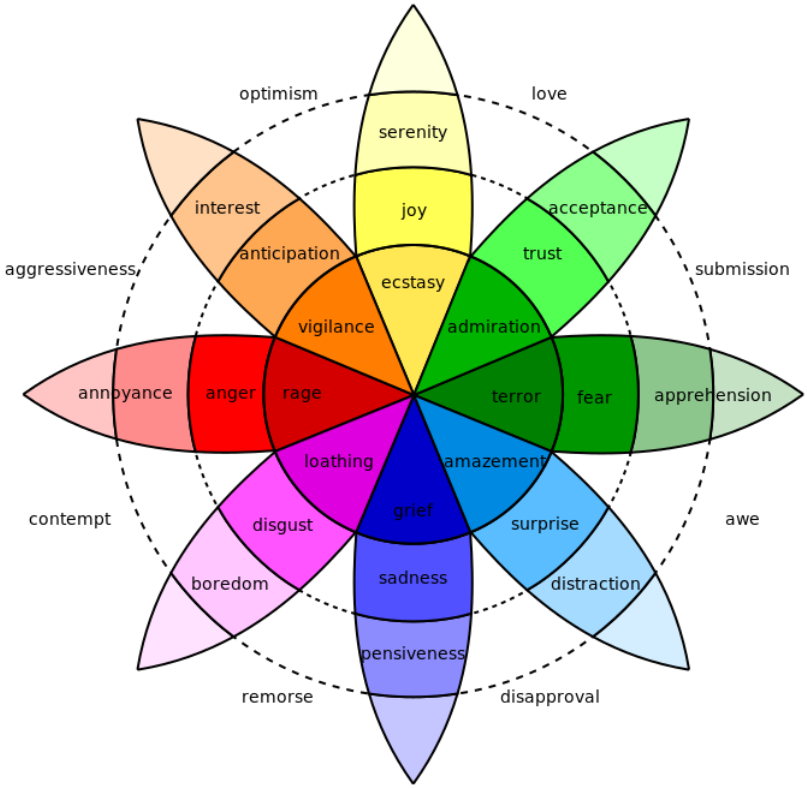


Figure 2.1: Plutchik's wheel of emotions. Source: (Plutchik, 2001).

Dimensional models

Dimensional models arrange emotions in a continuous space along several (usually two or three) principal dimensions. Russell (1980) proposed a model that consists of two dimensions: valence (ranging from unpleasant to pleasant) and arousal (ranging from passive to active) (Figure 2.2). This is the most widely used emotional model in Music Information Retrieval. A. Mehrabian and J.A. Russel proposed that perception, experience, and psychological responses to environmental stimuli are described by a three dimensional model: pleasure (valence), arousal and dominance (PAD) (Mehrabian & Russell, 1974).

The valence–arousal (V–A) model is often criticized for its lack of granularity. For instance, anger and fear are placed very close to each other in the upper left quadrant of the valence–arousal plane. Emotions can also be contradictory (e.g. bitter-sweetness) (Hunter, Schellenberg, & Schimmack, 2008). It is impossible to present these on the valence–arousal plane. Though the valence–arousal model is very popular with MIR researchers, many have concluded that V–A model fails to capture all the variance reflected by music (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005; Collier, 2007; Ilie & Thompson, 2006; Fontaine, Scherer, Roesch, & Ellsworth, 2007).

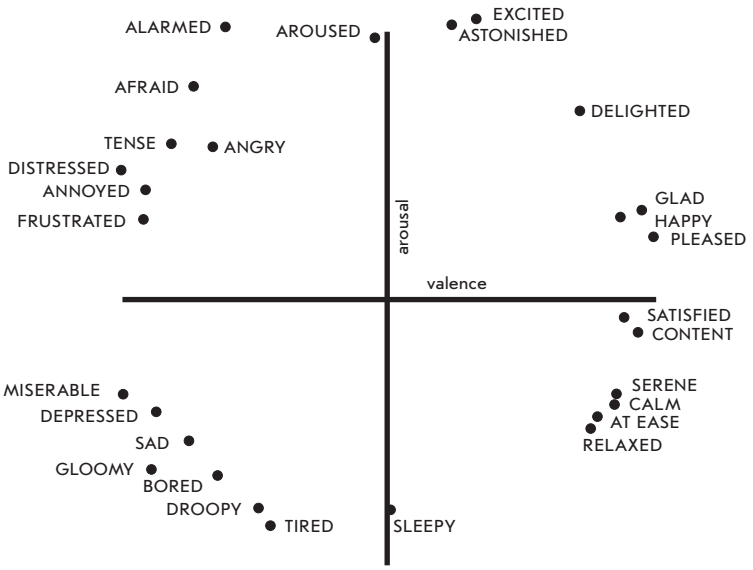


Figure 2.2: Valence and Arousal model. Multidimensional scaling of 28 affect words marked on the two-dimensional plane. Source: (Russell, 1980).

Music-specific emotional models

In the case of music it might be crucial for a model to be domain-specific. Scherer (2004) argues that everyday *utilitarian* emotions should be distinguished from *aesthetic* emotions, induced by works of art. Aesthetic emotions are usually much more

subtle, and don't coincide with everyday emotions (for instance, shame or guilt are almost never felt in response to music (Zentner et al., 2008)).

The earliest attempt to create a specifically musical categorical model of emotion was undertaken by Hevner (1936). She created an ontology of eight emotional clusters, where each cluster contained from six to eleven adjectives (see Figure 2.3). This list was later revised by Farnsworth (1958), the adjectives were rearranged and a ninth cluster was added (sacred, spiritual). The list was revised again by Schubert (2003). There were about a dozen other studies that identified between 3 and 12 clusters (Asmus, 2009). The clusters are very different in terms of vocabulary and structure, which should not be very surprising as over time words change their meaning (gay), fall out of fashion (stately, benevolent), and the emotional content of music changes as well.

The categorical models used in Music Information Retrieval research usually possess few (4–6), but sometimes as many as 18 classes (Y.-H. Yang & Chen, 2012). Some of these models are derived from online tags or surveys using techniques such as hierarchical clustering or latent semantic analysis (Skowronek, McKinney, & Par, 2006; Laurier, Sordo, Serra, & Herrera, 2009; Hu et al., 2008).

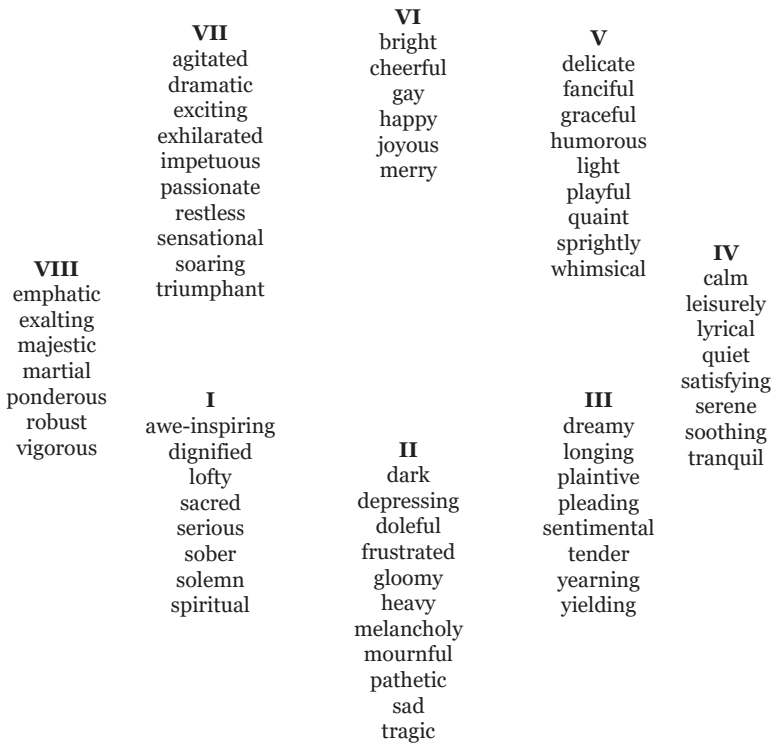


Figure 2.3: K. Hevner's adjective circle (Hevner, 1936).

Several two or three-dimensional emotional models were experimentally derived for music. The dimensions are similar in meaning: tension–energy, gaiety–gloom, and

solemnity–triviality (Wedin, 1969, 1972a), intensity–softness, pleasantness–unpleasantness, and solemnity–triviality, obtained using multi-dimensional scaling (Wedin, 1972b), tension–relaxation, gaiety–gloom, and attraction–repulsion (Nielzen & Cesarec, 1969); gay/vital–dull, excited–calm (Gabrielsson, 1973); 'Heiterkeit–Ernst' (cheerful–serious) and 'Robustheit–Zartheit' (strong/powerful–soft/tender) (Kleinen, 1968). For a more comprehensive review of models of musical emotion we refer to (Juslin & Sloboda, 2011).

None of the models described before were specifically developed for induced emotion. In 2008, a new domain-specific categorical emotional model called GEMS (Geneva Emotional Music Scale) was proposed by Zentner et al. (2008). GEMS is unique in that it addresses induced emotion, was created specifically for describing musical emotion, and has a level of granularity that other models do not provide. Zentner et al. conducted four consecutive studies to derive the model. First, a list of music-related terms was compiled both for induced and perceived emotion. It showed that these two types of emotion differ from each other, the major difference being the bias for positive emotions in case of induced emotions. In the following studies, a structure of music-induced emotions was examined through factor analysis of questionnaire data. As a result, the GEMS scale was created.

Through further factor analysis, shorter versions of the scale were added. The full GEMS scale consists of 45 terms, with shorter versions contain 25 and 9 terms. These nine terms can in turn be grouped into 3 superfactors: vitality, sublimity and unease. Originally, the terms were collected in French, and later translated to English. In 2012, an additional research was conducted to improve the short GEMS scale (Coutinho & Scherer, 2012). In this research, the problem of classical music overrepresentation in the original work behind GEMS was addressed. The experiment confirmed the nine-factor structure of GEMS. It was suggested to add new terms related to feelings of harmony, interest and boredom. The final results from the study were still unpublished at the time when the game was created, so we used the original short nine term version of GEMS for our online game (Wonder, Transcendence, Tenderness, Nostalgia, Peacefulness, Power, Joyful Activation, Tension and Sadness).

Induced vs perceived emotion

There is no doubt that music can reliably communicate emotions and listeners can reach some degree of agreement on them, as demonstrated, e.g., in (Eerola & Vuoskoski, 2011; J. Vuoskoski & Eerola, 2013; Bigand et al., 2005; Torres-Eliard et al., 2011). Some doubts were expressed whether instrumental music can induce emotions, most notably by P. Kivy, but also by E. Payne and C. Pratt (Pratt, 1952; Payne, 1961; Kivy, 1990, 1993). This issue is still sometimes debated, but in contemporary music psychology, the most supported position is that music can induce emotion (Konečni, 2008; Scherer, 2004; Krumhansl, 1997; Rickard, 2004). In a study conducted by Alf Gabrielsson, more than thousand respondents provided reports on their strong experiences with music, and almost any emotion can be found in those reports as having been elicited by music (Gabrielsson, 2011). Wager et al. demonstrated in an extensive meta-analysis of PET (Positron Emission Tomography) and fMRI (Functional Magnetic Resonance Imaging) studies that perception and induction of emotions involve

'peak activations' of different areas of the brain, supporting the idea that these are different processes (Wager et al., 2008).

The relationship between perceived and induced emotion is not straightforward. Gabrielsson (2002) argues that perceived and induced emotion can relate in four ways: *positive*, *negative*, *no systematic relation* or *no relation*. Positive relation, though being most common, should not always be assumed. Also, though a qualified listener can nearly always recognize emotion expressed in the music, emotion induction is less frequent. Recent studies suggest that listeners experience strong emotions only about 55% of the time they spend listening to music (Juslin & Laukka, 2004), or that in 65% of the musical episodes music affects how they feel (Juslin, Liljeström, Västfjäll, Barradas, & Silva, 2008). It was also demonstrated that negative types of emotions (anger, fear) are less likely to be induced by music, though music can express them (Zentner et al., 2008).

Emotional responses can be measured from self-report, expressive behaviour and physiological responses (heart rate, skin conductivity, blood pressure, as well as biochemical responses) (Krumhansl, 1997; Rickard, 2004). In case of music, pronounced expressive behaviour is not the rule. Measuring physiological responses can help to verify the presence of induced emotion, because in case when emotion is only perceived, but not felt, the response may be more subtle. Most studies show that listening to music can create pronounced physiological response — change in heart rate, hormone levels, skin conductivity (many such studies are reviewed in (Hodges, 2011)). Unfortunately, the results from different studies so far are very contradictory and it is difficult to make definitive conclusions about the actual influence of music on physiological response. Self-report is the most widely used measure. For obtaining fine-grained measurements on musical emotion, self-report is also arguably the most informative measure, because it provides information on the otherwise inaccessible cognitive part of emotion (Zentner & Eerola, 2011b).

2.2.4 Experiments involving the GEMS model

In this section we will describe some of the research, where GEMS was used as an underlying model for data collection.

The biggest experiment, involving nearly 4000 participants, took place in 2010 (Jaimovich, Coghlan, & Knapp, 2012) in Dublin. Participants listened to music and reported their emotional state, using several self-assessment methods, GEMS among them. Physiological measurements were also recorded. The dataset contained 53 songs from different genres (rock, classical, pop, jazz, world etc.), specially selected for their emotional content. The analysis of the collected data is presented in the PhD thesis of Jaimovich (2013). Unfortunately, due to a software error, the answers to the GEMS questionnaire had to be discarded.

In 2010, Vuoskoski et al. performed a comparison of three emotional models: one dimensional (valence–arousal–tension), one categorical (anger, fear, happiness, sadness, tenderness), and GEMS. Only 16 excerpts from movie soundtracks were used (J. K. Vuoskoski & Eerola, 2010, 2011). The most consistent ratings were produced in the case of the two-dimensional valence–arousal model, while basic emotions and GEMS were less consistent, with GEMS's possessing both the most consistent (joyful activation, tension) and inconsistent (wonder, transcendence) categories. It was also

found that GEMS categories are redundant, and valence and arousal dimensions account for 89% of variance. That experiment, though, was performed on 16 musical excerpts only, and the excerpts were selected using criteria based on V–A model, which might have resulted in a bias. In 2011, K. Torres-Eliard et al. used GEMS for continuous emotion measurements (Torres-Eliard et al., 2011). Every rater controlled one GEMS dimension. Data on emotion *expressed* in 36 musical excerpts were collected. The inter-rater agreement (based on the extent to which a single emotion was present in the music at a given moment of time) was found to be in the range of good agreement (Cronbach's alphas ranged from 0.84 to 0.98). In (Lykartsis, Pysiewicz, Coler, & Lepa, 2013), the GEMS-25 model was tested for invariance across genre (classical vs. electronic music) and language (English vs. German version GEMS-28-G), by examining the data using confirmatory factor analysis. The model was found to be configurally invariant across language (same configuration of factor loadings on latent variables) and weakly invariant across genre (configurally invariant and the values of loadings are also equal) for German version only.

In (Choppina et al., 2016), both patients with bypolar disorder and a healthy control group could assign musical excerpts the "correct" (defined by researchers) emotion on GEMS scale for slightly more than half of the excerpts. In (Baltes et al., 2011), GEMS was used in a study of operatic performance, and this self-report measure showed significant correlation with physiological parameters (such as systolic blood pressure, respiratory sinus arrhythmia, etc.). In (Kaelen et al., 2015), it was shown that emotional response to music is enhanced by LSD, especially the emotions from the GEMS scale named wonder, transcendence, power and tenderness. Pearce and Halpern (2015) studied age differences in emotional reactions to music, measuring those reactions using GEMS-9, and found significant age-related effects in peacefulness, sadness and tension.

In the original study that introduced GEMS (Zentner et al., 2008), a small-scale experiment with 16 classical pieces showed that GEMS equips listeners with a more adequate instrument to measure musical emotion and results in better agreement than V–A or basic emotions. As this experiment was very small, based on one genre only, and the questions asked for V–A model were unconventional, this finding needs further investigation. In all the studies that we described above the datasets were not big, and the data are not publicly available.

2.3 Emotify: game design

In this section we will start describing the experimental part of this chapter. We created a GWAP called Emotify to collect labels using GEMS.

Emotify was launched on the 1st of March 2013. As a platform, we used both a social network (a Facebook application[6]) and a stand-alone website[7]. Using a social network as the platform for a GWAP simplifies dissemination, but for those who do not possess or want to use or create a Facebook profile, we provided a stand-alone version. Figure 2.4 shows a screenshot of the game interface. Involving a social network gave

[6] apps.facebook.com/emotify
[7] emotify.org

us the possibility to provide users with inter-player comparison in a non-competitive manner.

Figure 2.4: Emotify interface. Calmness and tenderness are selected and highlighted. An explanation for the hovered button is shown below the buttons (Sensuality, affect, feeling of love).

The gameflow is as follows.

1. The player authenticates through Facebook (or alternatively, enters the game from the stand-alone website) and provides personal details: age, gender, musical preferences, first language, level of English, and current mood. At this stage, the player is also provided instructions and is asked to report his or her personal emotions in response to music (i.e., induced emotion).

2. The player is randomly assigned to one of four musical genres (rock, pop, classical and electronic music) and can switch to any other if he or she so wishes. The player may also switch at any later time.

3. In every genre, the player is presented with a random sequence of musical excerpts, each one minute in length. If a player is invited by a friend through Facebook, he or she is presented with the same (whenever possible) sequence as the player who sent the invitation. This constraint is necessary in order to enable comparison between them.

4. After listening to the one-minute fragment, the player selects up to three emotions from a list of nine. This limitation should encourage players to think more carefully about the choices and name only the strongest emotions.

5. The player also may indicate whether he or she liked or disliked the music and whether he or she knows the song. The player may also provide a new emotion definition if none of the nine corresponds to what he or she is feeling.

6. At any time, it is possible to skip listening and go to another song or another genre.

7. There is a countdown from 10 to 1, saying that after 10 fragments the player will receive final feedback on his or her emotional perception of music. The countdown should encourage players to listen to at least 10 fragments to earn a "reward". Players may continue after listening to 10 fragments, but we prefer them not to do so, because feeling emotional content of music requires concentration and sensitivity, which is difficult to maintain for a long period of time.

Before starting with the game, the players were explained that they will be asked to describe what they feel in response to music, and also they were encouraged to skip the song if it fails to elicit any emotions in them. We tried to encourage more personal induced emotion responses by providing feedback in a style of psychological questionnaire.

2.3.1 Incentives

When discussing related work, we mentioned MoodSwings, a game for emotional annotation of music. MoodSwings focuses on perceived emotion (as opposed to induced emotion), which is also apparent in their choice of emotional model and their method of data collection. In our game, in contrast, we are trying to collect induced emotion annotations. For a game with a purpose, induced emotion creates a design problem. A standard type of player engagement in GWAPs is making players compete over giving the most uniform answer possible (which is also supposed to be the correct one). In case of induced emotion there is no correct answer, and it would be misleading to encourage the listener to look for one. Induced musical emotion is by its nature personal and subjective.

This is why we introduced a different fun element than competition. We decided to create engagement by providing a feedback on a player's answers in the manner of a psychological quiz. In addition, we decided to use a social network in order to give the player a possibility to compare his musical tastes and perception to those of his or her friends in the social network. There were three feedback elements in the game:

- **Continuous** feedback: A score calculated as a correlation of the player's answers to the averaged answers of other players. This score is recalculated after every answer and averaged over all answers.

- **Final** feedback: A histogram of the player's emotional responses for songs that the player liked or disliked. This feedback was provided only after completing 10 songs and was used to stimulate the user to continue playing.

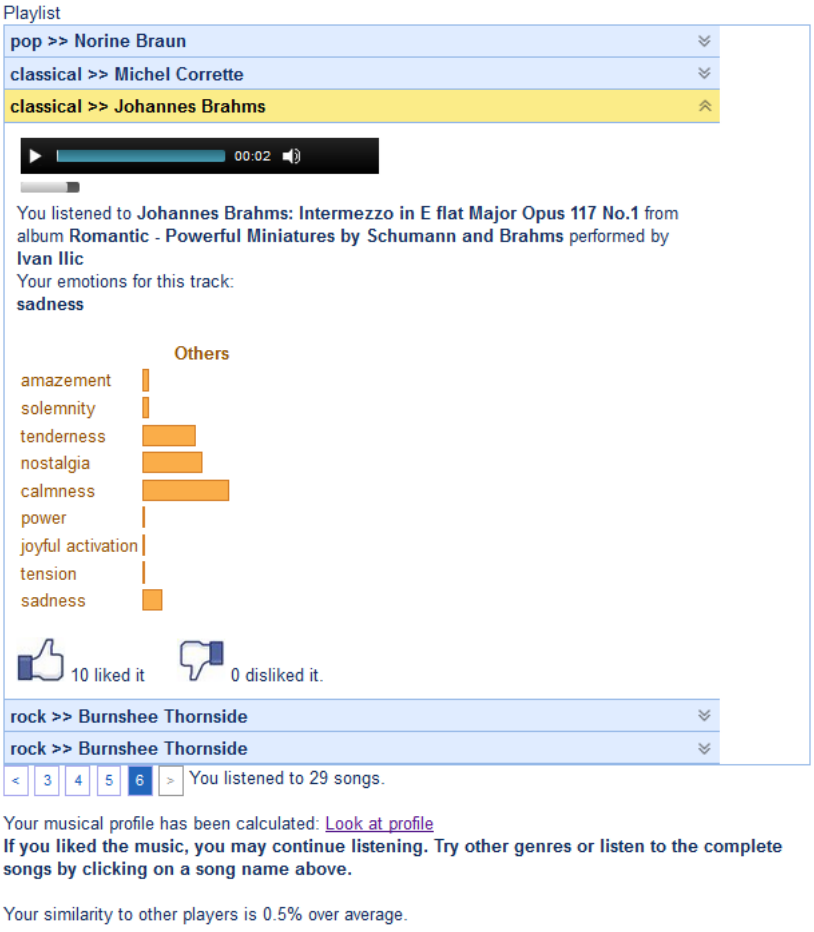


Figure 2.5: Emotify interface. Playlist feedback.

- **Playlist** feedback: Feedback on every song to which the player listened. Players had the possibility to listen to the whole song (not just the initial one-minute excerpt), and to see a detailed comparison to other players (see Figure 2.5).

We hope that by designing such a feedback scheme, we encouraged players to give sincere and serious answers and at the same time provide a reward for their contribution.

2.4 Methods

In this section, we will explain the details on the selection of the musical material and the modifications to the GEMS questionnaire that we made.

2.4.1 Music

In existing research on musical emotion, music is often selected for its strong and obvious emotional content (J. K. Vuoskoski & Eerola, 2010; Jaimovich, 2013; Zentner et al., 2008). In such a case, it is unclear how obtained results are comparable to non-preselected music. In our experiment we intentionally chose music randomly from a larger collection. We assembled a set of 400 musical pieces from the Magnatune recording company (magnatune.com), 100 pieces from each of four selected genres (**classical**, **rock**, **pop** and **electronic**). Genres were assigned by the recording company. The resulting dataset contains music from 241 different albums by 140 performers. There were several reasons to choose music from Magnatune: the quality of sound recordings is generally good and the music little known, except for classical genre (familiar music might precondition induced emotion (Schubert, 2007)). The music was reviewed manually and some recordings (around 2%) were removed because of insufficient quality.

We randomly divided our musical corpus into two subsets, maintaining the genre ratio (15 songs from each of the four genres). The smaller subset of the data (which will be called **subset A**) consists of 60 songs. The remaining 85% of the corpus (**subset B**) consists of 340 songs. We collected different amounts of annotations for **subset A** and **subset B**. The smaller subset was intended to be used to calculate the listener agreement on the GEMS categories, the bigger subset was intended to be distributed as a public dataset of annotated music together with smaller subset. In **subset A**, each song is annotated with at least 10 measurements per variable, which makes it at least 90 labels per song, since there are nine questions in the questionnaire. We count all labels given to a song independently, thus if a person assigns 2 labels to a piece of music, we count each. On average there were 48 annotators per song in **subset A**, and for **subset B**, at least 10 people listened to and annotated each song (on average — 18) (see Section 2.5.2).

2.4.2 Questionnaire

GEMS questionnaire adaptations

In order to adapt the GEMS questionnaire to an online game, we made several modifications. Originally, GEMS is designed to be answered using a Likert scale ranging

Superfactor	Emotional category	Explanation
Sublimity	Amazement*	Wonder, happiness
	Solemnity*	Transcendence, inspiration, thrills
	Tenderness	Sensuality, affect, love
	Nostalgia	Dreamy, melancholic, sentimental
	Calmness*	Relaxation, serenity, meditativeness
Vitality	Power	Strong, heroic, triumphant, energetic
	Joyful activation	Dancing, bouncy, animated, amused
Unease	Tension	Nervous, impatient, irritated
	Sadness	Depressed, sorrowful

Table 2.1: GEMS categories with explanations as used in the game. The categories marked with asterisk were modified.

from 1 to 5. The Likert scale is a psychometric scale commonly used in questionnaires. When answering a question using a Likert scale, a participant has to choose one of several items typically ranging from "Not at all" to "Very much". This way of data collection is, however, very slow, requires quite some mental effort, and is not suitable for a dynamic online game. Therefore we modified the task and asked to select several labels from a list instead. This means that for each emotion we obtain one value, which is either 1 or 0 (emotion is present or not), which results, for each song, in a vector of 9 binary values.

We also restricted users on how many labels they could select, by instructing them to select no more than 3 labels. We did this because we wanted the players to select only the strongest emotions. As we abandoned the Likert scale, limiting the number of responses was the only way to measure the strength of emotion.

Following the findings from (J. K. Vuoskoski & Eerola, 2010; Torres-Eliard et al., 2011), where it was discovered that participants have trouble with understanding certain categories of GEMS, we changed the wording of three GEMS categories by replacing them with one of the emotions from the list of explanatory synonyms that accompany each GEMS category. *Transcendence* was changed to *solemnity*, *wonder* to *amazement*, and *peacefulness* to *calmness* (see Table 2.1).

Personal questions

We collected the following personal data about participants: age, gender, first language, level of English (Beginner, Intermediate, Advanced), musical preferences (we asked the participants to report their preferences on the four selected genres (binary, whether a genre is liked by participant or not), and added an open question where other preferred genres could be indicated), and current mood (on a Likert scale from 1 (*in a very bad mood*) to 5 (*in a very good mood*)).

Other information

For every piece of music the participant listened to we collected, apart from the data described above:

- Whether the participant is familiar with the piece (binary).
- Whether the participant liked or disliked the piece (binary).
- The order in which GEMS categories were presented to the participant (randomized between participants).
- Optionally, a suggestion of a new emotion label not represented by GEMS, or an explanation of choices that participant made.

2.5 Annotations

Creating a publicly accessible dataset was one of the main motivations of this study. All the data is available at the following link[8]. Below, we list statistics on game players and describe the size and contents of our dataset.

2.5.1 Participants

1778 participants (747 females, 1031 males) took part in the study and 16191 labels were collected for 400 songs during 8358 listening sessions. The average age of participants was 30.32 years (sd = 11.74). Participants listed different languages as their first language: 38% English, 19% Dutch, 19% Russian, the remaining 24% of the participants indicated 41 other languages (mostly European, with some Chinese, Hindi, etc.). The style preferences were as follows: 61% Rock, 55% Classical, 44% Pop and 43% Electronic (multiple genres were allowed). 11% of the participants reported that their English language proficiency was on the beginner level, 26% were on intermediate level and 63% were advanced. On average, they listened to 8 songs, and spent 13 minutes and 40 seconds playing the game (sd = 12.62). The actual time spent in the game differed a lot over all players. As we were advertising a game through online media, there were many players who merely examined the game and quit almost immediately, but there were also devoted players who spent a lot of time listening to music. Overall, the players gave positive feedback to the game. In the experiment, participants had to select one, two or three main emotions they felt after listening to a one minute excerpt. For 37% of samples they selected only one emotion, 30% obtained two emotional labels and 33% three emotional labels. There were no complaints about not being able to select more than three labels, but some participants reported not being able to find exact emotion they felt in GEMS model, and about 7% of participants suggested new emotion labels.

2.5.2 Amounts of annotations

The annotations produced by the game are spread unevenly among the songs, which is caused both by the design of the experiment and the design of the game. Participants could skip songs and switch between genres, and they were encouraged to do so, because induced emotional response may not occur on every music listening occasion. Therefore, less popular (among our particular sample of participants) genres received

[8] www.projects.science.uu.nl/memotion/emotifydata

fewer annotations, and the same happened to less popular songs. A density histogram on Figure 2.6 illustrates the spread of annotations among 400 songs. On average, each song from **subset A** was annotated by 48 participants (sd = 4.54) and each song from **subset B** by 18 participants (sd = 7.8).

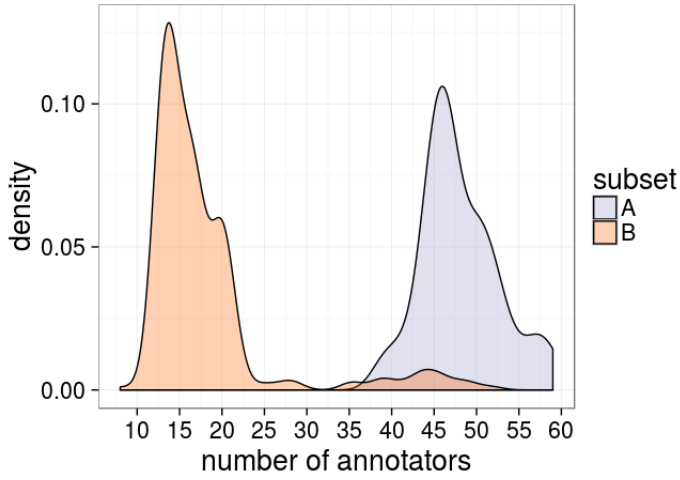


Figure 2.6: Density curves for number of annotators per song for subsets A and B.

2.5.3 Confounding factors

Influence of button order on the frequency of button selection

For each of the participants, positions of the buttons (the nine buttons with emotional labels on them) in the game interface were randomized. The buttons were placed in a three by three grid, as shown on Figure 2.4. We need to verify whether the buttons in certain positions were selected more often than buttons in other positions (regardless of the text on the button). Table 2.2 shows the frequencies of button selections in listening sessions. The position in the table corresponds to the button position on the screen. Since up to three emotions could be selected during one listening session, the percentages do not sum up to 100.

2021 = 19%	1840 = 17%	1969 = 19%
1916 = 18%	2027 = 19%	1946 = 19%
1804 = 17%	1862 = 18%	1816 = 17%

Table 2.2: Frequency of button selection (the absolute number of clicks and a percentage from all the listening sessions).

The buttons in the lowest row were selected less frequently than the buttons in the upper and middle row (about 7% less than the buttons above them). We conducted a Chi-Squared test on these frequencies and found that these differences are higher than chance: $\chi^2 = 13.309$, df = 4, p-value = 0.0098. We conclude that the quality of

the annotations would be compromised, should the order of the buttons have not been randomized.

Influence of English language proficiency

For almost 62% of the game participants, the first language was another language than English. From these participants, 15% indicated that their level of English fluency is "Beginner", 36.5% indicated "Intermediate" and 48.5% "Advanced".

The group of beginner-level participants was too small for their answers to be separately compared to the other groups. This is why we studied the effect of removing those participants and computed intra-class correlation coefficients for each of the songs with and without beginner-level participants. Removing "beginners" didn't affect intra-class correlation coefficients significantly. Also, in Section 2.6.4, we estimated the expertise of each labeler. The expertise estimation is based on consistency of responses. If there were language understanding problems for Beginner level participants, their responses would be less consistent. We used expertise estimation to compare Beginner level participants with the other groups and didn't find any significant difference in means of expertise.

This leads us to believe that the level of their understanding was satisfactory enough for their answers to not degrade the quality of our dataset.

2.6 GEMS model comprehensibility

One of the objectives of our experiment was to test whether the GEMS model is suitable for large-scale music retrieval systems. In our game, we involve a varied sample of participants from different age groups and linguistic backgrounds, which resembles an actual composition of the users in online music services. We ask them to provide feedback on using GEMS, and we also compute implicit consistency measures.

2.6.1 Feedback questionnaire

After completing 10 excerpts, game players could view their scores and were asked to fill in a feedback questionnaire. 556 participants did so. They were asked to rate how difficult it was to use GEMS on a scale from 1 to 5 (where 1 means "very easy") and on average they gave rating of 2.92 (sd = 1.07, mode = 3). On average they rated their liking of music on a scale from 1 to 5 (where 1 means "disliked completely") as 3.16 (sd = 1.08, mode = 4). Also, participants were asked to indicate which GEMS categories were most difficult to understand and to associate with the emotions they felt (see Table 2.3, column 2).

From the feedback we can see that we did not manage to improve the situation (J. K. Vuoskoski & Eerola, 2010) with categories *wonder* and *transcendence* by giving them synonymous names. A very big number of participants (one third) considered them unclear. The most easy to understand categories were calmness and sadness. Those were also the most often selected ones as well. The rest of the categories were considered unclear by approximately one tenth of the players.

Emotion	Considered unclear	Frequency of selection
Amazement	31%	13%
Solemnity	31%	20%
Tenderness	12%	18%
Nostalgia	10%	26%
Calmness	3%	30%
Power	11%	18%
Joyful activation	11%	25%
Tension	13%	23%
Sadness	4%	30%

Table 2.3: GEMS categories and feedback questionnaire statistics. Second column: considered unclear by percentage of respondents (n = 556). Third column: how often an emotion was selected in the listening sessions (n = 8358).

Emotion	Classical	Rock	Pop	Electronic	Average
Amazement	.70	.36	.31	.48	.46
Solemnity	.48	.70	.72	.72	.65
Tenderness	.75	.86	.85	.85	.82
Nostalgia	.81	.81	.64	.60	.71
Calmness	.92	.72	.90	.78	.83
Power	.90	.87	.82	.82	.85
Joyful activation	.96	.92	.91	.87	.91
Tension	.55	.75	.83	.75	.72
Sadness	.78	.81	.46	.70	.69

Table 2.4: Cronbach's α values per category per genre (subset A).

2.6.2 Listener agreement on emotional categories

We collected especially a big amount of data for one subset of songs (**subset A**), in order to examine the inter-rater agreement. Here we will analyze these songs.

From the 60 songs of **subset A**, only 25 songs possessed at least one emotional category that was selected by the majority (more than a half) of the respondents (the highest percentage of respondents to select a category unanimously was 77%). The most frequent highly selected categories were calmness and joyful activation (both for 8 songs), tension (7 songs), and the least frequent were power, nostalgia and tenderness. The rest of the categories (amazement, solemnity and sadness) in most cases weren't selected by more than one third of the participants unanimously. Though most of the songs failed to reach a majority vote on any of the emotional categories, all of the songs demonstrate agreement that is much better than chance.

To assess agreement, we calculated Cronbach's α per category (see Table 2.4). Cronbach's α is a coefficient of internal consistency, commonly used in psychometric tests (Cronbach, 1951). Technically, Cronbach's α is the mean of all possible split-half reliabilities, corrected for test length. Split-half reliability test divides the test in two halves, computes their means and calculates the correlation between them.

Standardized Cronbach's α is calculated as follows:

$$\alpha = \frac{K\bar{r}}{1+(K-1)\bar{r}}, \tag{2.1}$$

where K is the number of components in the test, and $\bar{r}$ is the mean of all the non-redundant correlation coefficients (i.e., excluding self-correlations and correlations computed in reverse order). In our case, we are interested in correlations between the answers of the players for the same song.

In psychological research, Cronbach's α above 0.7 is viewed as acceptable agreement, and three categories do not pass that threshold: amazement, solemnity and sadness. There were known problems with categories amazement and solemnity and they were also marked as least understood categories on feedback questionnaire. Sadness, however, didn't arise questions in most participants (96%). In recent research by Peltola and Eerola, it was shown that a range of emotions, experienced in response to music, is named under a joint umbrella word "sadness": Sweet Sorrow, Melancholia and Grief (Peltola & Eerola, 2016). We hypothesize that our participants did not really agree on the meaning of that category. We conducted the Tukey HSD test on Cronbach's α values between genres and didn't find any significant differences. Tukey HSD test is essentially a t-test which allows multiple comparisons and corrects for an increased probability of making a Type I error.

2.6.3 Suggestions to modify the model

Players were given the opportunity to suggest a new emotional term that was missing from the model, or comment on existing ones. We received 437 such comments. Of them, 125 comments suggested new emotional terms, and the rest explained the reasons behind choosing from a list of GEMS terms or contained other notions.

Table 2.5 lists the most frequent semantic groups of comments, ordered by popularity. As we can see from the table, by far the most frequent suggestion is not related to emotion induced by music but to disliking it — boredom. In groups 1 and 2 we placed all the comments which referred to the fact that music failed to induce any emotion in the respondent. Though we asked the participants to skip the fragments which did not induce any emotion in them, not all the participants did so. Group 3 contains comments on liking the music. Groups 1–3 confirm the findings of Coutinho and Scherer (2012), who discovered that feelings of interest, boredom ("bored", "indifferent", "weary") and feelings of harmony and clarity are lacking from the model. Indeed, when reporting their induced emotion, participants find it very important to be able to report their interest, engagement and enjoyment (or, on the other hand, indifference, boredom and irritation from disliked music). These emotions can also be regarded as music-induced emotions and should be included in the model.

Other semantic groups of comments, not related to liking or disliking the music, are introduced in groups 4 to 8. *Anger* (group 6) and *fear* (group 7), according to Zentner et al. (2008), are often expressed by music, but are unlikely to be induced by it. It is possible that respondents were confusing what they perceive in music and what it induces in them. Impetus (group 4) was the next most suggested semantic group after feelings of interest and boredom. Less frequently suggested semantic groups were humour (group 5) and contentment (group 8). These emotions, along with boredom and interest, are not covered by GEMS as well.

Group	Category	Examples	Occurrence frequency
1	Disliking the music	boring, boredom, bored annoyance, annoyed, ennui	68
2	Neutral	neutral, no emotion, indifferent	10
3	Liking music	interesting, nice, good	10
4	Impetus	anticipation, determined, hopeful, impatient, call to action	8
5	Humour	humour, humorous, sarcastic, silly	7
6	Anger	aggression, anger, wild	6
7	Fear	scared, fear, tense scene in a movie	6
8	Contentment	content, contented, satisfied	5

Table 2.5: Emotions that were suggested by players of Emotify.

Some suggestions that only occurred once were "religious" and "awkward".

2.6.4 Factor analysis and correlation analysis of the data

In this section we analyze the relationships between GEMS categories and conduct factor analysis of the data.

Averaging labels across participants

The data produced by the game is in form of vectors with binary values per player per song. There are several possible ways to average these data per song across players.

In the original GEMS questionnaire, the responses are collected using Likert scale. In our game, we ask the players to indicate only the strongest of the experienced emotions. Therefore, we can assume that the emotions that are mentioned more often with regard to a particular song, are the ones that are most strongly associated with it, and their strength is proportional to how often they are mentioned.

On this assumption, there are two possible ways to average the scores. An emotion could be given a fixed weight regardless of how many other emotions are selected, according to Formula 2.2:

$$\text{score}^1_{ij} = \frac{1}{n}\sum_{k=1}^{n} a_k, \tag{2.2}$$

where score^1_{ij} is an estimated strength of emotion i for song j, a_k is the answer of the k-th participant on a question whether emotion i is present in song j or not (answer is either 0 or 1), and n is the total number of participants who listened to song j. The other way of averaging is to inversely weight each individual answer based on the number of selected emotions, according to Formula 2.3

$$\text{score}^2_{ij} = \frac{1}{n}\sum_{k=1}^{n}\frac{a^k}{\sum_{z=1}^{9} a^k_z}, \tag{2.3}$$

where a^k_z are the answers to all emotions for song j by participant k, and the rest of the variables are the same as explained for the first score. To find out which score is closer to the original Likert scales questionnaire, we asked 5 labellers to label 15 songs on original Likert scale, averaged their answers, and compared the result to the one obtained by averaging using both methods. On basis of this, we chose the first method to average our labels as it was more similar with Likert-scale answers.

We could also try to infer the strongest emotions for each song, based on the votes from the players. We will do that using a method by Whitehill, Wu, Bergsma, Movellan, and Ruvolo (2009) for integration of labels from labelers with unknown expertise. This method is based on an assumption that labeling process depends on two hidden factors: difficulty (ambiguity) of an item to be labeled $\frac{1}{\beta}$ and expertise of the labeler α. From that we can calculate the probability of a label L_{ij} being equal to a true label Z_j as follows:

$$\Pr(L_{ij} = Z_j \,|\, \alpha_i, \beta_j) = \frac{1}{1 + e^{-\alpha_i \beta_j}}. \tag{2.4}$$

Now we can infer L_i (strongest emotions per song), α and β, using expectation maximization. For more information on implementation see (Whitehill et al., 2009).

We consider an emotional category present in a song, if the estimated probability of this is bigger than 0.5. Only half of the songs (218) have at least one identifiable strongest emotion that labellers agreed on. Moreover, for some of the emotions, the number of songs where these emotions are a "majority vote", is very scarce: 0 for amazement, 11 for solemnity, 16 for tenderness, 9 for sadness, and between 27 and 75 for the rest of the categories. As far as this averaging method seems to lose too much detail from our data, we only use the result of this method in Section 2.5.3 to compare participants' expertise. The expertise in our case is not the expertise of labeling music with emotion, but the expertise of using English language to express the felt emotion with words.

Correlation analysis

We used the score described in Formula 2.2 to average the annotations, and calculated correlations (Spearman's correlation coefficient, as the data is strongly positively skewed) between the GEMS categories (see table 2.6). Before doing correlation analysis, we excluded the annotations from those listening sessions where participants indicated that they disliked the music.

Strong positive correlations mean that the correlated categories were either often selected together by the same annotator, or were often selected by different people for the same music. Therefore, these emotions can either be co-occurring, or could also be confused and potentially redundant categories. Prominent examples are: tenderness and nostalgia with $r = 0.55$ and $p < 0.001$ (compare to (Zentner et al., 2008) where $r = 0.5$), power and joyful activation with $r = 0.41$ and $p < 0.001$ (compare to (Zentner et al., 2008) where $r = 0.38$). The strongest correlations are negative: sadness and joyful activation with $r = -0.64$ and power and calmness with $r = -0.64$.

	Sol.	Tend.	Nost.	Calm.	Power	J. act.	Ten.	Sad.
Amazement	−.08	−.13	−.23	−.26	.16	.41	−.11	−.32
Solemnity		−.17	−.18	.08	.06	−.26	−.20	.14
Tenderness			.51	.53	−.59	−.38	−.50	.26
Nostalgia				.43	−.49	−.41	−.45	.42
Calmness					−.64	−.51	−.41	.25
Power						.41	.42	−.28
J. activation							.03	−.64
Tension								−.03

Table 2.6: Correlations between emotional categories.

Factor analysis

Zentner et al. (2008) used factor analysis on the GEMS ratings and found that 9 GEMS factors could be accounted for in terms of three higher-order superfactors, which they called Sublimity (highly correlated with wonder, transcendence, tenderness, nostalgia, peacefulness), Vitality (joyful activation and power) and Unease (tension and sadness). J. K. Vuoskoski and Eerola (2010, 2011) conducted principal component analysis on data from ratings of 16 film music excerpts and obtained a two-dimensional solution (the third factor only accounted for 6% of variance and was difficult to interpret, so it was discarded). The orthogonally rotated solution accounted for 89.9% of variance and the factors were named *Valence* and *Energy*. Pearce and Halpern (2015) collected another set of responses for the same 16 excerpt dataset, and factor analyzed these data. This resulted in retaining three factors with eigenvalues bigger than 1, accounting for 82% of variance. The factors were named Animacy (wonder, transcendence and power), Valence (sadness and tension and joyful activation (negative loading, the axis varies from positive to negative valence and not vice versa as usual)) and Arousal (tenderness, peacefulness, joyful activation (positively) and tension (negatively)).

We perform maximum likelihood exploratory factor analysis to compare our findings with the results described above.

Determining the number of factors to retain is usually a rather subjective procedure, based both on graphic and non-graphic tests, and interpretability criteria. Figure 2.7 shows the eigenvalues (computed from the correlation matrix) sorted from biggest to smallest (Scree plot). Keiser criterion suggests retaining components where eigenvalue exceeds one. There are three factors with eigenvalue bigger than 1 (eigenvalues are 3.39, 1.76 and 1). Examining the Scree plot for an elbow also suggests three factors. Figure 2.7 also illustrates three other diagnostic tests: Optimal Coordinate (OC), Acceleration Factor (AF) and Parallel Analysis (PA). OC method attempts to determine the location of the scree by measuring the gradients associated with eigenvalues and their preceding coordinates. AF determines the coordinate where the slope of the curve changes most abruptly. Parallel analysis of Monte Carlo simulations uses a comparison with random datasets of the same size. OC method indicates 3 factors, AF method — 2 factors, PA — 3 factors. We also use Very Simple Structure (VSS) test (Revelle & Rocklin, 1979), which compares the fit of a number of factor analyses with the loading matrix simplified by deleting all except the c greatest loadings per

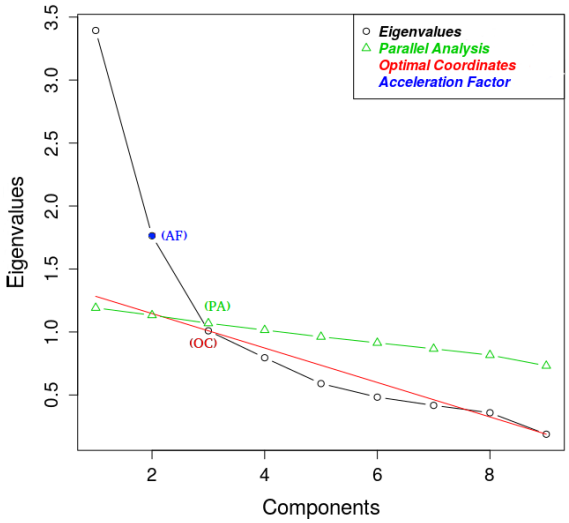


Figure 2.7: Scree plot that illustrates four of the test criteria — Scree test, AF, OC and PA.

item, where c is a measure of factor complexity. VSS and MAP criterion (which involves a complete PCA followed by the examination of a series of matrices of partial correlations) both suggest 2 factors. Four out of seven tests suggest retaining three factors.

In the end, we decided to retain three factors, which together explain 84.2% of variance in our data. We extracted and rotated them using orthogonal varimax rotation. Table 2.7 shows the factor loadings. The first factor (named **Arousal**) strongly correlates with calmness and tenderness, and negatively with power. The second factor (**Valence**) correlates with sadness and negatively with joyful activation and amazement. The third factor (**Tension**) correlates with nostalgia and negatively with tension. All the names of our factors are given by the direction of the strongest negative correlation of the factor. Three factors of a similar nature have been already discovered in musical emotion research before (Wedin, 1969, 1972b; Nielzen & Cesarec, 1969), making these findings highly plausible.

In the original study by Zentner et al. (2008) tension and sadness contribute to one factor **Unease** and are correlated with $r = 0.22$. In our case, sadness and tension are not correlated at all ($r = 0.03$). Factor **Sublimity** is similar to our factor **Tension**. But *Solemnity* (former transcendence) in our study is negatively correlated with *tenderness* and *nostalgia*, while it was positively correlated in (Zentner et al., 2008) ($r = 0.42$ and $r = 0.33$, respectively). See Table 2.6 for correlation values.

2.7 Influence of personal factors on induced emotion

Personal and situational factors can significantly affect the emotion induced by music in the listener (Thompson, Graham, & Russo, 2005; Dibben, 2004). In this section, we will examine the degree of this influence for various factors.

	Arousal	Valence	Tension
Amazement	−.15	−.40	.16
Solemnity	−.01	.18	−.51
Tenderness	.54	.27	.33
Nostalgia	.37	.49	.52
Calmness	.88	.21	−.09
Power	−.66	−.21	−.19
Joyful activation	−.37	−.81	.25
Tension	−.49	.07	−.54
Sadness	.05	.72	.07

Table 2.7: Factor loadings of the GEMS categories.

2.7.1 Influence of mood

Emotion induction by music might produce a different effect in the listener, depending on his mood (not necessarily so with perceived emotion). Cantor and Zillmann (1973) found that induced musical emotion is influenced by prior emotion induction. Schellenberg, Corrigall, Ladinig, and Huron (2012) found that listener's induced emotion is stronger when a listener experienced a contrasting emotion induction (through another piece of music) before listening to music. In (Dibben, 2004), participants' arousal was manipulated with physical exercise prior to listening to music. Their self-reported induced emotion changed, while perceived emotion did not differ between groups that did or did not exercise. Likewise, in (Zagrodski, 2013) no effect of previous mood was observed for recognizing perceived emotion from music.

We construct a linear mixed effects model with mood as fixed effect and musical excerpts as random effects for each of the emotions.

Linear mixed effects model allows to model both fixed effects (our explanatory variable) and random effects (the extra factor in our data that we want to account for). The responses for the same song are dependent, which violates the independence assumption necessary to use a linear model. Mixed model accounts for this by estimating different random intercepts per song. The model has a form:

$$y = X\beta + Z\theta + \varepsilon, \tag{2.5}$$

where y is a vector of observations (in our case, strength of an emotional category), β is a vector of fixed effects (mood), θ is a vector of random effects (ID of the song), ε is a random error, and X and Z are the design matrices related to y, β and θ.

To find which of the emotions were influenced by mood, we perform likelihood ratio test with ANOVA. This test compares two models: without the factor of interest and with the factor. We find significant differences for the categories sadness, tension, tenderness and amazement (Table 2.8). The clearest tendency is observed for sadness. The lower the participant's mood, the more often he or she selects sadness as an emotion induced by music. Participants who indicated that their mood was "very bad" selected sadness almost twice as often as the participants whose mood was "very good". A similar trend is observed for tension — the lower the mood, the more ten-

sion the music induces. An opposite trend is observed for amazement — the better the person feels, the more amazement is induced by music.

Emotion	Participant's mood					$\chi^2(4)$	p-value
	1	2	3	4	5		
Amazement	**14%**	**12%**	**15%**	**16%**	**18%**	**37.89**	$\mathbf{1.2 \cdot 10^{-7}}$
Solemnity	17%	21%	22%	22%	24%	0.90	0.92
Tenderness	**23%**	**19%**	**20%**	**23%**	**18%**	**9.69**	**0.04**
Nostalgia	25%	27%	26%	28%	26%	2.00	0.57
Calmness	56%	43%	40%	44%	44%	2.32	0.67
Power	20%	17%	19%	20%	21%	5.10	0.28
Joyful activation	23%	25%	29%	27%	29%	5.29	0.25
Tension	**21%**	**15%**	**14%**	**15%**	**16%**	**10.01**	**0.04**
Sadness	**28%**	**17%**	**14%**	**15%**	**15%**	**34.11**	$\mathbf{7.1 \cdot 10^{-7}}$

Table 2.8: Mood and frequency of selection of emotional category.

2.7.2 Influence of genre

Again, we used a linear mixed model in the same way as in the previous section to investigate the influence of genre on emotion selection. We found that all the emotions except Joyful Activation are induced with different frequency depending on genre. Figure 2.8 shows the boxplot of the distribution of emotion strength as computed by Formula 2.2 per emotion per genre.

2.7.3 Influence of age and gender

We did not find any influence of gender on emotion induction. However, we observed an effect of age for three emotional categories: for amazement ($\chi^2(1) = 17.52, p = 2.84 \cdot 10^{-5}$), for solemnity ($\chi^2(1) = 4.55, p = 0.03$) and the biggest effect for calmness ($\chi^2(1) = 42.75, p = 6.2 \cdot 10^{-11}$). Table 2.9 shows the trends — for every emotion, where an effect was observed, the emotion is induced less frequently as the age increases (the data was discretized into 10 year lapses).

2.7.4 Influence of musical preference

Liking and disliking the music appears to be very important for induced emotions, and is even sometimes considered to be a musical emotion per se. From table 2.10

Emotion	< 20	20 to 30	30 to 40	40 to 50	50+
Amazement	16.7%	14.2%	11.3%	14.4%	10%
Solemnity	26%	19.6%	22.8%	20.5%	19.1%
Calmness	36.9%	31.6%	27.3%	26.7%	26.2%

Table 2.9: Emotion induction frequency with age.

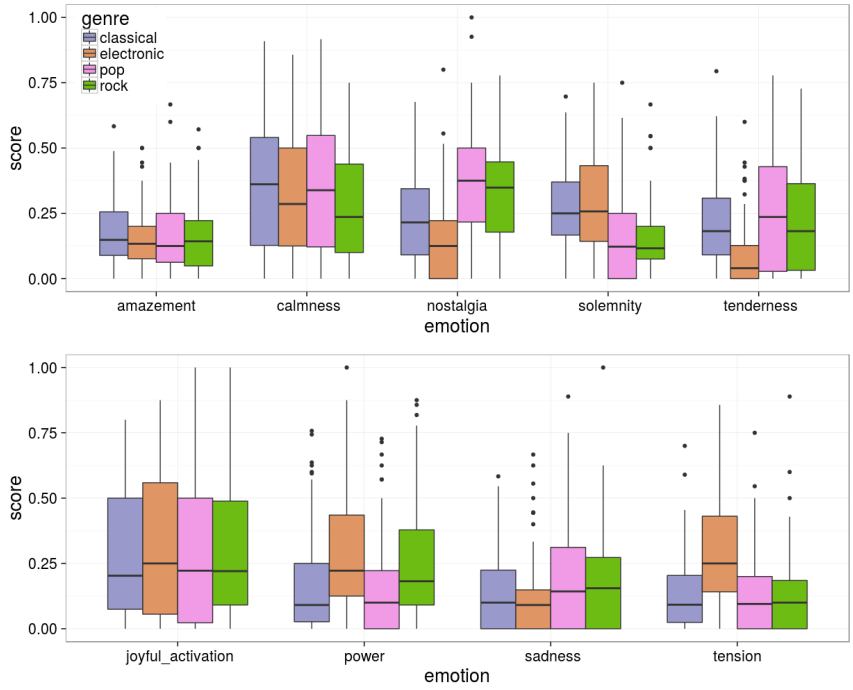


Figure 2.8: Boxplots of emotion distribution by genre.

Emotion	When music is liked	When music is disliked
Amazement	8%	2%
Solemnity	11%	6%
Tenderness	11%	5%
Nostalgia	12%	11%
Calmness	17%	12%
Power	10%	8%
Joyful activation	15%	8%
Tension	5%	27%
Sadness	6%	27%

Table 2.10: Category assignment by percentage of song listenings (n = 8358) depending on liking or disliking the music.

we can see that selection of emotional category is strongly dependent on whether participant liked or disliked the music, especially for such categories as amazement, joyful activation, tension and sadness. It is important to understand whether we can rely on the knowledge about the preferred genres to predict whether someone will enjoy the music.

Genre	Regular Listeners		Non-Listeners	
	Liked songs	Disliked songs	Liked songs	Disliked songs
Classical	60% *	4% *	48% *	12% *
Rock	40% *	24% *	30% *	35% *
Pop	39%	26%	30%	29%
Electronic	37%	25%	27%	30%

Table 2.11: Liking and disliking music by genre preference. Significant on a 5% level on a Chi-Squared test, if marked with asterisk.

We asked the players to indicate their genre preferences before the game (see a list of questions in Section 2.4.2). From Table 2.11 we see that in all cases people who report frequently listening to genre X tend to like songs in genre X more and dislike them less than those who do not prefer this musical genre. Though this difference exists, it is not as big as might be expected, and for pop and electronic music the differences between liking and disliking the music depending on genre preferences were not even statistically significant on a Chi-Squared test.

2.7.5 Effect of liking the music on the response consistency

For more than half of the listening sessions, participants reported whether they liked or disliked the music (the question was optional). There was a positive dependency between the consistency of the ratings (as measured by intraclass correlation coefficients) and liking the music. When the disliked listening sessions were excluded, the data showed more consistency (mean ICC (Intraclass Correlation) = 0.18 as compared to ICC = 0.16, significant on a t-test with p-value < 0.01).

Figure 2.9 shows a correlation between liking the music and response consistency. Such a correlation might mean that people can understand an emotion of the song better when they like the song, or it might mean that people like the song more when it is easier to understand its emotion.

2.8 Discussion

In this chapter we described a GWAP for music induced emotion annotation and analyzed the data collected using this GWAP. We were aiming at improving automatic music emotion recognition methods by creating a new sizeable and public dataset, providing further testing to the GEMS model, and studying the extra-musical factors that contribute to emotion induction.

The game that we presented is a serious game created for the purpose of data collection, but in contrast with the other GWAPs, it does not make players compete to guess the correct answer. Emotify rewards the players by giving them feedback on

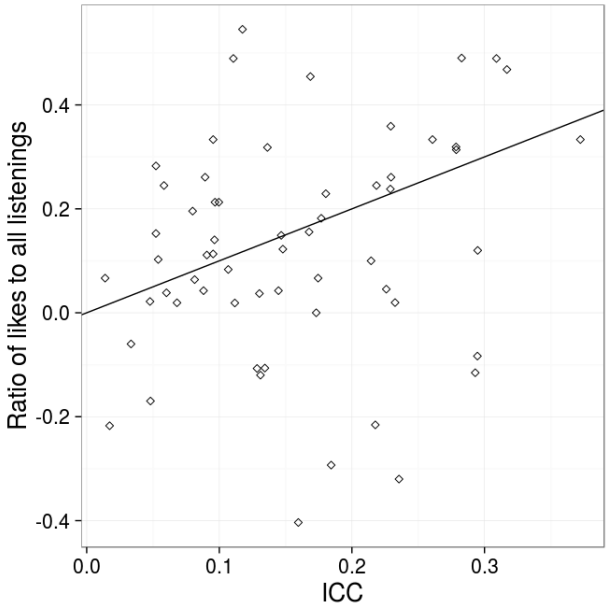


Figure 2.9: Scatterplot of song ICC vs. likes ratio. Pearson's $r = 0.26$.

their own input and comparing their answers to other players' input. Emotify uses Facebook as a platform. We faced some limitations, however, when using the Facebook application platform. The Emotify game tried to use invitations to involve new players into the game, but the acceptance rate of the invitations was very low: of the invitations sent by the players of Emotify, only 6% were accepted. Moreover, people were reluctant to use the Facebook version of the game. After running the pilot of the game on Facebook, we launched an independent website for hosting the game, and both the Facebook version and the independent website were advertised together. Having a choice, more than 90% of the players preferred to use the independent website.

We modified the GEMS model because it was found that participants find some of the categories confusing (J. K. Vuoskoski & Eerola, 2010; Torres-Eliard et al., 2011). Two of the modified categories (*amazement* and *solemnity*, previously *wonder* and *transcendence*) still resulted in a low agreement between participants (0.46 and 0.65 in terms of Cronbach's alpha, interpreted as Unacceptable and Acceptable, respectively), which might be caused by two issues. Firstly, low agreement might imply that these categories are inherently more subjective and depend on situational, cultural and other factors. Secondly, the feedback questionnaire showed that these two categories are less understood (*amazement* and *solemnity* were considered unclear by one third of the participants), which might be the second cause of low agreement. For the majority of the categories, such as *tenderness*, *joyful activation*, *power*, and *calmness*, the inter-rater agreement is high, and these categories also are comprehensible enough according to the feedback questionnaire.

When conducting factor analysis on our data, we found three factors — Arousal, Valence and Tension — similarly to (Wedin, 1969, 1972b; Nielzen & Cesarec, 1969). The structure of the factors we found was not similar to the original GEMS study by Zentner et al. (2008). We did not observe that *tension* and *sadness* jointly contribute to any of the factors, and *amazement* and *solemnity* were not correlated with the same factor with other emotions that contributed to factor **Sublimity** in (Zentner et al., 2008). The reason may be that our experiment was conducted in a different language, and the meaning of emotional words shifted after translation. Also, the music genres differed from the original study.

We did not find any significant differences in inter-rater agreement across genres, and therefore we conclude that GEMS is equally suitable for describing all the four studied genres.

In this chapter we also studied factors external to music. We found that the most important factor that should be taken into account when predicting induced emotion is liking or disliking the music. However, for our particular selection of broad music genres, the self-reported genre preferences failed to predict liking of the music accurately. In our study we did not intend to control participants' liking of the music using their self-reported genre preferences, but this finding might be important for designing further experiments. We also found that participants' induced emotions were affected by their mood (to a considerable extent) and age, but not by gender.

2.9 Conclusion

One of the open research questions that we addressed with this study was whether music can express and induce a complex fine-grained range of emotions, or it is only possible to find crude counterparts of verbally expressible emotions in music. On basis of our study we conclude that there indeed is enough variety and expressive power in music to convey and induce such emotions as tenderness, nostalgia or peacefulness in such a way, that they can be distinguished by participants with sufficient inter-rater agreement. We also concluded that the GEMS model can be successfully used by participants from various linguistic backgrounds, though there obviously exists a lack of understanding concerning categories *wonder* and *transcendence*. It is a direction for future research to find out what should be done with them.

Apart from this modification, we would also leave for future work to research how the GEMS model should be augmented with more emotional categories. Our study suggests that some of the nuances of emotional experience might be absent from the GEMS model (8% of our participants were not able to use GEMS to describe their induced emotions). Our findings agree with (Coutinho & Scherer, 2012) that feelings of boredom and interest must be added to the model, but also suggest that more semantic categories are lacking from it. Such semantic groups as impetus (call to action), humour and contentment were repeatedly named by the players of our game.

Another motivation for our study was collecting a dataset of music annotated with induced musical emotion which could be used as a ground-truth for MER research. The size of the dataset makes it possible to apply computational methods to explore the mechanisms underlying music emotional expressiveness, and to use these methods

for automatic music classification and retrieval. We will describe research on that in the next chapter.

We hope that this work will contribute to solving the problem of finding the most appropriate model of musical emotion. Though this problem is important both for research on music psychology and music industry, currently it is far from being solved.

CHAPTER 3

Computational modeling of induced musical emotion

3.1 Introduction

As was demonstrated in the previous chapter, induced emotion is highly subjective and depends on listener's perception, which is influenced by musical taste, mood and age of the listener. There is also a substantial agreement across the listeners, which is based on the shared understanding of the relationship between musical audio (musical cues) and emotion.

A recently proposed multi-level framework BRECVEMA, which we already mentioned in the previous chapter, explains emotion induction by music through a set of eight psychological mechanisms (Juslin, 2013). The mechanisms that are related only to musical audio (and not to a particular listener's memories and associations) are *brain stem reflex* (reaction to startle stimuli), *rhythmic entrainment* (locking up with the rhythm of music), *emotional contagion* (mimicking the perceived emotion of music) and *musical expectancy* (when a specific feature of the music violates, delays, or confirms the listener's expectations). The acoustic cues in music that trigger these mechanisms are loudness (through the brain stem reflex), tempo and rhythm stability (through the rhythmic entrainment), structure and repetition (through the musical expectancy), and all the numerous cues that influence perceived emotion (through the contagion mechanism). The latter cues have been extensively studied.

Gabrielsson and Lindström (2011) reviewed more than 100 empirical studies on a relationship between musical expressive mechanisms and perceived emotion, conducted between the end of the 19th century and the present time, and found that the most important cues are *tempo*, *mode*, *loudness* (dynamics), *timbre*, *harmony* (consonance/dissonance, chords, melodic and harmonic intervals), *melody*, *pitch* (pitch height and pitch range), *rhythm* and *articulation*.

The cues might interact, and the relationship between the cues and emotions might not be linear. Whether this is so, was studied in (Juslin & Lindström, 2010; Ilie &

Thompson, 2006; Eerola, Friberg, & Bresin, 2013). Juslin and Lindström (2010) manipulated eight cues (all of them are listed above in italics as the most important cues for emotion) to create musical pieces. Listeners rated emotions expressed in these pieces (basic emotions). A total of 77–92% of listeners' ratings could be predicted by a linear combination of cues, accounting for interactions only provided a small (4–7%) improvement. Ilie and Thompson (2006) manipulated three cues and found interactions between intensity and pitch height, and pitch height and tempo. Eerola et al. (2013) manipulated six cues and found that most of the variance in ratings was explained by linear influence of the cues, and no interactions were found. These findings have limited ecological validity, as they all were conducted on specially composed pieces which are bound to be much simpler than real music.

In this chapter we investigate the problem of audio features for automatic music emotion recognition and their adequacy. First, we conduct a perceptual experiment to find a set of musical cues which can describe emotional categories in the GEMS model through cognitive constructs. Then, we extract a comprehensive set of audio features, using several open-source toolboxes for audio signal processing, speech processing and MIR for feature extraction. We also develop new features that describe the harmonic content of music (interval features and chord features). We extract low-level spectral features related to timbre and energy with OpenSmile feature extration tool, and a more musically motivated feature set, containing high-level features related to mode, rhythm, and harmony, with MIRToolbox, PsySound, SonicAnnotator and Essentia. We compare the performance of extracted audio features and perceptual features.

As induced emotion is a highly subjective phenomenon, the performance of the model is confounded by the amount of agreement between listeners who have provided the ground-truth. The scenario when no other input (such as listener's mood and demographic details) from the listener is available is the most common application scenario for MER. We show that a good performance can still be achieved for induced emotion recognition with better high-level features, describing rhythmic and harmonic content of the music.

3.2 Background

In this section we first review the findings from musicological studies on acoustic cues important for emotion in music, then we review the audio features used for MER, and, lastly, we review approaches to computational modeling of musical emotion and the results achieved so far.

3.2.1 Musical structure and emotion

Gabrielsson and Juslin (2002) summarized the findings from 79 studies on the relationship between musical structure and expression. In Table 3.1 we list the emotions that are represented in the GEMS model and the cues associated with them. We will not cite the original studies, but we will indicate in parenthesis the number of studies in which the relationship was found. The references to the original studies can be found in (Gabrielsson & Juslin, 2002). Two of the emotions from the GEMS model are not found in studies and are not represented in the table: amazement and nostalgia. The

other emotions from the GEMS model appeared in the studies much less frequently than the basic emotions or Valence–Arousal dimensions, and the list of cues might be inconclusive.

In the Introduction we mentioned an adaptation of the Brunswick's lens model by Juslin (2000), which explains how emotion is decoded through a set of probabilistic redundant cues. In the Table 3.1 we can see how one cue applies to many emotions (e.g., slow tempo is associated with sadness, calmness, tenderness and solemnity), one emotion is interpreted through a set of many cues, sometimes even contradictory (according to different studies, solemnity can be associated with major or minor mode). Contradictory cues might mean that either different combinations of cues can create the same emotions, and across the combinations the cues can be contradictory, or, simply and more plausibly, there was no agreement on the interpretation of a certain emotion between the participants of different studies.

It must be noted that in most of the studies classical music or specially composed music was used, and it is unknown to which extent these results can be generalized to popular music.

3.2.2 Audio features for MER

Acoustic features extracted from the audio rarely have a straightforward counterpart in musicology. Due to complexity of human auditory perception and music cognition, and complexity of the audio signal itself, designing meaningful audio features requires a lot of work. Separate areas of MIR research are chord recognition, onset detection, tempo estimation, fundamental frequency detection, key detection and beat tracking, each developing or modifying audio processing pipeline for their purposes. The performance of MER methods depends on the progress in all of these areas. In Table 3.2 we list the main features that are used in MIR and explain their musical meaning. For the in-detail explanation of feature extraction pipeline and description of more audio features we refer to (Müller, 2015), (Y.-H. Yang & Chen, 2011a) and (Balen, 2016).

In this chapter we propose new interval and chord features, which is why we will review other chord and interval features below. Then, we will review the studies on the relationship between musical emotion and audio features.

Common mid-level audio features are chords. Chords are able to convey emotion even without musical context (Sollberge, Rebe, & Eckstein, 2003; Lahdelma & Eerola, 2016). Various chord-based statistical measures have been employed for different MIR tasks, such as music similarity or genre detection. Mckay and Fujinaga (2004) successfully applied chord and melody based features to genre recognition of symbolically represented music. Cheng, Yang, Lin, Liao, and Chen (2008) used chord features (longest common chord sequence and histogram statistics on chords) to find similar songs and to estimate their emotion (in terms of valence) based on chord similarity. Schuller, Dorfner, and Rigoll (2010) used chord statistics for MER. In Schuller's histograms, the duration of chords was not taken into account, which we found important and account for in this paper. B. Yang and Lugger (2010) improved the accuracy of emotion detection from speech using interval-based features. These features were calculated using circular autocorrelation of the pitch histogram (calculated from the estimated pitch contour of an utterance) on the logarithmic semitone scale, and occurrence of different two-pitch intervals was measured. The reason why interval features

Emotion	Musical cues
Solemnity	Slow tempo (7), major mode (1), minor mode (1), consonance (4), high loudness (1), moderate or high loudness (2), few loudness changes (1), low pitch (3), narrow melodic range (1), regular rhythm (2), firm rhythm (1), legato articulation (2), sharp envelope (1), small timing variation (1)
Tenderness	slow tempo (4), moderate tempo (1), soft loudness (5), rising intonation (1), little loudness variation (1), legato articulation (3), round envelope (2), soft timbre (2), large timing variation (2), softened contrasts between long and short notes (2)
Calmness	slow tempo (4), consonance (3), soft loudness (3), high pitch (1), low pitch (1), narrow pitch range (2), regular rhythm (1), flowing rhythm (1), legato articulation (1), small vibrato extent (2), low formal complexity (1)
Potency	fast tempo (1), high loudness (1), high pitch (1), rising pitch contour (1), many harmonics (1), round envelope (1)
Joy	fast tempo (25), moderate tempo variation (2), major mode (8), consonance (4), high loudness (12), small loudness variation (3), high pitch (5), large pitch range (3), ascending melody (2), regular rhythm (2), varied rhythm (1), fluent rhythm (2), staccato articulation (8), large articulation variation (3), bright timbre (2), sharp envelope (3), moderate timing variation (2), sharpened contrasts between long and short notes (2), low formal complexity and average dynamism (2)
Tension	dissonance (2), high sound level (2), ascending melody (2), increased note density (2), harmonic complexity (1), rhythmic complexity (2), lack of melody (1), various formal properties (2)
Sadness	slow tempo (24), minor mode (8), dissonance (4), soft loudness (12), small loudness variation (1), moderate loudness variation (1), low pitch (8), narrow pitch range (2), descending melody (2), firm rhythm (2), legato (7), little articulation variation (3), soft timbre (4), round envelope (4), large timing variation (2), softened contrasts between long and short notes (2), slow vibrato (2), high formal complexity combined with low dynamism (2)

Table 3.1: Musical structure and emotion, from (Gabrielsson & Juslin, 2002).

worked for speech is probably related to consonance/dissonance rather than to harmony, as in the case of music.

Intervals (both melodic and harmonic) were shown to be important for musical emotion in musicological studies (Gabrielsson & Lindström, 2011). Costa, Fine, and Bitti (2004) studied melodic intervals and found that activity was expressed by melodies with a greater occurrence of minor seconds, tritones, and intervals larger than the octave; potency (vigour, power) was associated with more frequent occurrences of unisons and octaves; pleasant melodies had greater occurrence of perfect fourths and minor sevenths and no or less tritones.

Though a wealth of audio features is available, most of them are far away from actually explaining the cognitive mechanisms of music processing, which are essential for decoding emotion (Wiggins, 2009; Aucouturier & Bigand, 2012). Currently, low-level spectral audio features are a staple of MIR. Song, Dixon, and Pearce (2012) evaluated 55 audio features related to dynamics, rhythm, spectrum and harmony, and found that low level spectral features give the best performance in a four class classification task. The most plausible interpretation, in our opinion, is that high level features are not up to the task yet. Eerola (2011) showed that models built using extracted audio features are not robust across genre (when tested on a different genre they perform much worse). This might mean that current audio features are very sensitive to particularities of sound production, or it might mean that music in different genres expresses emotions in a different way. Trochidis, Delbe, and Bigand (2011) found that for contemporary western classical music valence was best explained by pulse clarity, articulation and brightness, and arousal was best explained by periodicity amplitude of flatness and entropy of the magnitude of the highest peak in the chromagram. Evaluating these sort of findings is always a challenging task, because often particular transformations of the spectrum used as audio features do not have any perceptual counterpart and hence it is difficult to find a meaningful motivation why these particular transformations should work well or not. These concerns about the evaluation of MIR methods are extensively raised by Bob Sturm (Sturm, 2013, 2014).

3.2.3 MER approaches

Automatic MER can be formulated both as a regression and a classification problem, depending on the underlying emotional model, making it possible to apply almost any machine learning algorithm to MER. There are also different ways to represent the ambiguity (the subjectiveness of human emotional experience), such as multi-label classification or representing emotion of a song as a Gaussian distribution (J.-C. Wang, Yang, Wang, & Jeng, 2015). Below we will describe some of the approaches and report the accuracy achieved using these approaches. It must be noted that different classification, regression, multi-label classification approaches use different evaluation metrics, which makes it difficult to compare the achieved accuracy. And in any case we would like to stress that it is anyway inappropriate to compare the performance on different datasets, because of different amount of inherent ambiguity in these datasets and different emotional models used (and number of classes to distinguish). In order to compare different feature sets and different learning algorithms, a benchmark is necessary. We will develop and describe a benchmark for MER in Chapter 4.

Musical concept	Features
Timbre	MFCCs, Δ MFCCs, ΔΔ MFCCs, spectral features (centroid, shape, spread, skewness, kurtosis, contrast, flatness), tristimulus, brightness, 95% rolloff, zero crossing rate, octave-based spectral contrast, Daubechies wavelets coefficient histogram (DWCH), auditory modulation features, inharmonicity, roughness, dissonance, odd to even harmonic ratio
Loudness	RMS energy, specific loudness on Bark critical bands
Harmony	chromagram, chromagram peak, key, mode, key clarity, harmonic change, chords, HCDF
Pitch height	spectral centroid, low energy
Rhythm	tempo (bpm), beat histograms, rhythm regularity (autocorrelation on onset detection curve), rhythm strength, onset rate
Articulation	attack slope, attack time

Table 3.2: Acoustic features and musical concepts.

The first research on automatic music emotion recognition dates back to year 2003 (Li & Ogihara, 2003), and the problem was formulated as a 13-class multi-label classification. In 2008, Y.-H. Yang et al. (2008) first used Valence–Arousal model in a MER system. SVR was trained on low-level acoustic features (spectral contrast, DWCH and other low-level features from Marsyas and PsySound), achieving performance of 0.76 for arousal and 0.53 for valence (in terms of Pearson's r here and further). Laurier, Lartillot, Eerola, and Toiviainen (2009) modeled five dimensions (basic emotions) with a set of timbral, rhythmic and tonal features, using SVR. The performance varied from 0.59 to 0.69. In (Guan, Chen, & Yang, 2012), pleasure, arousal and dominance were modeled with AdaBoost.RM using features extracted from audio, MIDI and lyrics. A feature set consisting only of audio features performed worse than multimodal features (audio + MIDI + lyrics) (0.4 for valence, 0.72 for arousal and 0.62 for dominance). J. C. Wang, Lee, Chin, Chen, and Hsieh (2015) used a hierarchical Dirichlet process mixture model, whose components can be shared between models of different emotions, for multi-label classification. In (Xue, Xue, & Su, 2015), two modalities were combined (lyrics and audio) and fused with Hough forest to create a deep network, achieving 60% classification accuracy on a four class classification task. In (Hu & Downie, 2010b) it was shown that lyrics outperformed audio features on almost all of the 18 mood categories (except calmness). However, it is not clear whether the reason for a good performance of lyrics features is that lyrics are more important than audio (and listeners base their emotional judgements mostly on lyrics) or that audio features were just not up to the task. This is possible, because only low-level spectral audio features (MFCC, spectral flux and roll-off, etc.) were used. However, this still proves that state-of-the-art lyrics features are better than audio features. And whatever the precedence between lyrics and sound, lyrics undoubtedly are responsible for a large part of emotional meaning in songs.

In most of the MER studies it was noted that modeling the Valence dimension (or emotions that are related to valence, such as happiness or sadness) is always more difficult (Y.-H. Yang et al., 2008; Laurier, Lartillot, et al., 2009; Guan et al., 2012). This

might happen because energy (Arousal) can generally be successfully predicted from loudness and roughness of the timbre. These features are easier to extract. Valence, on the other hand, is related to harmonic and melodic content and expectation, and it is impossible to predict valence-related emotions using only low-level spectral features. In this chapter we will work towards improving the situation by developing new interval-based features and chord statistics to strengthen the harmony-related features.

3.3 Data preprocessing

The dataset was described in detail in Chapter 2. We clean the data and remove some annotations and some songs. We are not interested in modelling irritation and boredom from listening to non-preferred music, so we remove the ratings of the listeners who reported that they disliked the music from the annotations. We also remove the songs which were most confusing for the listeners, and they failed to reach a certain level of agreement on them (33 songs). To do that we compute Fleiss's kappa, which is a statistical measure designed to estimate agreement, when the answers are binary or categorical. For each song, a number of people (players of the game) rate some items (emotions) on a binary scale (emotion is either present or not). Fleiss's kappa is calculated as follows:

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}, \tag{3.1}$$

where $1 - \bar{P}_e$ gives the degree of agreement that is attainable above chance, $\bar{P} - \bar{P}_e$ gives the actual degree of agreement above chance. If the agreement is at the chance level, then $\kappa \leqslant 0$. $\bar{P}$ is the mean of all P_i, which indicate the degree of agreement per subject. In our case, subjects are emotional categories and there are nine subjects. Agreement per subject is calculated as follows:

$$P_i = \frac{1}{n(n-1)} \sum_{j=1}^{k} n_{ij}(n_{ij}s - 1), \tag{3.2}$$

where n is the total number of ratings per subject, k is the number of categories (in our case, there are two categories). The categories are indexed $j = 1, ...k$. The subjects (emotions) are indexed $i = 1, ...N$. Therefore, n_{ij} is the number of raters who assigned the i-th emotion to the j-th category (categories are "emotion selected" and "emotion is not selected"). $\bar{P}_e$ is calculated as follows:

$$\bar{P}_e = \sum_{j=1}^{k} (\frac{1}{Nn} \sum_{i=1}^{N} n_{ij})^2. \tag{3.3}$$

We compute kappa on all the annotations for every musical excerpt, and discard the songs with negative kappa (this indicates that the answers have chance-level consistency).

We retain the remaining 367 songs for analysis (44 100 Hz, 128 kbps, converted to mono). Each excerpt is 1 minute long, except for 4 classical pieces which were shorter than 1 minute.

3.4 Perceptual acoustic cues and GEMS emotions

In this section we will describe an experiment we conducted to explain the GEMS categories through a set of musically motivated cues. We selected part of the dataset (60 pieces out of 367) for this purpose, consisting of 15 pieces from each genre. We chose pieces from the **subset A** (see section 2.4.1), because these pieces have more emotion annotations.

3.4.1 Procedure

Three musicians (26–61 years, each with over 10 years of formal musical training) annotated 60 pieces from the dataset with 10 cues, on a scale from 1 to 10. The list of cues was adapted from the study of Wedin (1972b): tempo (slow—fast), articulation (staccato—legato), mode (minor—major), intensity (pp—ff), tonalness (atonal—tonal), pitch (bass—treble), melody (unmelodious—melodious), rhythmic clarity (vague—firm), harmonic complexity (simple—complex). We added rhythmic complexity (simple—complex) to this list, and eliminated style (date of composition) and type (serious—popular) from it. Later, the harmonic complexity cue was discarded because of the lack of agreement among annotators.

Cue	Amz	Slm	Tnd	Nst	Clm	Pwr	Jfl	Tns	Sdn
Tempo	.50	−.44	−.48	−.47	−.64	.39	.76		−.45
Articulation	−.37	.39	.56	−.57	.48	−.35	−.70	−.36	.51
Rhythmic compl.	*.27						*.27		−.38
Mode		−.45	.30			*−.27	.24	−.36	−.23
Intensity	*.27		−.48	−.30	−.50	.51	.41		−.24
Tonalness			*.29	*.28				−.47	
Pitch			.44	*.27	.36	−.47		−.44	
Melody			.54	.50		−.43		−.66	*.27
Rhythmic clarity		−.34					.31		

Table 3.3: Correlations between musicological cues and emotional categories. Amz — amazement, Slm — solemnity, Tnd — tenderness, Nst — nostalgia, Clm — calmness, Pwr — power, Jfl — joyful activation, Tns — tension, Sdn — sadness.

3.4.2 Analysis

Interactions between cues

We tested for interactions between perceptual cues using linear models with added interactions. The presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different values of the other predictor variable. Interactions have implications for the way that the categories should be modeled using the cues.

We found few interactions, mostly with mode. Mode interacted with tempo (for joyful activation and calmness), intensity (for joyful activation), pitch height (for ten-

derness) and rhythmic clarity (for power and nostalgia). Also, there was an interaction between intensity and tonalness (for joyful activation and nostalgia).

Explaining categories through cues

Table 3.3 shows the correlations (Spearman's ρ) between acoustic cues and GEMS emotional categories. We used a non-parametric test, because distribution of emotional categories is positively skewed (emotion was more often not present than present). All the correlations are significant with p-value < 0.01, except for the ones marked with asterisk, which are significant with p-value < 0.05. The values that are indicated in grey failed to reach statistical significance, but are still listed, because they conform with the trends previously found in the literature and are likely to reach statistical significance on a bigger sample.

The most important cues that influence almost every emotion in GEMS, are tempo, articulation, intensity (loudness) and mode. For each of the emotions, the most prominent cues are listed below, the new findings are highlighted in bold font, the finding that contradicts previous ones in italic font (there is only one such finding):

- Amazement — **fast tempo** and **staccato articulation.**
- Solemnity — slow tempo, legato articulation, minor mode, *flowing rhythm.*
- Tenderness — slow tempo, legato articulation, **major mode**, soft loudness, **high pitch**, **melodious.**
- Nostalgia — **slow tempo**, **staccato articulation**, **soft loudness**, **tonal**, **high pitch**, **melodious.**
- Calmness — slow tempo, legato articulation, soft intensity, high pitch.
- Power — fast tempo, **staccato articulation**, **minor mode**, high intensity, **low pitch**, **unmelodious.**
- Joyful activation — fast tempo, staccato articulation, **high rhythmic complexity**, major mode, high intensity, high rhythmic clarity.
- Tension — staccato articulation, **minor mode**, atonal, **low pitch**, unmelodious.
- Sadness — slow tempo, legato articulation, **low rhythmic complexity**, minor mode, soft loudness, **melodious.**

As we described in the previous chapter, several GEMS categories were strongly correlated (*tenderness* and *nostalgia*: $r = 0.51$, *tenderness* and *calmness*: $r = 0.53$, *power* and *joyful activation*: $r = 0.41$). All of these have, however, musical characteristics that allow to differentiate them.

Both *nostalgia* and *tenderness* correlate with slow tempo, but *tenderness* is also correlated with major mode, and legato articulation (as opposed to staccato for nostalgia). *Calmness* is characterized by slow tempo, legato articulation and smaller intensity, similarly to *tenderness*. But *tenderness* has a correlation with melodiousness and major mode as well. Both *power* and *joyful activation* are correlated with fast tempo, and intensity, but *power* is correlated with minor mode and *joyful activation* with major mode.

3.5 Audio feature extraction

In the previous section we examined the perceptual features annotated by human experts. In this section we will move to automatically extracted features. First, we will extract audio features using a number of different toolboxes (some of the features are implemented in multiple toolboxes and we will not extract them several times). Then, we will develop some new features.

3.5.1 Feature sets

We extract a comprehensive set of both high-level and low-level audio features. Before extracting any harmony-related features (such as mode, key, key clarity, chords), we perform harmonic-percussive source separation using the method of Fitzgerald (Fitzgerald, 2010) and then extract features from the harmonic part of the spectrum. The method uses median filtering across successive frames to suppress percussive components (and uses median filtering across frequency bins to suppress harmonic components).

1. **OpenSmile**[1]: 260 low-level spectral features related to timbre and energy (mean and standard deviation of 65 low-level acoustic descriptors, and their first-order derivatives) (Eyben, Weninger, Gross, & Schuller, 2013). OpenSmile combines features from Speech Processing and MIR and has shown good performance on cross-domain emotion recognition (Weninger, Eyben, Schuller, Mortillaro, & Scherer, 2013). The 260 feature set that we extract consists of features selected for their good performance on music emotion recognition in the MediaEval 2014 Emotion in Music benchmark. Features developed to describe voice were computed using 60 ms frames and Gaussian windows ($\sigma = 0.4$). Other features were calculated using 25 ms frames and Hamming windows. In both cases, overlapping windows were used with a step size of 10 ms.

2. **MIRToolbox**[2]: 43 features from MIRToolbox (Lartillot & Toiviainen, 2007) (low level spectral features and high-level features: HCDF, mode, inharmonicity, key clarity, tempo). MIRToolbox was conceived as a tool for investigating the relationship between emotions and features in music. Low level features are extracted with 50 ms frames and Hamming window, and high-level features related to harmony are extracted with 200 ms frames.

3. **PsySound**[3]: 4 features related to loudness from PsySound (using the loudness model of Chalupper and Fastl), and perceptual roughness.

4. **Essentia**[4]: Essentia's functionality and audio features are similar to the ones from the MIRToolbox. We only extract onset rate with Essentia.

[1] opensmile.sourceforge.net
[2] jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/mirtoolbox
[3] psysound.wikidot.com
[4] essentia.upf.edu

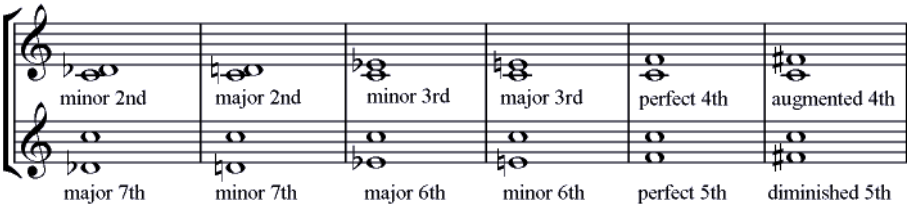


Figure 3.1: Intervals and their inversions.

3.5.2 Harmonic features

In this section we describe the statistics that we suggest to compute on the intervals and chords in each song. Before extracting these features, we also apply a harmonic-percussive source separation step, as described in the previous section.

Interval Features

We segment the audio, using local peaks in the harmonic change detection function (HCDF) (Harte & Sandler, 2006). HCDF describes tonal centroid fluctuations. The segments that we obtain are mostly smaller than 1 second and reflect single notes, chords or intervals. From these segments we compute the spectrum between 100 Hz and 6400 Hz (corresponding to the integer number of octaves, in order to avoid the range of the spectrum covering particular pitches more than others). From this spectrum we compute chromagrams. Then, from the 12 values (one octave) that we obtain, we select the two peaks with the highest energy and compute the interval between them. For each type of interval, we compute its combined duration, weighted by its loudness (expressed by energy of the bins). Because the chromagrams only span one octave, it is impossible to distinguish between, for instance, a fifth spanning **A2**–**E3** and a fourth spanning **E3**–**A3**. Therefore we combine intervals and their inversions. Figure 3.1 illustrates the concept (each bar corresponds to the musical representation of a feature that we obtain). As there are 6 distinct intervals with inversions, we obtain 6 features.

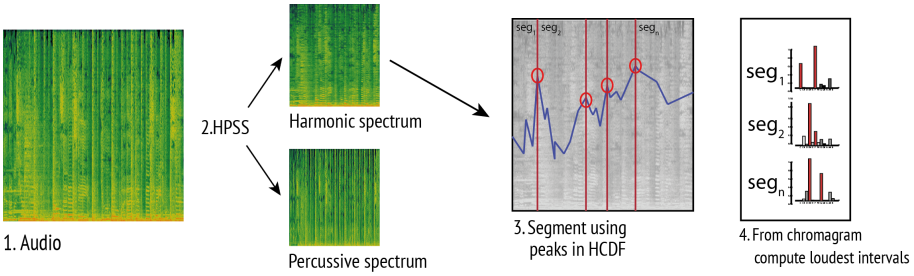


Figure 3.2: Feature extraction pipeline for interval features.

Figure 3.2 shows the feature extraction pipeline. We expect that augmented fourths and fifths (tritone) could reflect tension, and perfect fourths and fifths should have opposite meaning. The proportion of minor thirds and major sixths, as opposed to

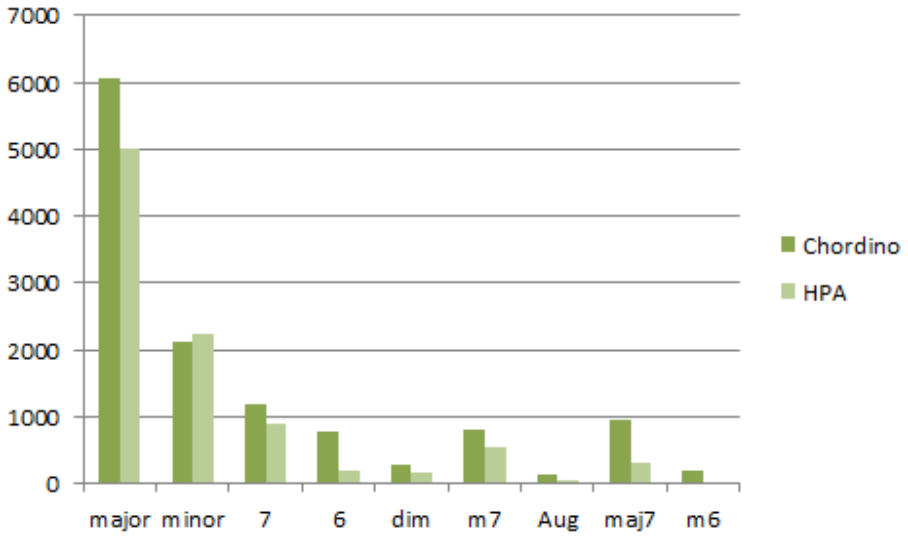


Figure 3.3: A histogram of chonord distribution for Emotify dataset.

proportion of major thirds and minor sixths, could reflect the modality. The interval-inversion pairs containing seconds are, hypothetically, rather unrestful.

Chord Features

To extract chord statistics, we used 2 chord extraction tools, HPA[5] (Harmonic Progression Analyzer) and Chordino[6] plugins for Sonic Annotator[7]. HPA provides 8 types of chords: major, minor, seventh, major and minor seventh, diminished, sixth and augmented. Chordino, in addition to these eight types, also provides minor sixth and slash chords (chords for which bass note is different from the tonic, and might as well not belong to the chord). The chords are annotated with their onsets and offsets.

After comparison we discarded the chords from HPA, because the chords from Chordino could explain our data better. We computed the proportion of each type of chord in each song, obtaining nine new features. The slash chords (the chords where the base note is not the root of the chord, and can sometimes belong to the triad and sometimes not) were merged with their base chord (e.g., Am/E chord is counted as a minor chord). The distribution of chords was disparate, with major chords being in majority (see Figure 3.3).

Weighting the chords by their duration was an important step, which improved the performance of chord histograms and has not been suggested before.

[5] patterns.enm.bris.ac.uk/hpa-software-package
[6] isophonics.net/nnls-chroma
[7] isophonics.net/SonicAnnotator

3.6 Evaluation

In this section we evaluate the perceptual features and the automatically extracted features on emotion recognition. We have several goals. Firstly, modeling induced emotion without taking into account any personal information about the listeners. Secondly, comparing what can be achieved with automatically extracted features and perceptual features.

3.6.1 Learning Algorithm

The focus of this chapter is not on choosing an algorithm for our regression task, but on evaluating audio features. We tried the state-of-the-art algorithms which have already shown good performance on MER: Support Vector Regression (SVR) and Gaussian Processes Regression (GPR). GPR and SVR had comparable performance and we chose SVR.

SVR brings training examples into higher-dimensional space and maps them so that distinct examples are divided by a clear gap that is as wide as possible. SVR is also called a large-margin classifier (we use the regression version). We trained SVR with three different kernels: linear, radial basis function and polynomial kernels. The SVR with linear kernel was performing much worse than the kernels that can learn non-linear functions. In the following evaluation we will use the LIBSVM implementation[8]. The best performance was achieved using the RBF kernel, which is defined as follows:

$$k(x_i, x_j) = \exp\left(-\gamma \left\|x_i - x_j\right\|^2\right), \tag{3.4}$$

where γ is a parameter given to SVR. All the parameters, C (error cost), epsilon (slack of the loss function) and γ, are optimized with grid-search on the training data for each feature set (but not for each emotion). To select an optimal set of features, we use recursive feature elimination (RFE). RFE assigns weights to features based on output from a model, and removes attributes until performance is no longer improved.

3.6.2 Evaluation

We evaluate the performances of the feature sets by incrementally adding more high level features. We start with low-level OpenSmile features, then add more low-level and some high-level features from MIRToolbox, PsySound and Essentia, then the chord features and interval features that we designed. We evaluate using 10-fold cross-validation (in a proportion 75% training set and 25% test set), splitting the dataset by artist (there are 140 distinct artists per 400 songs). If a song from artist A appears in the training set, there will be no songs from this artist in the test set. Splitting by artist is particularly important for features from openSMILE, which tend to be able to discriminate timbral differences very well and overfit to the dataset. Table 3.4 shows evaluation results. With each addition of more features the performance is a little bit improved. The accuracy of the models differs greatly per category, while all the feature sets demonstrate the same pattern of success and failure (for instance, perform badly

[8] csie.ntu.edu.tw/~cjlin/libsvm

Feature set	OpenSmile		OS, MT, PS, E		OS, MT, PS, E, HF	
Emotion	r	RMSE	r	RMSE	r	RMSE
Amazement	.21 ± .01	.10 ± .01	.28 ± .01	.10 ± .01	.32 ± .00	.09 ± .00
Solemnity	.43 ± .01	.14 ± .01	.44 ± .01	.14 ± .01	.44 ± .01	.13 ± .01
Tenderness	.52 ± .01	.15 ± .02	.47 ± .01	.15 ± .01	.51 ± .01	.15 ± .01
Nostalgia	.46 ± .01	.17 ± .01	.44 ± .01	.16 ± .01	.52 ± .01	.15 ± .00
Calmness	.51 ± .01	.20 ± .01	.54 ± .01	.19 ± .01	.58 ± .01	.18 ± .01
Power	.53 ± .01	.16 ± .01	.55 ± .01	.17 ± .01	.56 ± .01	.16 ± .01
Joyful activation	.56 ± .01	.19 ± .01	.62 ± .01	.20 ± .01	.60 ± .01	.19 ± .01
Tension	.52 ± .01	.16 ± .01	.48 ± .01	.17 ± .01	.55 ± .01	.15 ± .01
Sadness	.28 ± .02	.16 ± .01	.31 ± .02	.16 ± .01	.43 ± .01	.14 ± .01

Table 3.4: Evaluation of 3 feature sets on the data. Pearson's r and RMSE with their standard deviations (across cross-validation rounds) are shown. OS — OpenSmile, MT — MIRToolbox, PS — PsySound, E — Essentia, HF — harmonic features.

on *amazement* and well on *joyful activation*). This reflects the fact that these two categories are very different in their consistency, as we showed in Chapter 2. Figure 3.4 illustrates the performance of the systems (r) for each of the categories along with Cronbach's alpha (which measures agreement), and shows that the performance metric and consistency metric are highly correlated. The low agreement between listeners results in conflicting cues, which limit model performance.

In table 3.4 we observe a counter-intuitive trend — the emotional categories that have better performance according to the r metric (larger r), such as joyful activation or power, have worse performance in terms of the Root Mean Squared Error (RMSE) metric (larger RMSE). This trend can be explained by the particular property of the ground truth. The range of values for each of the emotional categories is in theory between 0 and 1, but in practice it is different per emotional category. For instance, for joyful activation the range is indeed between 0 and 1, but for amazement it is only between 0 and 0.52. The players of Emotify game could unanimously agree on the presence of joyful activation for some songs, but it never happened for amazement. Different emotions have different inherent ambiguity and resulted in different consistency of the annotations. The emotional categories with better consistency could be predicted with more accuracy (higher r), but they also have wider range of values (and therefore higher RMSE). Because of these complications, r is generally a more accurate metric for this dataset, but for the sake of completeness we display both of them in the table. We only show correlation coefficient on the Figures 3.5 and 3.4.

In general, the accuracy of our system is lower than that achieved for perceived emotion by others (Y.-H. Yang et al., 2008; Laurier, Lartillot, et al., 2009; Guan et al., 2012). This might be caused by the fact that all the categories contain both arousal and valence components (as we have seen in factor analysis in previous chapter), and also induced emotion annotations are less consistent. In (Laurier, Lartillot, et al., 2009), *tenderness* was predicted with $r = 0.67$, as compared to $r = 0.52$ in our case. For *power* and *joyful activation*, the predictions from the best systems demonstrated 0.56

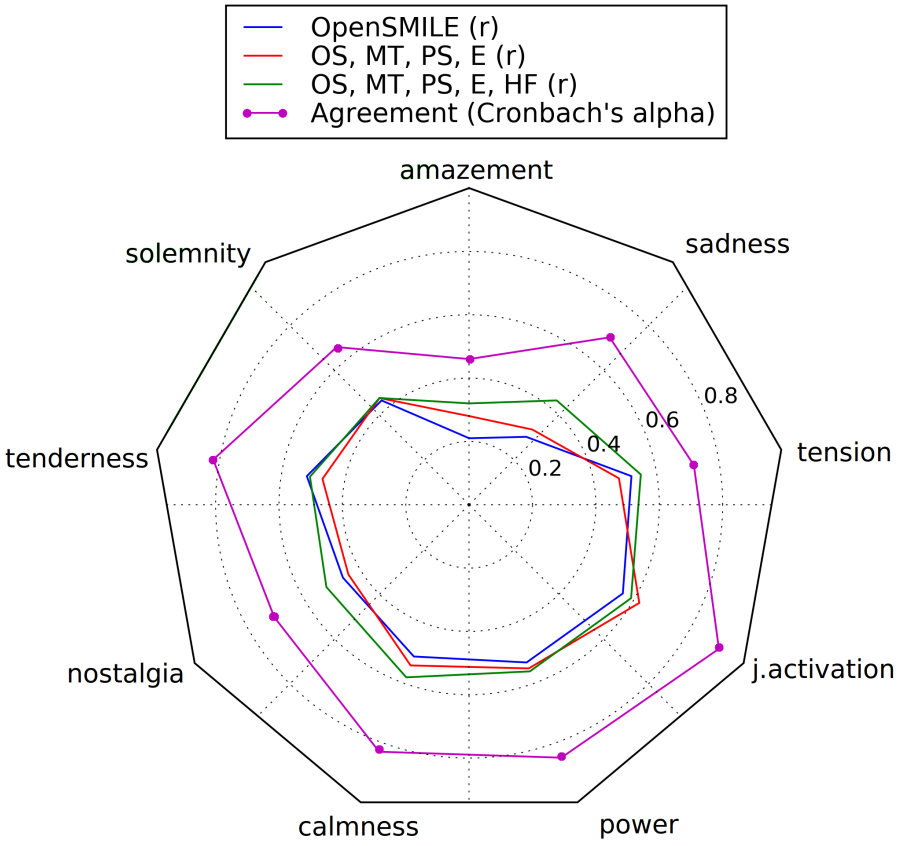


Figure 3.4: Comparison of systems' performance (r), and Cronbach's alpha per category.

and 0.62 correlation with the ground truth, while in (Guan et al., 2012; Y.-H. Yang et al., 2008) it was 0.72 and 0.76 for arousal.

The performance of all the three feature sets is comparable, though the system with most high-level features performs slightly better in most of the categories. Adding harmonic features improves average performance from 0.45 to 0.50, and performance of the best feature set decreases to 0.35 when answers from people who disliked the music are not discarded. As we were interested in evaluating the new features, here we show a list of features that were considered important by RFE. The 10 most important features are displayed for each emotion, the new features are highlighted in bold font:

- Amazement — spectral centroid and brightness, loudness, 95% rolloff, first MFCC, zerocross, entropy, **major chords**, **major seconds and inversions.**
- Solemnity — spectral features (centroid, brightness, skewness, spread), flatness, 95% rolloff, first MFCC, loudness, hcdf, **major chords.**
- Tenderness — spectral features (centroid, brightness, kurtosis, spread, skewness, flux), 95% rolloff, hcdf, **major chords**, **minor chords.**
- Nostalgia — RMS, spectral features (centroid, spread, skewness, flux), first MFCC, zerocross, 95% rolloff, **major chords**, HCDF.
- Calmness — RMS (and standard deviation), spectral features (brightness, skewness, spread), entropy, 95% rolloff, **major chords**, HCDF, **minor seconds and inversions.**
- Power — spectral features (centroid, skewness, brightness, kurtosis), 95% rolloff, RMS, roughness, loudness, **fourths and inversions**, HCDF.
- Joyful activation — spectral features (spread, skewness, centroid, brightness, kurtosis), loudness, onset rate, HCDF, **seventh chords**, 95% rolloff.
- Tension — spectral features (centroid, brightness, kurtosis, spread), 95% rolloff, first MFCC, key clarity, **major chords**, **minor chords**, entropy.
- Sadness — spectral features (centroid, spread, kurtosis, skewness), flatness, onset rate, **minor chords**, loudness, **minor seconds and inversions**, HCDF.

Low level spectral features perform very well and are selected as very important. High level features mode and tempo did not appear on the list for any emotion, though the other (but relying on more low-level computations) features related to mode and tempo are there: minor and major chords, onset ratio.

3.6.3 Perceptual cues and computationally extracted features

As we saw in the previous section, the most important audio features (as selected by RFE) are low level spectral features, describing the timbre. We do not yet know whether this is so because in popular music timbral qualities are the most important, or because high-level feature extraction does not perform well enough yet. To find out, in this section we will compare the performance of the audio features with the performance of perceptual cues, manually annotated by musicians. None of the perceptual

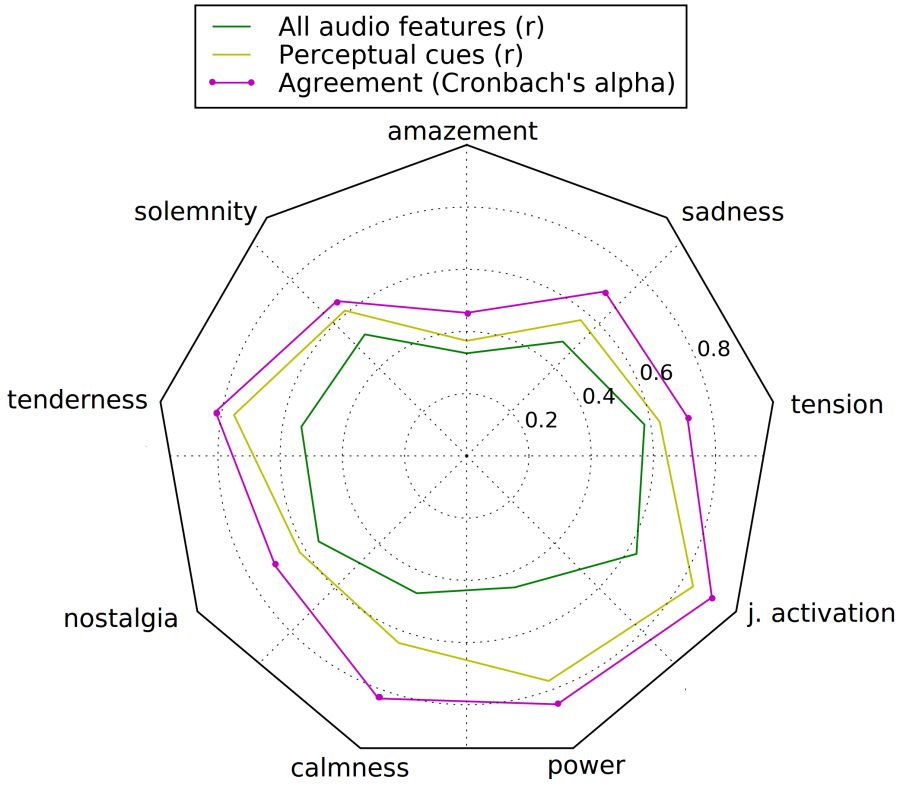


Figure 3.5: Comparison of performance of audio features and perceptual manually annotated cues, and Cronbach's alpha per category.

cues describes the timbre. We use SVR as described in the previous section on a subset of 60 songs (subset A) with 10-fold cross-validation. Table 3.5 shows the results, and figure 3.5 visualizes our main metric (correlation coefficient) along with Cronbach's alpha.

Feature set	OS, MT, PS, E, HF		Perceptual cues	
Emotion	r	RMSE	r	RMSE
Amazement	.33 ± .01	.08 ± .01	.37 ± .01	.08 ± .01
Solemnity	.51 ± .00	.09 ± .02	.61 ± .01	.10 ± .01
Tenderness	.54 ± .02	.12 ± .02	.76 ± .02	.11 ± .02
Nostalgia	.55 ± .02	.13 ± .01	.62 ± .02	.13 ± .02
Calmness	.47 ± .02	.18 ± .03	.64 ± .02	.15 ± .02
Power	.45 ± .03	.16 ± .03	.77 ± .02	.13 ± .02
Joyful activation	.63 ± .02	.19 ± .02	.84 ± .02	.14 ± .02
Tension	.58 ± .02	.13 ± .04	.63 ± .02	.12 ± .03
Sadness	.48 ± .01	.12 ± .01	.57 ± .01	.16 ± .01

Table 3.5: Evaluation of 2 feature sets on the data. Pearson's r and RMSE with their standard deviations (across cross-validation rounds) are shown.

From this evaluation we see that the perceptual cues can predict the emotional categories much better than audio features. On average $r = 0.65$ for perceptual cues, and $r = 0.50$ for audio features. It is likely that high-level feature extraction is too noisy and inaccurate as of yet, and this is why the low-level features are still demonstrating better performance.

3.7 Conclusion

In this chapter we analyzed the performance of different audio features on the prediction of induced musical emotion. We extract as wide variety of audio features as possible, and suggest new features that describe the harmonic content of the music. However, even with this very comprehensive feature set, we could not reach the performance that could be reached with only 9 attributes manually annotated by musicians. These attributes performed very well, despite being identified to be relevant for classical music, and applied to mostly popular (rock, pop and electronic) music.

The advancement of MER methods depends on the advancement of all the other areas of MER, such as source separation, chord recognition, onset detection, beat tracking, audio-to-score transcription. Given that the ground truth is consistent enough, we predict that the 'glass ceiling' has not been reached for music emotion recognition yet. However, predicting complex concepts, such as rhythmic complexity, tonalness or melodiousness, from acoustic signal, is utterly non-trivial and will require models of human cognition (Wiggins, 2009).

In this chapter we analyzed which musicological concepts are important for induced emotion. These findings could be used as a guideline for future feature development.

Part II

Benchmarking MEVD algorithms

CHAPTER 4

Emotion in Music benchmark at MediaEval Evaluation Campaign

This chapter contains a description of the benchmark design and an analysis of the outcomes of the benchmark for Music Emotion Variation Detection (MEVD) algorithms that was jointly organized by Mohammad Soleymani, Yi-Hsuan Yang and myself in the years 2013–2015. The winning algorithms and feature sets over the years are analyzed, and the design, evaluation metrics and data that we used are described. We also release the largest available dataset of continuous annotations of music with emotion, and suggest some transformation and data cleaning procedures which improve the quality of these data.

4.1 Introduction

In the last decade, many new MER methods have been proposed (Y.-H. Yang & Chen, 2012). However, it is difficult to compare their performance because methodological differences in data representation result in a choice of different evaluation metrics. Figure 4.1 shows 14 different data annotation and representation choices in a form of a labyrinth (each choice is a way to go through the labyrinth). In addition to these choices, a wide variety of categorical and dimensional models are used, such as basic emotions (Laurier, Lartillot, et al., 2009), valence and arousal model (Y.-H. Yang et al., 2008; Eerola, 2014; Barthet, Fazekas, & Sandler, 2012), GEMS (J. K. Vuoskoski & Eerola, 2011; Aljanaki, Wiering, & Veltkamp, 2014), or custom mood clusters (Hu et al., 2008; Schubert, 2003). Despite differences in data representation, most of the methods are essentially solving the same problem of mapping audio features (or lyrics and metadata-based features) to the emotional annotations. A specific learning algorithm cannot always be adapted to other representations (though many algorithms, such as SVM or different types of neural networks, are versatile), but audio features

are most certainly transferable. A benchmark can therefore enable a comparison of different methods and feature sets, by fixing the data representation choice and releasing a dataset.

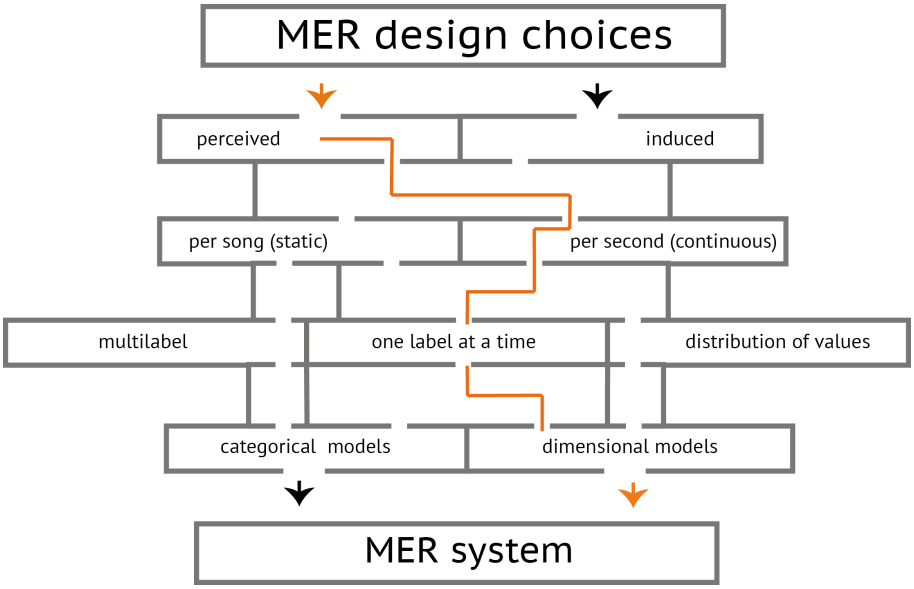


Figure 4.1: A labyrinth of data representation choices for a MER algorithm. The choices that we made for the benchmark are connected with a line.

Another problem of MER is that due to audio copyright restrictions, the datasets used in various studies are seldom made public and reused in other studies. Annotations are often obtained by crawling the tags from social music websites, such as last.fm or allmusic.com. In this case, the audio is usually copyrighted and cannot be redistributed by the researchers. The music that is distributed for free under a license such as Creative Commons, usually is less well-known and has fewer tags, and therefore needs to be annotated. Annotation is a huge burden, because with such a subjective task many annotations are needed for every item ro bring out the consensus.

Therefore, a benchmark for MER is needed that would:

1. Fix a data representation choice, an emotional model and an evaluation metric.
2. Release a public dataset and keep a separate test set for evaluation. Some of the benchmarks do not release any data, leaving it to the participants to find the training data. This approach would not work well for MER, because the internal subjectivity in the data can vary a lot across the different datasets, and could compromise an otherwise worthwhile approach.
3. Design the evaluation procedure that would explain the meaningful differences between approaches.

In our benchmark we fixed the data representation to continuous emotion annotation. A fundamental property of music is that it unfolds over time. An emotion ex-

pressed in the song may also change over time, though it is always possible to reduce this variety to a single value. The online music websites, such as moodfuse.com, musicovery.com, allmusic.com, usually represent songs in a mood space by a single label, which is always an approximation of the emotional content of the song. In the design of the benchmark we recognize the time-dependent nature of music by setting out to predict the emotion of the music dynamically (per-second), i.e., the main purpose of the benchmark is to compare music emotion variation detection algorithms.

We also collected and released new data every year. The MediaEval Database for Emotional Analysis in Music[1] (DEAM) is the combination of the datasets developed in three years (with data transformation and cleaning procedures applied to them), in addition to the raw annotations. DEAM database contains 1802 songs (58 full-length songs and 1744 excerpts of 45 seconds) from a variety of Western popular music genres (rock, pop, electronic, country, jazz etc.). Part of the data was annotated in the lab and part using Amazon Mechanical Turk crowdsourcing platform[2]. We made a very traditional and safe choice for the emotional model: we used the Valence and Arousal model (Russell, 1980; Thayer, 1989) to annotate the data.

The evaluation procedure that we designed allows to compare the two crucial components of a machine learning system (algorithms and the feature sets) separately.

The benchmark was first organized in 2013, and over three years of activity 21 teams participated in the task. In this chapter we will systematically evaluate the feature sets and the algorithms.

4.2 Background

In this section we will describe the annotation interfaces for MEVD and the datasets collected using these interfaces. Then, we will review the only other benchmark for MER and its design choices. Lastly, we will describe the algorithms suggested for MEVD, which we will benchmark and compare later.

4.2.1 Datasets and annotation interfaces for MEVD

Since the late 1980s, time-varying responses to music have been measured using the Continuous Response Digital Interface (Robinson, 1988; Gregory, 1989). Usually, only one dimension (such as tension, musical intensity or emotionality) was measured. Schubert proposed to use a two-dimensional interface (Valence–Arousal plane) to annotate music with emotion continuously (Schubert, 1996). This approach was adopted by MER researchers as well.

Speck, Schmidt, Morton, and Kim (2011) used an interface very similar to the one suggested by Schubert to create a game with a purpose MoodSwings that was already described in Chapter 2. The interface of the game is shown on Figure 4.2. The data from MoodSwings was released publicly[3]. The dataset comprises 240 segments of US pop songs (each 15-second long) with per-second V–A annotations, collected through MTurk. After an automatic verification step that removes unreliable annotations, each clip in this dataset is annotated by 7 to 23 subjects.

[1] cvml.unige.ch/databases/DEAM/
[2] mturk.com
[3] music.ece.drexel.edu/research/emotion/moodswingsturk

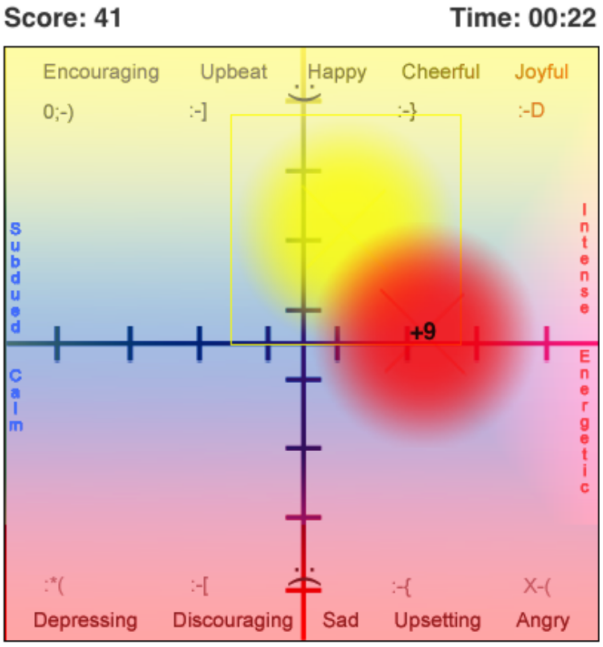


Figure 4.2: MoodSwings game interface.

A discrete interface to collect continuous measurements with a discrete categorical model, was also suggested by Schubert, Ferguson, Farrar, Taylor, and McPherson (2012). The interface is shown on Figure 4.3.

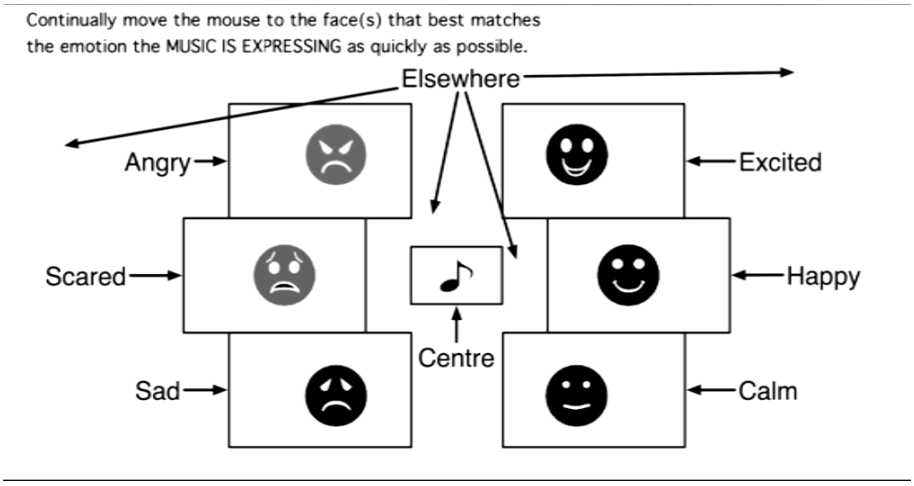


Figure 4.3: Emotion face clock. Black boxes, arrows and labels were not visible to the participant.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Rowdy	Amiable	Literate	Witty	Volatile
Rousing	Good natured	Wistful	Humorous	Fiery
Confident	Sweet	Bittersweet	Whimsical	Visceral
Boisterous	Fun	Autumnal	Wry	Aggressive
Passionate	Rollicking	Brooding	Campy	Tense/anxious
	Cheerful	Poignant	Quirky	Intense
			Silly	

Table 4.1: MIREX mood clusters.

The experiments involving continuous music annotations are numerous, but only MoodSwings game resulted in publicly available data. In Section 2.2.1 we also describe the public datasets with static (per song) music emotion annotations.

4.2.2 MIREX benchmark

The only other benchmark that exists for MER methods is the audio mood classification (AMC) task, organized by the annual Music Information Retrieval Evaluation eXchange[4] (MIREX) (Hu et al., 2008). In this task, the train set consists of 600 audio files which are provided to the participants of the task. The test set is not disclosed. Since 2013, another set of 1438 segments of 30 seconds clipped from Korean pop songs has been used in MIREX as well.

The benchmark uses five discrete emotion clusters instead of the more widely accepted dimensional or categorical models of emotion. The clusters are derived from cluster analysis of online tags from All Music Guide. The five clusters are shown in Table 4.1. AMC has been criticized for using an emotional model that is not based on psychological research. It was noted that there exists semantic or acoustic overlap between clusters (Laurier & Herrera, 2007). Probably, due to this overlap and the inconsistency that it creates in the data, the performance in MIREX mood recognition has not improved in the last 5 years, indicating a possible glass ceiling. We have already talked about this problem in the Introduction.

What is more, MIREX dataset only applies a singular static rating per audio clip, which belies the time-varying nature of music.

4.2.3 MEVD methods

The first study that models (using linear regression) musical emotion (arousal and valence) unfolding over time with musical features (loudness, tempo, melodic contour, texture, and spectral centroid) was conducted by Schubert (2004). The model could explain from 33% to 73% of variation in emotion. Korhonen, Clausi, and Jernigan (2006) suggested a method to model musical emotion as a function of musical features using system identification techniques. Korhonen et al. used the low-level

[4] music-ir.org/mirex/wiki

spectral features extracted using Marsyas software[5], and perceptual features extracted with PsySound software (Cabrera, 1999). The system reached a performance of 21.9 for valence and 78.4 for arousal in terms of R^2. Schmidt and Kim (2010) used Kalman filtering to predict per-second changes in the distribution of emotion over time on 15 second music excerpts. Schmidt and Kim (2011) suggested to apply a new method — Conditional Random Fields — to model continuous emotion with a resolution of 11×11 in valence–arousal space. A very small feature set was used — MFCCs, spectral contrast and timbre — and the system reached performance of 0.173 in terms of Earth Mover's Distance (between the true 11×11 2D histogram of Arousal–Valence values and predicted one). Panda and Paiva (2011) used Support Vector Machines and features extracted with Marsyas and MIRToolbox to track music over quadrants of Valence–Arousal space. Imbrasaite, Baltrušaitis, and Robinson (2013) combined Continuous Conditional Random Fields with a relative representation of features. Later, Imbrasaite, Baltrušaitis, and Robinson (2014) showed that using Continuous Conditional Neural Fields offers improvement over the previous approach. In (J.-C. Wang, Yang, Wang, & Jeng, 2012), the ambiguity of emotion was represented through a Gaussian distribution, and tracking of emotion over time was implemented using a mapping between music emotion space and low-level acoustic feature space through a set of latent feature classes. Markov, Iwata, and Matsui (2013) used Gaussian Processes for MEVD. The bidirectional Long Short-Term Memory Recurrent Neural Networks were first applied to continuous emotion recognition not in the domain of music, but in the domain of multi-modal human emotion prediction (from facial expression, shoulder gesture, and audio cues) (Nicolaou, Gunes, & Pantic, 2011).

It is also possible to detect changes in emotion by segmenting a song using a sliding window and applying the static MER method to each of the windows. This approach was suggested by Y.-H. Yang, Liu, and Chen (2006). A sliding window of ten seconds with a 1/3 overlap was used to segment a music piece, and a fuzzy k-NN classifier was trained to detect the emotion of the segments. This method would give a distorted result when a sliding window has an emotional boundary in it.

Most of the algorithms mentioned in this section were employed in the benchmark: Support Vector Regression, linear regression, Kalman filtering, Gaussian Processes, Conditional Random Fields, Continuous Conditional Neural Fields and Long Short-Term Memory Recurrent Neural Networks.

4.3 Music database

Our dataset consists of royalty-free (Creative Commons license enables us to redistribute the content) music from several sources: `freemusicarchive.org` (FMA), `jamendo.com`, and the medleyDB dataset (Bittner et al., 2014). There are 1744 clips of 45 seconds from FMA and 58 full length songs, half of which come from medleyDB and another half from Jamendo.

FMA contains music in a multitude of genres: *rock, pop, soul, blues, electronic, classical, hip-hop, international, experimental, folk, jazz, country* and *pop*. We tried to balance the genre distribution, but it was not always possible because songs are annotated with multiple genres and the more common genres appear more often. The

[5] `marsyas.info`

music from the MedleyDB dataset in addition had music in *world* and *rap* genres, and the music from Jamendo also had *reggae* music. For the 2014 and 2015 datasets, we manually checked the music and excluded the files with bad recording quality or those containing speech or noise instead of music. For each artist, we selected no more than 5 songs to be included in the dataset. From medleyDB and Jamendo we selected full-length songs, which had emotional variation in them. In order to detect emotional variation, we used a simple SVR-based MEVD algorithm for automatic filtering of emotionally heterogenous songs, and then manually made the final selection (three annotators voted for inclusion of the shortlisted songs).

4.4 Annotations

Getting high quality reliable annotations with sufficient agreement and consistency is a crucial step for a highly subjective task. To collect annotations, we have turned to crowdsourcing using Amazon Mechanical Turk (MTurk),[6] which was successfully used by others to label large libraries (Speck et al., 2011; Chen et al., 2015). We developed a procedure to filter out poor quality workers, following current state-of-the-art crowdsourcing approaches (Soleymani & Larson, 2010). The workers passed a test to demonstrate a thorough understanding of the task, and an ability to produce good quality work. The test contained several automatically scored multiple choice questions, and several free-form questions and assignments, which were evaluated manually if the automatically scored part was passed correctly. In years 2013 and 2014, each excerpt was annotated by a minimum of 10 workers. In 2015, each song was annotated by five workers, three of which were recruited among the most successful workers from previous years, and two were working in the lab. The dynamic annotations were collected using a web-interface on a scale from −10 to 10, where the Mechanical Turk workers could dynamically annotate the songs on valence and arousal dimensions separately while the song was being played. The static annotations were made on the nine-point scale on valence and arousal for the whole 45 seconds excerpts after the dynamic annotations. Figure 4.4 shows the interface used for annotation. The annotator had to keep the mouse in a box with the slider, otherwise the annotation paused. Three songs were presented one by one in one batch. It was possible to rest between doing the annotation, because we provided a generous time to the annotators to finish every batch. Additional questions about the song were presented after completing the main part. The extra questions are listed in Table 4.2.

In 2015 we also introduced a preliminary listening round. This is the round which looks exactly like the normal annotation round, but the measurements are not recorded. The annotator can familiarize himself or herself with the music and prepare to react when needed. We hoped that listening to a song once before starting the annotation would help to reduce the reaction time.

In addition to the audio features, we also provide meta-data covering the genre labels obtained from FMA, medleyDB and Jamendo, and, if available, folksonomy tags crawled from last.fm.

[6] mturk.com

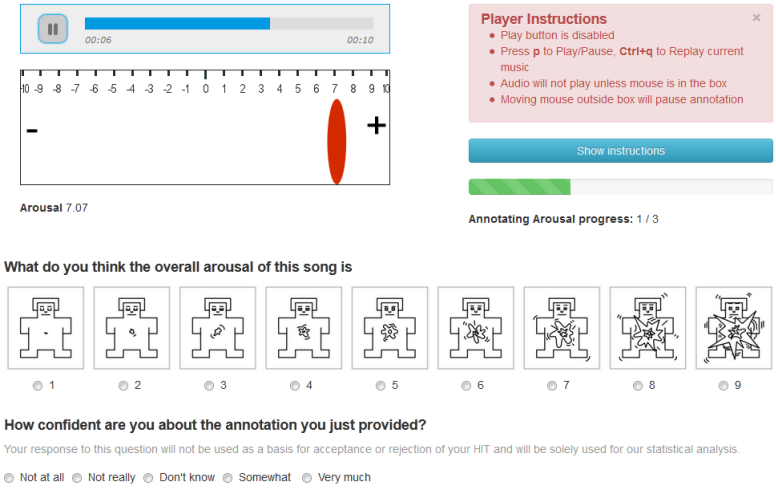


Figure 4.4: Annotation interface.

Year	Number of songs	Source	Extra data
2013	1000 (744 unique)	MTurk	Time of the day, mood
2014	1000	MTurk	Confidence in rating, familiarity of music, liking of music, free emotion label, Big Five personality, preferred genre
2015	58	MTurk/Lab	Liking, maximum and minimum arousal/valence value, free emotion label

Table 4.2: The data overview.

4.4.1 Annotation consistency

We resample the annotations from every annotator to the same rate (2 Hz). There are between 60 to 600 samples per song, depending on the length of a song and whether it is an excerpt or a full song.

To compare the consistency across annotators, we use two measures: we compute Cronbach's α on the sequences of annotations for each of the songs for all the annotators, and a coefficient of determination of a Generalized Additive Model that generalizes song's annotations across annotators.

We normalize the annotations for each song as follows:

$$a^*_{j,i} = a_{j,i} + \left(\overline{A}_j - \overline{A}\right), \tag{4.1}$$

where $a_{j,i}$ is an annotation by annotator j at timestamp i, $\overline{A_j}$ is the mean of the annotations by annotator j, and $\overline{A}$ is a mean of all annotations for this song by all annotators (global mean). This transformation shifts the annotations to the same range (they are very often off-place if the annotator started from the wrong point, which is difficult to correct in real time, especially for a short excerpt). See Figure 5.1 for an example of such a shift. This problem is also discussed in the next chapter.

Cronbach's α is used to estimate the degree to which a set of items measures a single unidimensional latent construct. This measure should theoretically range between 0 and 1, but in practice it can be negative when inter-item correlations are negative. There is no lower bound on negative values of this measure. Only positive values are informative and accurately report the degree of agreement. Therefore, we clip the negative tail by assigning the value of 0. Table 4.3 shows the averaged Cronbach's α for each year's annotations. To test whether annotation consistency improved with a change of experimental design, we will compare the three groups. The groups' sample sizes and variances are different, therefore we will use a non-parametric test based on ranks. The Kruskal–Wallis test (one way ANOVA on ranks) shows that there are significant differences between groups for arousal ($\chi^2(2) = 81.24$, p-value $= 2.2 \times 10^{-16}$) and the Dunnett–Tukey–Kramer test shows that the differences are significant between all three years on a 1% significance level. For valence, the differences exist ($\chi^2(2) = 57.91$, p-value $= 2.6 \times 10^{-13}$), but only in 2015 annotations are significantly different from the other groups.

The Cronbach's α test has some deficiencies, such as being sensitive to the number of items on the test (a greater number of items in the test can artificially inflate the value of alpha). In the year 2015 dataset, the songs were much longer. We will conduct an additional consistency test with GAMs.

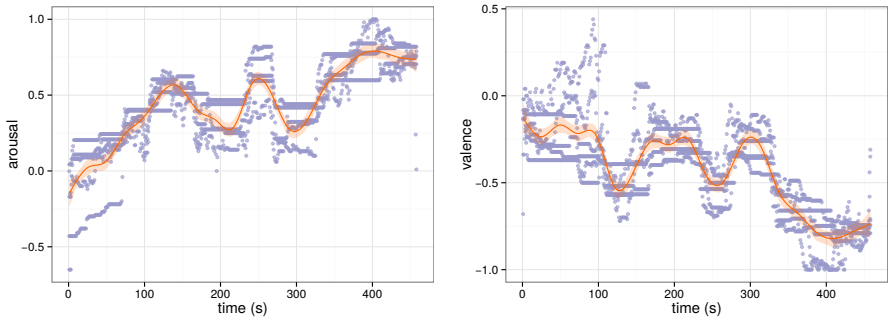
A generalized additive model is a generalized (i.e., allowing non-normal error distributions of the response variable) linear model with a linear predictor involving a sum of smooth functions of covariates. The model is defined as follows:

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \ldots + f_n(x_n), \tag{4.2}$$

where g is a link function (a function defining a relationship between the linear predictor and the mean of the dependent variable); $\mu = E(Y)$, where Y is a dependent variable; and $f_i(x_i)$ are non-parametric smooth functions, estimated, e.g., via scatterplot smoothing techniques, or can also be parametric functions or factors.

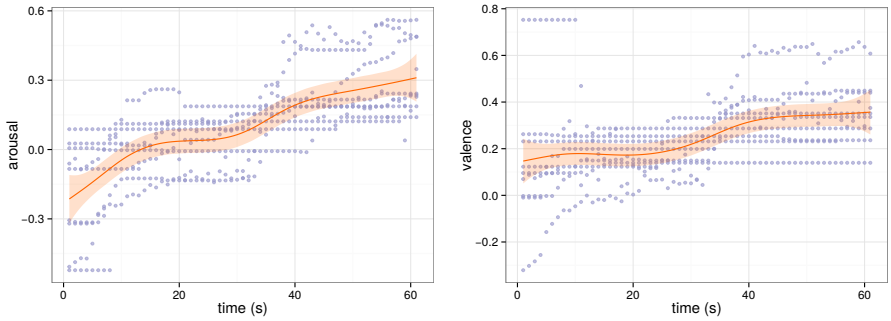
GAMs are very suitable for modeling continuous annotations of emotion, because these annotations are usually non-linear in nature and don't have abrupt changes, which makes it possible to model them using smooth functions. McKeown and Sneddon (2014) describe how GAMs and their mixed model extension can be used to model continuous emotion annotations and make inferences concerning linear differences between groups. In this paper we will only use GAMs to assess the effect size of shared perceived emotion. We will do that by building a model for each of the songs, and calculating the coefficient of determination (R^2) of the model.

There is only one smooth component in the model — time. We use penalized cubic regression splines with basis dimension of 20 and identity link function. The results are shown in table 4.3. Figure 4.5 shows scatterplots of annotations and fitted GAMS for 2 songs.



(a) Clip 2033, 2015, arousal. $R^2 = 0.72$

(b) Clip 2033, 2015, valence. $R^2 = 0.76$



(c) Clip 1008, 2014, arousal. $R^2 = 0.51$

(d) Clip 1008, 2014, valence. $R^2 = 0.23$

Figure 4.5: Fitted GAMs for arousal and valence annotations (annotations shown as a scatterplot).

There are significant differences between groups for arousal according to Kruskal–Wallis test ($\chi^2(2) = 121.03$, p-value $= 2.2 \times 10^{-16}$) and Dunnett–Tukey–Kramer test shows that the differences are significant between year 2015 and other groups on a 1% significance level. For valence, the outcome is the same: differences exist ($\chi^2(2) = 134.37$, p-value $= 2.2 \times 10^{-16}$), and only year's 2015 annotations are significantly different from the other groups.

According to both consistency measures, in 2015 we could achieve better consistency, which can be attributed to employing lab workers, choosing complete songs over excerpts and introducing preliminary listening.

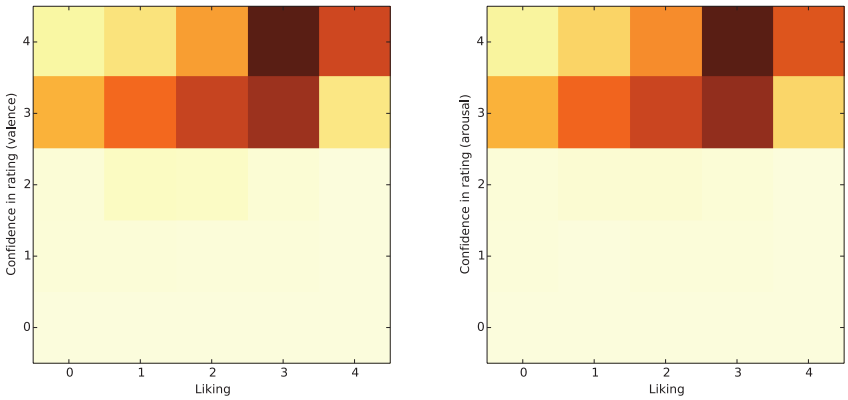
Despite all attempts, the consistency of the annotations is worse than what could be achieved for static (per song) annotations. In Chapter 2, we achieved a Cronbach's α of more than 0.7 for most emotional categories. The possible reasons for this will be discussed in the next chapter.

4.4.2 Influence of music familiarity, liking and other factors on the annotations

The Creative Commons music that we selected was largely unfamiliar to the participants (only in 1% of the listening sessions the participant reported having heard the

Year	2013	2014	2015
Total length	9 h 18 min	12 h 30 min	3 h 46 min
Cronbach's α for arousal	.27 ± .28	.31 ± .30	.66 ± .26
GAM's R^2 for arousal	.13 ± .10	.14 ± .11	.44 ± .19
Cronbach's α for valence	.27 ± .28	.20 ± .24	.51 ± .35
GAM's R^2 for valence	.13 ± .10	.10 ± .08	.37 ± .21

Table 4.3: Annotation consistency. Cronbach's α and GAM's coefficient of determination (mean and standard deviation) per year.



(a) Confidence in ratings of valence – liking. Spearman's $\rho = 0.37$, p-value $= 2.2 \cdot 10^{-16}$

(b) Confidence in ratings of arousal – liking. Spearman's $\rho = 0.29$, p-value $= 2.2 \cdot 10^{-16}$

Figure 4.6: Liking of the music and confidence in rating. The darker the color, the more measurements fall inside the square on a heatmap.

piece before). There was not enough data to derive any patterns regarding the familiarity of the music.

We found that liking influenced self evaluation of the confidence in rating. Figure 4.6 shows the 2D histogram for self-reported confidence in rating and liking the music. The confidence in rating is on average very high (the workers never reported being completely uncertain), which is, probably, caused by the fact that the data was collected from paid workers who did not want to be suspected of incompetence. Liking the music influenced perceived self-reported confidence. A similar effect was found in (Aljanaki, Wiering, & Veltkamp, 2016), when there was a positive dependency between liking the music and annotation consistency. We could not find any effect of averaged music liking on actual (and not self-reported) rating consistency.

4.4.3 Convergence of annotations

It is a known issue that the annotators need some initial orientation time (IOT), before their continuous annotations become meaningful and consistent. In (Schubert, 2013), median IOT was found to be 8 seconds for valence and 12 seconds for arousal. Also, afterglow effects — large outliers in the spread of scores just after the end of a piece — were identified. In (Bachorik et al., 2009), participants required on average 8.31 seconds to begin giving emotional judgements on music on a two-dimensional plane. The length of the delay was influenced by familiarity, genre and tempo of music.

To measure the IOT of the annotators in the beginning of the song, we calculate the average Krippendorff's α for every sample of the corresponding second for the whole dataset of year 2015. The Krippendorff's α is calculated as follows:

$$\alpha = 1 - \frac{D_o}{D_e}, \tag{4.3}$$

where D_o is the observed disagreement, and D_e is the disagreement expected by chance. We calculate D_o and D_e using the following formulas.

$$D_o = \frac{1}{n} \sum_{c \in R} \sum_{k \in R} (c-k)^2 \sum_{u \in U} m_u \frac{n_{cku}}{P(m_u, 2)}, \tag{4.4}$$

where n is the total number of elements (annotations by a single rater in our case), R is the set of acceptable ratings, m_u is the number of items in a unit (annotations by all raters of the same timestamp), n_{cku} number of (c,k) pairs in unit u, U represents items in a unit, and P is the permutation function.

$$D_e = \frac{1}{P(n,2)} \sum_{c \in R} \sum_{k \in R} (c-k)^2 P_{ck}, \tag{4.5}$$

where P_{ck} is the number of ways the pair (c,k) can be made.

The songs in the dataset had different length. Figure 4.7 shows that the annotations start to converge around the 13$^{\text{th}}$ second. A similar result was obtained in the 2013$^{\text{th}}$ annotations. So, despite preliminary listening stage, the IOT did not diminish.

We remove the first 15 seconds of the annotation from the benchmark data.

4.5 Benchmark history and design

The benchmark for music emotion recognition algorithms was organized in the years 2013–2015 as part of the MediaEval Benchmarking Initiative for Multimedia Evaluation[7]. MediaEval is a community-driven endeavour dedicated to evaluating algorithms for multimedia access and retrieval that has been organized annually since year 2008 (as VideoClef, in the years 2008 and 2009). The list of tasks offered at the benchmark is renewed every year based on interest and feedback from the multimedia retrieval community. Alongside the Emotion in Music task, 10–11 other tasks related to speech, music, image and video processing were held at MediaEval in the years 2013–2015.

[7] multimediaeval.org

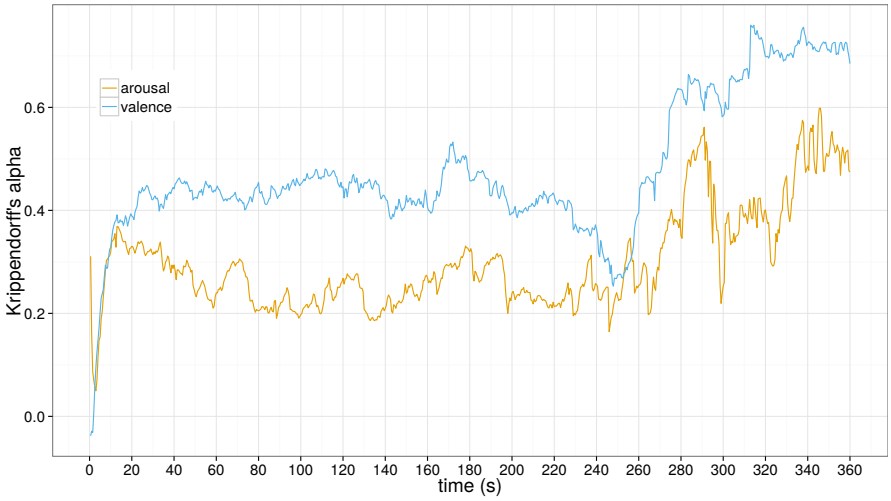


Figure 4.7: Krippendorff's α of dynamic annotations in the year 2015, averaged over all dynamic samples.

We followed the MediaEval benchmarking tradition by developing a separate development and evaluation set for each year. Every year, the data collected the previous year was publicly released.

4.5.1 Task description 2013

In 2013, the task was first proposed and organized by Mohammad Soleymani, Yi-Hsuan Yang and Erik Schmidt (Soleymani, Caro, Schmidt, & Yang, 2013). The task consisted of two subtasks: dynamic and static emotion characterization. In dynamic emotion characterization, the participating algorithms predicted emotion (arousal and valence) of the music dynamically per-second. In the static task, the arousal and valence of the complete music clip (45 seconds) were predicted. The training dataset consisted of 700 excerpts of 45 seconds, which were labelled both with dynamic annotations (1 Hz) and the static annotations, where static ratings were not derived from the dynamic ones, but were given separately. 300 clips were left out for the evaluation set. The music came from the Free Music Archive. Later, duplicates (excerpts sampled from the same song) were discovered and removed from these data, leaving 744 clips out of a 1000.

4.5.2 Task description 2014

In 2014, the static emotion characterization task was removed and a new subtask — feature design — was added instead (Aljanaki, Soleymani, & Yang, 2014). In the feature design task, new features, which had not been developed before, were proposed and applied to the arousal and valence prediction task. The feature design task was not popular and only one team submitted to it (Kumar et al., 2014). The training set

consisted of 744 clips from the previous year and 1000 new clips, all from the Free Music Archive. The time resolution for the dynamic task was changed to 2 Hz.

4.5.3 Task description 2015

In 2015, the feature design subtask was removed, leaving only the dynamic emotion characterization task. The training set consisted of 431 clips, which were selected out of 1744 clips from the previous years based on consistency measures:

1. We deleted the annotations for which Pearson's correlation with the averaged annotations for the same song is below 0.1. If fewer than 5 annotators remain after the deletion, we discarded the song.

2. For the remaining songs and remaining annotations, we calculated the Cronbach's α. If it was bigger than 0.6, the song was retained.

3. The mean (bias) of every dynamic annotation was changed to match the averaged static annotation for the same song (see Formula 4.1.

The evaluation set consisted of 58 full length songs, one half from the medleyDB dataset (Bittner et al., 2014) of royalty-free multitrack recordings and another half from the `jamendo.com` music website, which provides music under Creative Commons license. The songs were ≈ 4 minutes (234 ± 107 s) long on average. The time resolution for the annotations was 2 Hz. The participants had to submit:

- Features that were used in their approach. These features we used to train a baseline regression method (SVR with a linear kernel) to estimate dynamic affect. Any features automatically extracted from the audio or the metadata provided by the organizers were allowed.

- Results using baseline features.

- Any combination of the features and machine learning methods.

4.5.4 Evaluation metrics

We used two evaluation metrics to compare the performance of different methods: Pearson's correlation coefficient between the ground truth and predicted values for each song, averaged across songs, and root mean square error (RMSE), averaged the same way. In 2013 and 2014, we used the correlation coefficient as the main metric and RMSE as an auxiliary metric to break the ties. The tie is a situation where the difference between two methods adjacent in the ranking is not significant based on the one sided Wilcoxon test ($p < 0.05$). In 2015, we used RMSE as our primary metric. The RMSE metric measures how far is the prediction of the emotion from the annotated emotion of the song, and correlation measures whether the direction of change is predicted correctly. We will also report concordance correlation coefficient (ρ_c) as an evaluation metric. This metric was proposed by Lin (L. I.-K. Lin, 1989) in 1989 and is defined as follows:

$$\rho_c = \frac{2 s_{xy}}{s_x^2 + s_y^2 + (\overline{x} + \overline{y})}, \tag{4.6}$$

where x and y are the vectors of numbers to compare, s_x^2 is the variance of x, s_{xy} is the covariance of x and y, and $\overline{x}$ is the mean of vector x. Concordance correlation coefficient takes into account both components — RMSE and r. Bigger ρ_c indicates better performance.

4.6 Analysis of the proposed systems

In this section we will analyse the best systems suggested over the three years of the benchmark. In the last edition of the benchmark (year 2015) we asked the participants to provide their feature sets, and to run their algorithms on the baseline feature set. In this way we can conduct a systematic evaluation of the algorithms and feature sets separately.

4.6.1 Task participation

Three teams participated in the task in the year 2013 and the results were analyzed in (Soleymani, Aljanaki, et al., 2014). In 2014, there were six teams and in 2015, twelve teams. Every team wrote a working notes paper, and all the papers are available in the proceedings of the corresponding year's MediaEval benchmark. The last edition of the benchmark had most participating teams, and most of the algorithms from the previous years featured in the last edition. Here we will mostly analyse the results of the benchmark held in the year 2015.

4.6.2 Performance in the challenge over the years

Tables 4.4, 4.5 and 4.6 show the results of the benchmark in 2013, 2014 and 2015. The results are sorted by RMSE of arousal ascending (best solutions on top). The column "Method" shows the abbreviation of the machine learning algorithm used by a particular team, and a reference to the working notes paper that was published in the proceedings, where the details of the approach are explained. All the methods beat the baseline, shown on the bottom row. The baseline method is multi-linear regression.

In 2013 and 2015, LSTM-RNN based solutions were the best both for arousal and valence, in the year 2014 LSTM-based solution was second best for arousal, but best for valence.

In 2013, all the teams used different feature sets. The results are analyzed in detail in (Soleymani, Aljanaki, et al., 2014).

In 2014, solutions (Imbrasaite et al., 2014) and (Coutinho, Weninger, Schuller, & Scherer, 2014) used openSMILE feature sets. The rest of the teams used other features. The combination that produced the best result for arousal (but worse than baseline result for valence), was a combination of a Kalman filter and low-level features: MFCCs, zero-crossing rate, spectral flux, centroid, rolloff, and spectral crest factor.

Table 4.6 shows the 10 best solutions for year 2015 (each of the 12 teams submitted 3 runs, which produced more than 30 different solutions, and we will only show the ones that performed best). All of the solutions listed use the baseline openSMILE feature set, but it is usually transformed in some way (feature selection, dimensionality reduction, deep learning), or more features are added.

Method	Arousal		Valence	
	RMSE	ρ	RMSE	ρ
BLSTM-RNN Weninger et al. (2013)	.08 ± .05	.31 ± .37	.08 ± .04	.19 ± .43
GPR Markov et al. (2013)	.10 ± .05	.11 ± .36	.09 ± .05	.06 ± .28
SVR Aljanaki et al. (2013)	.10 ± .06	.14 ± .28	.12 ± .07	−.01 ± .27
Baseline	.25 ± .11	.16 ± .36	.23 ± .10	.06 ± .30

Table 4.4: Performance of the algorithms for arousal and valence in 2013. **BLSTM-RNN** — Bi-directional Long-Short Term Memory Recurrent Neural Networks. **GPR** — Gaussian Processes Regression. **SVR** — Support Vector Regression.

Method	Arousal		Valence	
	RMSE	ρ	RMSE	ρ
KF Markov et al. (2014)	.08 ± .05	.21 ± .57	.14 ± .07	.17 ± .5
LSTM Coutinho et al. (2014)	.10 ± .05	.35 ± .45	.08 ± .05	.20 ± .49
CCRF Yang et al. (2014)	.12 ± .05	.23 ± .56	.09 ± .05	.12 ± .55
CCNF Imbrasaite et al. (2014)	.12 ± .07	.18 ± .60	.10 ± .06	.07 ± .29
MR Fan et al. (2014)	.12 ± .05	.17 ± .41	.09 ± .05	.10 ± .37
PLSR Kumar et al. (2014)	.13 ± .07	.28 ± .50	.10 ± .06	.15 ± .5
Baseline	.14 ± .06	.18 ± .36	.10 ± .06	.11 ± .34

Table 4.5: Performance of the algorithms for arousal and valence in 2014. **KF** — Kalman Filter. **LSTM** — Long-Short Term Memory Recurrent Neural Network. **CCRF** — Continuous Conditional Random Fields. **CCNF** — Continuous Conditional Neural Fields. **MR** — Multi-level regression. **PLSR** — Partial Least Squares Regression.

Method	Arousal		Valence	
	RMSE	ρ	RMSE	ρ
BLSTM-RNN Xu et al. (2015)	.12 ± .06	.66 ± .25	.17 ± .09	.12 ± .54
BLSTM-ELM Xu et al. (2015)	.12 ± .05	.63 ± .27	.15 ± .08	.15 ± .47
LSTM-RNN Coutinho et al. (2015)	.12 ± .06	.61 ± .28	.19 ± .10	.03 ± .50
LSTM-RNN Coutinho et al. (2015)	.12 ± .06	.60 ± .29	.19 ± .10	.02 ± .49
LS Gupta et al. (2015)	.12 ± .05	.65 ± .22	.17 ± .09	.01 ± .50
LSB Gupta et al. (2015)	.12 ± .05	.59 ± .23	.17 ± .09	.05 ± .43
SVR Liu et al. (2015)	.12 ± .05	.56 ± .27	.19 ± .10	−.02 ± .45
SVR+CCRF Cai et al. (2015)	.12 ± .05	.54 ± .27	.17 ± .09	.02 ± .43
AE-HE-BLSTM Xu et al. (2015)	.12 ± .06	.52 ± .37	.17 ± .09	.02 ± .51
LSTM-RNN Coutinho et al. (2015)	.12 ± .06	.61 ± .25	.19 ± .10	.00 ± .50
Baseline	.14 ± .06	.37 ± .26	.18 ± .09	−.01 ± .38

Table 4.6: Performance of the algorithms for arousal and valence in 2015. **BLSTM-ELM** — BLSTM-based multi-scale regression fusion with Extreme Learning Machine. **AE-HE-BLSTM** — BLSTM + features created through deep learning. **LS** — Linear regression + Smoothing. **LSB** — Least Squares Boosting + Smoothing. **SVR + CCRF** — SVR + Continuous Conditional Random Fields.

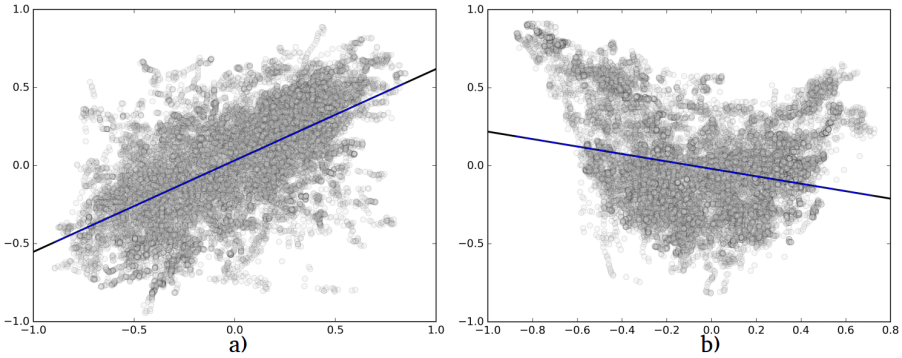


Figure 4.8: The distribution of the annotations on the Valence–Arousal plane. a) development set 2015, b) test set 2015.

4.7 Evaluation of the algorithms

In this section we describe an evaluation of the algorithms that use the same feature set. The baseline features were extracted using the openSMILE toolbox (Eyben et al., 2013). We obtained 260 low-level features (mean and standard deviation of 65 low-level acoustic descriptors, and their first-order derivatives) from non-overlapping segments of 500 ms, with a frame size of 60 ms with a 10 ms step.

Table 4.7 shows the evaluation of 10 algorithms participating in the year 2015 challenge on this feature set. The 10 best approaches are shown. The performance in terms of RMSE for arousal is the same for all the solutions (though the correlation coefficient is different), indicating that the algorithms might have reached some sort of ceiling in performance with this combination of annotations and features.

The algorithms are sorted by their performance according to RMSE on arousal ascending (RMSE increases and performance decreases). The algorithms show very good performance on arousal and completely unsatisfactory performance on valence. It is a known issue that valence is much more difficult to model than arousal, but not to the extent that we observe.

In 2013 and 2014, arousal and valence annotations were highly positively correlated. In 2015, they were not. We hypothesize that because of the high correlation the algorithms did not train to recognize valence-specific cues and could not perform well on the test set. Figure 4.8 shows the scatter plots of all the annotations (every second of every song for every annotator) along with regression lines.

Almost all the solutions listed in Table 4.7 (best solutions from 2015) are either based on LSTM-RNN networks or SVR. Solutions suggested by the team SAILUSC (Gupta & Narayanan, 2015) are exceptions from this rule. Their solutions are based on linear regression with smoothing, or least squares boosting. LSTN-RNN networks are capable of incorporating local context in their predictions. A smoothing step also incorporates the context, though it can not learn the dependencies in time-series. To sum up, the approaches that only use a context of 500 ms loose to the approaches that use bigger contexts. This problem is also discussed in the next chapter.

Method	Arousal			Valence		
	RMSE	ρ	ρ_c	RMSE	ρ	ρ_c
BLSTM-RNN Xu et al. (2015)	.12 ± .06	.66 ± .25	.30 ± .24	.15 ± .08	.15 ± .47	.06 ± .17
BLSTM-ELM Xu et al. (2015)	.12 ± .05	.63 ± .27	.25 ± .22	.15 ± .08	.15 ± .47	.06 ± .17
LR+S Gupta et al. (2015)	.12 ± .05	.65 ± .22	.32 ± .23	.17 ± .09	.01 ± .50	.01 ± .19
LSB Gupta et al. (2015)	.12 ± .05	.59 ± .23	.30 ± .24	.17 ± .09	.05 ± .43	.01 ± .18
LSTM-RNN Coutinho et al. (2015)	.12 ± .06	.61 ± .25	.31 ± .26	.19 ± .10	.00 ± .50	.01 ± .20
Combo Gupta et al. (2015)	.12 ± .05	.64 ± .23	.28 ± .22	.17 ± .09	.00 ± .48	.01 ± .19
SVR Liu et al. (2015)	.12 ± .05	.56 ± .27	.31 ± .25	.19 ± .10	−.02 ± .45	.00 ± .18
SVR+CCRF Cai et al. (2015)	.12 ± .05	.52 ± .30	.22 ± .22	.17 ± .10	.00 ± .43	.00 ± .13
LSTM-RNN Pellegrini et al. (2015)	.12 ± .06	.59 ± .24	.30 ± .23	.18 ± .09	.03 ± .48	.00 ± .20
SVR Xu et al. (2015)	.12 ± .07	.56 ± .24	.08 ± .08	.15 ± .09	.01 ± .40	.00 ± .04

Table 4.7: Performance of the different algorithms for arousal and valence, using the baseline feature set. **Combo** — An unweighted combination of LS, LSB and Boosted ensemble of single feature filters.

4.8 Evaluation of the feature sets

In this section we will analyze the features proposed by the teams in 2015 by building a system using one machine learning algorithm, but different feature sets. We chose the best performing algorithm of the previous years — LSTM-RNN.

The network has three hidden layers with 250, 150 and 50 nodes (the architecture used by the ICL team). We used the parameters of the network which were optimised for our data by the ICL team (Coutinho et al., 2015), i.e., the number of memory blocks in each hidden layer, the learning rate (LR), and the standard deviation of the Gaussian noise applied to the input activations. Every layer was pretrained (in a supervised way) before the next layer was added and the network was trained again. We used 20-fold cross validation.

4.8.1 Proposed features

A variety of software for audio signal processing and feature extraction was used by participants: Marsyas, MIRToolbox for Matlab, PsySound, openSMILE, Essentia, jAudio. Mostly, participants used the features that are known to be important for emotion recognition, such as MFCC, tempo, loudness, low level spectral features related to timbre. Few novel features were proposed. Kumar et al. (2014) proposed two new types of features: compressibility features, which describe how much the audio can be compressed, median spectral band energy, which describes the spectral bandwidth of the audio. The compressibility of audio was strongly positively correlated with static arousal ratings (Pearson's $r = 0.656$). Cai et al. (2015) proposed edge orientation histograms on mel-frequency spectrogram.

4.8.2 Results on development and test set cross-validation

Table 4.8 shows the evaluation of the feature sets on Valence, ordered by Concordance Correlation Coefficient of the results on evaluation set, in descending order. The best performing feature set for valence (by JUNLP team) is a baseline feature set, with feature selection applied to it to select the features optimized for valence recognition. The

Method	Evaluation set			Development set		
	RMSE	ρ	ρ_c	RMSE	ρ	ρ_c
JUNLP (2) Patra et al. (2015)	.27 ± .13	.19 ± .35	.08 ± .15	.26 ± .15	.22 ± .51	.09 ± .24
PKUAIPL Cai et al. (2015)	.27 ± .14	.16 ± .35	.07 ± .20	.22 ± .13	.33 ± .50	.16 ± .27
HKPOLYU Liu et al. (2015)	.28 ± .14	.19 ± .36	.06 ± .17	.21 ± .13	.41 ± .53	.20 ± .28
JUNLP (3) Patra et al. (2015)	.28 ± .13	.17 ± .33	.06 ± .14	.26 ± .15	.23 ± .53	.09 ± .24
UNIZA (1) Chmulik et al. (2015)	.29 ± .14	.14 ± .37	.06 ± .14	.22 ± .14	.32 ± .50	.16 ± .27
ICL Coutinho et al. (2015)	.30 ± .14	.12 ± .40	.06 ± .16	.22 ± .13	.30 ± .50	.15 ± .27
JUNLP (1) Patra et al. (2015)	.28 ± .13	.12 ± .39	.05 ± .15	.22 ± .14	.32 ± .50	.15 ± .27
UNIZA (2) Chmulik et al. (2015)	.29 ± .16	.09 ± .40	.05 ± .17	.23 ± .14	.31 ± .49	.15 ± .26
IRIT-SAMOVA Pellegrini et al. (2015)	.29 ± .15	.08 ± .41	.05 ± .16	.23 ± .14	.33 ± .50	.16 ± .27
MIRUtrecht Aljanaki et al. (2015)	.29 ± .14	.11 ± .43	.04 ± .15	.24 ± .15	.30 ± .49	.13 ± .23

Table 4.8: Performance of the different feature sets on Valence, development and evaluation sets of year 2015, 20 fold cross-validation.

second best feature set, suggested by PKUAIPL team, consisted of the baseline feature set with an addition of three types of features: MFCC and Δ MFCCs, edge-orientation histograms and standard low-level spectral features. In addition, team PKUAIPL applied auto-regressive and moving average filters to the features to account for the temporal changes in music, and added the output as new features to the feature vector. Team HKPOLYU suggested a supervised transformation on the baseline feature set (Arousal–Valence similarity preserving embedding). This transformation maps high-dimensional feature vectors to a lower-dimensional space so that for similar songs (in terms of Valence or Arousal) the feature vectors are also closer in this low-dimensional space.

Table 4.9 shows the evaluation of the feature sets on Arousal, ordered by Concordance Correlation Coefficient of the results of development set, descending. Teams HKPOLYU, THU-HCSIL and IRIT-SAMOVA suggested the best features for arousal. The features by the team HKPOLYU were already described above. Team THU-HCSIL applied Deep Belief Networks to a set of features extracted with the openSMILE and MIRToolbox, in order to learn the higher representation for each group of features independently, which were then fused by a special Autoencoder with a modified cost function considering sparse and heterogeneous entropy, to produce the final features at a rate of 2 Hz for the succeeding regression. Team IRIT-SAMOVA could achieve a very good performance with a very simple feature set consisting of 6 measurements on bands of a Bark scale for spectral valley, and spectral flatness on ERB and Bark scale, for a total of only 8 features. Spectral flatness provides a way to quantify how noise-like a sound is. Spectral valley is a feature derived from the so-called spectral contrast feature, which represents the relative spectral distribution.

The general conclusions are that it was important to do feature selection or transformation for valence or arousal dimensions separately. Also, for arousal, very simple low-level spectral features worked quite well.

4.9 Discussions and perspectives

In this chapter we released the MediaEval Database for Emotional Analysis in Music (DEAM), the biggest available dataset of dynamic annotations (valence and arousal annotations for 1802 songs and song excerpts licensed under Creative Commons with

Method	Development set			Evaluation set		
	RMSE	ρ	ρ_c	RMSE	ρ	ρ_c
HKPOLYU Liu et al. (2015)	.20 ± .12	.48 ± .47	.23 ± .28	.22 ± .12	.39 ± .41	.24 ± .26
THU-HCSIL Xu et al. (2015)	.21 ± .13	.46 ± .42	.22 ± .26	.27 ± .12	.33 ± .40	.16 ± .22
IRIT-SAMOVA (3) (2015)	.21 ± .13	.49 ± .43	.21 ± .27	.24 ± .13	.52 ± .37	.25 ± .25
IRIT-SAMOVA Pellegrini et al. (2015)	.21 ± .12	.45 ± .44	.21 ± .27	.24 ± .12	.43 ± .30	.22 ± .22
JUNLP (1) Patra et al. (2015)	.21 ± .12	.45 ± .43	.19 ± .26	.24 ± .12	.52 ± .31	.26 ± .24
UNIZA (2) Chmulik et al. (2015)	.22 ± .12	.41 ± .44	.19 ± .25	.24 ± .12	.48 ± .32	.26 ± .24
PKU-AIPL Cai et al. (2015)	.21 ± .12	.40 ± .44	.19 ± .26	.23 ± .10	.52 ± .30	.32 ± .27
UNIZA (1) Chmulik et al. (2015)	.22 ± .12	.40 ± .44	.19 ± .25	.25 ± .13	.49 ± .30	.27 ± .23
JKU-Tinnitus (2) Weber et al. (2015)	.22 ± .13	.39 ± .45	.19 ± .26	.30 ± .14	.06 ± .38	.04 ± .17
JKU-Tinnitus (1) Weber et al. (2015)	.22 ± .12	.38 ± .43	.19 ± .26	.29 ± .14	.09 ± .39	.05 ± .15

Table 4.9: Performance of the different feature sets on Arousal, development and evaluation sets of year 2015, 20 fold cross-validation.

2 Hz time resolution). Using DEAM, we organized the 'Emotion in Music' task at MediaEval Multimedia Evaluation Campaign from year 2013 to 2015. The benchmark attracted in total 21 teams to participate in the challenge.

During the three years, changes were introduced to the data collection routine, which led to the improvement of the quality of the annotations. In the first two years of the benchmark, the size of the segment was chosen in such a way that both static and dynamic ratings were possible. This resulted in a compromise: we selected the window of 45 seconds. This window appears to be too short to capture a lot of emotional variation, and too long to make estimating the static emotion unambiguous. In 2015, we opted for full-length songs. In combination with preliminary listening and more careful selection of workers, the quality of the annotations was improved. However, full-length songs might also not be the optimal solution because the annotation procedure is very demanding and requires a lot of concentration, and there is a danger that full-length song annotation stretches the limits of what human annotators are capable of. This question requires more investigation. Also, in 2015 we employed a MEVD method and manual filtering to select songs with more emotional variation, in particular songs in the upper left and lower right quadrants of the V–A space. This led to a different distribution of labels, which made it possible to identify problems with valence recognition.

Estimating the absolute value of an emotion in real time is difficult for the annotators, and often, though the direction of change is indicated correctly, the magnitude is not. We proposed to alleviate this problem by resampling the annotations into the same range using the overall emotion of the song (annotated separately).

Since the first edition of the *Emotion in Music* task in 2013 we have opted for characterizing the per-second emotion of music as numerical values in two dimensions — valence (positive or negative emotions expressed in music) and arousal (energy of the music) (V–A) (Russell, 1980; Thayer, 1989), making it easier to depict the temporal dynamics of emotion variation. The V–A model has been widely adopted in affective computing (Kim et al., 2010; Huq, Bello, & Rowe, 2010; Eerola, 2014; Y.-H. Yang & Chen, 2012; Barthet et al., 2012; J.-C. Wang, Yang, et al., 2015; Koelstra et al., 2012; Soleymani, Larson, Pun, & Hanjalic, 2014; S. Wang & Ji, 2015). However, the model is not free of criticisms and some other alternatives may be considered in the future. For example, the V–A model has been criticized for being too reductionist and

that other dimensions such as dominance should be added (Collier, 2007). Moreover, the terms 'valence' and 'arousal' may be sometimes too abstract for people to have a common understanding of its meaning. Such drawbacks of the V–A model can further harm the inter-annotator agreement of the annotations for an annotation task which is already inherently fairly subjective.

In the benchmark, we resampled the annotations to either 1 Hz or 2 Hz. This led to benchmark participants using 1 or 0.5 second windows as the main unit of emotion prediction. As far as musical emotion is usually created on bigger time scales, the best algorithms for MEVD prediction were the ones that could incorporate the bigger context, either through algorithm design (LSTM-RNN) or through smoothing step applied at a later stage.

The best feature sets that were suggested for the task treated predicting valence and arousal separately, and suggested separate feature selection or dimensionality reduction steps for each emotional dimension. Again, it was shown that though arousal can be successfully modelled just with simple timbral features (spectral valley and spectral flatness), modeling valence is much more complex, and satisfactory performance was not achieved by any of the algorithms.

The benchmark was not continued in 2016. In all the three years, the same algorithm (LSTM-RNN) was winning, and the features suggested for the task were still rudimentary. Also, the problem with data consistency was not solved completely. In the next chapter we will discuss how the MEVD task could be continued further with a different approach.

Part III

Emotion based segmentation

CHAPTER 5

Emotion-based segmentation — problem statement

5.1 Introduction

The previous chapter described a benchmark for music emotion variation detection (MEVD) systems that track changes in musical emotion on the Valence–Arousal plane. Both data annotators and MEVD systems' designers faced certain challenges, which, as we will argue in this chapter, are intrinsically connected to the way in which the musical emotion was represented by means of a series of continuous measurements.

The benchmark involved collecting a new dataset each year to be provided to the participants to train and evaluate upon. As we have already emphasized in this thesis, collecting good quality ground truth data and reducing the internal ambiguity in the data are the crucial steps towards the solution of the automatic MER challenge. However, it has been very difficult to collect a ground truth for MEVD with a reasonable inter-annotator agreement. We argue in this chapter that the reason may lie in the fact that it is difficult and unnatural for the listeners to evaluate their emotional response to music in a continuous way.

During the three years of organizing the benchmark we made continuous efforts to improve the efficiency of our data collection and data cleaning procedures. However, the method of data annotation for MEVD (continuous real-time annotation with a slider) introduces obstacles to good quality data, which we will describe later in this chapter. The problems that emerged could only partially be alleviated by the techniques and post-processing steps that we suggested in the previous chapter.

A typical MEVD system tracks emotion of a piece of music over time using a short sliding window with a width of one second or smaller. The per-second MEVD approach actually does not make an unreasonable assumption that emotion in music should change every second. However, as we have seen in the previous chapter, when confronted with a task of tracking changes in emotion on a very short time scale,

the MEVD methods usually address very short audio excerpts as the target of the emotion prediction. Music needs time to unfold and create emotional meaning, and this meaning is normally communicated during bigger time spans than one second (surprise based on startle stimuli being the exception). From the results of the benchmark we saw that MEVD systems could predict changes in arousal expressed through variations in loudness and timbre with reasonable accuracy, but predicting changes in valence was much more difficult. Valence is more dependent on cues such as harmony, melody and rhythm, which only emerge on a bigger time scale than is possible to consider with short MEVD tracking windows. Therefore, though MEVD systems do detect important changes in dynamics and timbre, they are definitely missing a lot of musically meaningful cues because they tackle music in a way that is far from how music is perceived by human listeners.

Also, though it might still be interesting and important to track musical change over time with such a high time resolution, the question should be raised whether these changes are actually an expression of musical emotion or *the means of creating emotional expression* on a higher level. For example, a set of dynamic changes in loudness and tempo (i.e., *crescendo*, *diminuendo*, *sforzando*, *rallentando*) are the standard means in the repertoire of musical expressiveness. These expressive means can be related to the concept of "vitality affects" introduced by Daniel Stern in 1985 (Stern, 1985):

> ... many qualities of feeling that occur do not fit into our existing lexicon or taxonomy of affects. These elusive qualities are better captured by dynamic, kinetic terms, such as "surging", "fading away", "fleeting", "explosive", "crescendo", "decrescendo", "bursting", "drawn out", and so on.
>
> ...
>
> The expressiveness of vitality affects can be likened to that of a puppet show. The puppets have little or no capacity to express categories of affect by way of facial signals, and their repertoire of conventionalized gestural or postural affect signals is usually impoverished. It is from the way they move in general that we infer the different vitality affects from the activation contours they trace. Most often, the characters of different puppets are largely defined in terms of particular vitality affects; one may be lethargic, with drooping limbs and hanging head, another forceful, and still another jaunty.

What Daniel Stern says about puppet shows can be largely applied to music's expressiveness. In the same way, dynamic and temporal contours can be a way of communicating character and emotion, and, probably, should be analysed in a larger temporal context.

In addition to the shortness of the sliding window, the excerpts of music used as the ground truth for continuous annotation are usually also short. In the work of Schmidt et al. 15 second excerpts were used (Kim et al., 2008; Schmidt, Turnbull, & Kim, 2010; Schmidt & Kim, 2011), and in the first two years of the MediaEval Emotion in Music benchmark the excerpts were 45 seconds long (Soleymani et al., 2013; Aljanaki, Soleymani, & Yang, 2014). Ironically, the excerpts that are normally used for

song level MER and assume stability of emotion within the excerpt are usually about the same size (15 seconds to 1 minute). This conflict of assumptions happens because researchers try to annotate as much different music as possible, while reducing the annotation burden. Both assumptions can theoretically be met — some music has very little emotional variability over 1 minute, some music goes through a lot of transitions over 15 seconds.

To establish the typical length of emotionally stable segment for classical music, Liu et al. made an educated guess that emotion is stable over a duration of standard musical period of 16 bars, and hence the shortest emotionally stable segment should be 16 seconds (for very fast music) (D. Liu, Lu, & Zhang, 2003). Another attempt to establish the typical emotionally stable unit was made by Xiao et al., who classified excerpts of different lengths by emotion and found that excerpts of 8 or 16 seconds have a better classification accuracy than excerpts of 4 or 32 seconds (Xiao, Dellandrea, Dou, & Chen, 2008). Of course, the longer the excerpt, the better the chance for encountering some emotional variation, but there are drawbacks to lengthening the excerpts. Continuous emotion annotation in real time is straining for the annotators' concentration capability and abusing that capability may degrade the quality of the annotations. Also, collecting the annotations becomes very costly when the excerpts are long. However, the short excerpts (e.g., 15 seconds) usually have no serious musical development.

Due to all these issues, here we suggest that a continuous stream of per-second measurements is not a suitable representation of musical emotion for ground truth collection purposes. Therefore we propose a different approach to tracking emotion over time. In our approach we assume that music consists of a series of emotionally stable segments (which are normally much longer than 1 second), and transitions between them (unstable segments). An assumption that music consists mostly of long stretches of stable emotion, is the one employed by static (song-level) MER methods. It is natural for listeners to describe musical content by applying emotional labels to musical excerpts or complete pieces. This kind of labels are used by most music services to categorize their data. But in order for the MER method to work correctly, a classified excerpt must contain a single emotion. This problem is usually just neglected by static MER methods, which often use ground-truth excerpts picked by randomly sampling the audio and filtering out the excerpts that receive contradictory ratings from the experts. Also, sometimes the problem is circumvented by picking the most representative part of the song for classification (e.g., a chorus), assuming that emotion does not change during the chorus. In (Y.-H. Yang et al., 2008), 25-second segments that expressed a single dominant emotion were selected manually. In (Lu, Liu, & Zhang, 2006), 20-second segments representative of the song were selected by the experts. With bigger datasets, manual filtering of the excerpts might not work. In (Hu et al., 2008), with the dataset consisting of 1250 tracks, a heuristic was used — "the clips were extracted from the middle of the tracks which are assumed to be more representative than other parts".

The problem of music segmentation by emotion has received very little attention. There are still many questions to answer. What is a typical length of an emotionally stable fragment in music? Is emotional segmentation explained by structural segmentation? Which segmentation methods and features work best when applied to emotional boundary detection?

In this chapter we are going to deal with these questions. For this purpose we assemble a dataset of 52 triple-annotated pieces from the RWC music database (Goto, Hashiguchi, Nishimura, & Oka, 2002), which also were structurally annotated (these pieces are also in the SALAMI dataset (Smith, Burgoyne, Fujinaga, Roure, & Downie, 2011)). We obtain a little under 2000 annotated emotional boundaries (around 640 from each of the annotators). We compare emotional and structural segmentation of music, analyze the inter-annotator agreement and the average stable segment length. Then we apply four segmentation algorithms to emotional segmentation problem and benchmark them on our dataset.

5.1.1 Organization

The rest of the chapter is organized as follows. In section 5.2 we describe related research. In section 5.4 we describe the annotations of emotional boundaries and analyze them to answer some of the questions asked above. In section 5.5 we compare different segmentation methods when applied to a problem of detecting emotional boundaries in music. Section 5.6 concludes the chapter.

5.2 Background

In this section we will review the research related to reliability of continuous music emotion measurement, and the approaches that were suggested for emotional boundary detection or can be applied to this task.

5.2.1 Reliability of continuous music emotion measurement

We have already described the various interfaces that are used for collecting continuous responses to music in Chapter 3. All of these interfaces originate from the Continuous Response Digital Interface (CRDI), first introduced by Robinson (1988). This interface initially had a form of 256 degree dial that could be turned, or a lever which could be pulled up and down. The interface was used to measure a wide variety of phenomena related to sound, video and music, such as tension, aesthetic response, conductor evaluation, mood, intensity, preference, instrument family prominence, focus of attention etc. The various uses of the instrument, its validity and reliability are reviewed in (Geringer, Madsen, & Gregory, 2004), where it is also mentioned that the viability of using the CRDI for such complex multi-dimensional phenomena as musical emotion should still be studied. A study to measure such a viability was conducted by Schubert (1999). Four complete music pieces (from the romantic period of classical music) were annotated by 67 annotators using a two-dimensional Valence–Arousal plane. The annotations were analyzed, showing that they indeed reflect the emotion of the music (for instance, the annotators kept the cursor in the correct quadrant most of the time) and annotators react to changes. Also, a test-retest experiment was conducted, showing that the annotators are consistent with themselves 6 month earlier. A possible problem was detected with the “scaling” of responses: ”on the whole, people do not like to use large regions of the emotion-space during continuous response. They prefer to keep something ‘up their sleeve’ in case a more extreme episode is

expressed". Also, the strong positive correlation on the test-retest experiment does not exclude the possibility that there could be systematic biases in the annotations.

5.2.2 Segmentation by emotion

Liu et al. suggested the first method for emotional boundary detection (D. Liu et al., 2003). A sliding window of 16 seconds was used to extract features from the audio of a song. Then, feature distributions were compared on both sides of each timestamp, a novelty curve formed and peaks detected (suggesting a presence of a boundary). There was no systematic evaluation of the method (for the lack of data on emotional segmentation). We will benchmark this method along with other methods in section 5.5. A similar method (computing difference features using a sliding window) was suggested by Turnbull and Lanckriet for structural music segmentation (Turnbull & Lanckriet, 2007). Both unsupervised (based on peak picking) and supervised versions were tested, with a supervised version showing superior performance. In (Lu et al., 2006), the method described in (D. Liu et al., 2003) was modified to include loudness contours as a preprocessing step to detect potential boundaries. Also, an evaluation on 9 musical pieces (with 63 boundaries in them) showed 84.1% recall.

In (Wu, Zhong, Horner, & Yang, 2014) songs were segmented using aligned lyrics annotations on an assumption that most often emotion is stable within one sentence. Then, a hierarchical Bayesian model was built for multi-label classification. This method only would work for vocal music, and though it is plausible that emotion is stable within a sentence, this is definitely not guaranteed. Due to the absence of ground-truth on emotional boundaries in (Wu et al., 2014), it is unclear how well the sentences in the lyrics actually correspond to the emotional structure of the musical piece.

In (Deng & Leung, 2015), dynamic texture models were trained corresponding to the quadrants of resonance-arousal-valence model and applied to predict musical emotion continuously. A transition from one quadrant to another signified an emotional boundary. This approach can only detect very coarse changes in emotion due to lack of resolution in the emotional space.

Structural segmentation of music is a much more developed area of research than emotional segmentation. It is not yet clear how related they are, actually, but in this chapter we use concepts, evaluation methodology and methods of structural segmentation, for need of a place to start. For more information on structural segmentation aims and methodology, we refer the reader to (Paulus, Müller, & Klapuri, 2010) and a more recent source — a chapter in a MIR textbook dedicated to segmentaion of musical audio by Müller (2015).

5.3 Challenges of continuous emotion annotation

Dynamic MER relies on human ground-truth in the form of continuous emotional annotations, which are typically recorded by an annotator continuously moving their cursor in a one or two-dimensional space (see Chapters 3 for a review of interfaces). It seems that this task is extremely difficult for humans, which is, in particular, indicated by a low inter-annotator agreement as compared to static annotations (where, due to task subjectivity, it is also not very high). We will calculate Kendall's W for the

two public datasets. Kendall's W is a non-parametric statistical test that outputs values from 0 (no agreement) to 1 (total agreement). Kendall's W is computed on ranks and is linearly related to another non-parametric test on ranks, Spearman's correlation coefficient (but it is capable of correlating multiple pairs of ratings). For the MediaEval dataset (Aljanaki, Soleymani, & Yang, 2014), the average Kendall's W is 0.23 ± 0.16 for arousal and 0.28 ± 0.21 for valence, and for the MoodSwings Lite dataset (Speck et al., 2011) the mean Kendall's W is 0.21 ± 0.14 for arousal and 0.23 ± 0.17 for valence. All these numbers indicate weak agreement. As we showed in the previous chapter, through training of the annotators and investing extra effort, such as preliminary listening, it is possible to achieve satisfying results. Also, careful selection of music pieces with obvious emotional content might be a factor (Coutinho & Cangelosi, 2011). Chapter 3 gives more information on the agreement for continuous annotation. There are several typical problems arising when annotating music continuously:

1. A dimensional annotation interface has an absolute scale. Giving absolute ratings is easier when evaluating music statically (comparing a piece to all existing music). When comparing a piece with itself over time, humans tend to think of occurring changes relatively. This leads to a difference in position and magnitude, though the direction of change can be indicated uniformly (e.g., see Figure 5.1). This demand to annotate on an absolute scale also creates a "scaling" problem discovered in (Schubert, 1999) and described in Section 5.2.

2. Though it is not explicitly requested from the annotators to move their cursor at all times, the task demands (necessity to track and respond continuously, short music excerpts) lead to some of the annotators evaluating every single musical event (e.g., see Figure 5.1). This results in annotations on widely different 'zoom level' (sections, phrases and even individual notes or drum beats).

3. There is a variable time lag in the annotation. The problem of initial orientation time (the annotators need time to understand the mood of the music and start annotating) can be solved by clipping the beginning of the annotation. But with every change in emotion, annotators again react with a time lag, which often depends on how sudden the change was (whether there was a startle response). A similar effect was reported in (Schubert & Dunsmuir, 1999), where it was described that listeners react instantly to sudden bursts of loudness, whereas more gradual changes in loudness can delay the response by 2 to 4 seconds.

4. Short excerpts usually do not contain enough emotional variation. Long excerpts strain annotators' concentration capability.

Due to all these problems we argue that continuous annotation is so difficult because it is unnatural for humans to evaluate their emotional response on a per-second basis and on absolute scale.

5.4 Analysis of emotional boundaries

Here we will describe the dataset we collected to study the emotional boundaries and start looking for approaches to emotional segmentation. This is the first dataset for this kind of task.

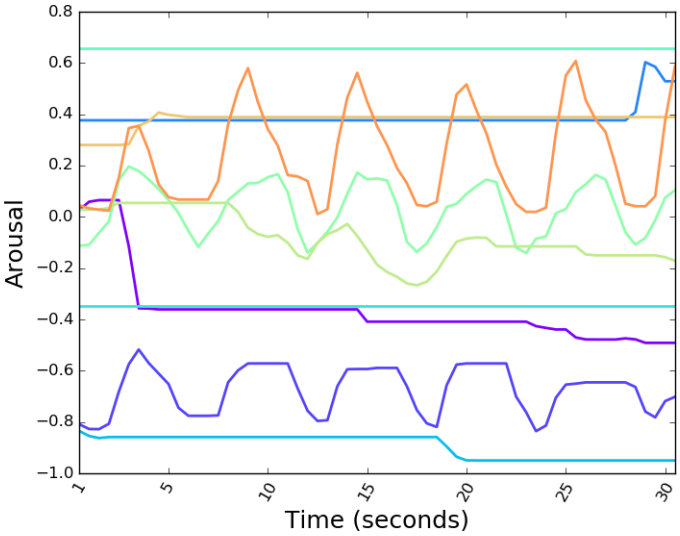


Figure 5.1: Dynamic annotation of 45 seconds of audio from the MediaEval 2014 dataset (Aljanaki, Soleymani, & Yang, 2014). Three of the annotators react to every beat of slow music by a peak in arousal. Also, these annotators agree on the direction of change, but not on the magnitude or absolute position.

5.4.1 Data

The dataset consists of 52 complete pieces from Pop, Jazz and Genre (the latter contains rock, soul, world etc. music) collections of RWC music database (Goto et al., 2002). We picked pieces that already had SALAMI (Smith et al., 2011) annotations in order to be able to compare structural and emotional segmentation. The SALAMI annotations for these pieces are single-keyed, our annotations are triple-keyed in order to enable measuring agreement.

The three annotators received instructions to mark a boundary when emotion of the piece changes. There were no explicit instructions as to what could be interpreted as an emotional boundary. They were also instructed to mark the *transitions* between stable emotional states as separate sections, in case those were long enough to be perceived as 'transition states'. In practice, this meant for instance marking a long diminuendo (fade-out) at the end of a musical piece as a separate transitional section, or any segment where emotion was not stable. In order to measure the prominence of the boundary, the annotators were also indicating the valence and arousal value (perceived, not induced emotion) on a scale from 1 to 10 for each of the sections (except for the transitional sections, which were indicated as transitional). The annotators used Sonic Visualizer to do the annotation.

In total, annotators marked 545, 676 and 702 emotional boundaries, respectively. The dataset is available from `osf.io/jpd5z`.

The mean number of boundaries per piece was 12.3 ± 5.11. The average segment length was 18.92 ± 17.76 for emotionally stable segments, and 7.16 ± 7.79 for tran-

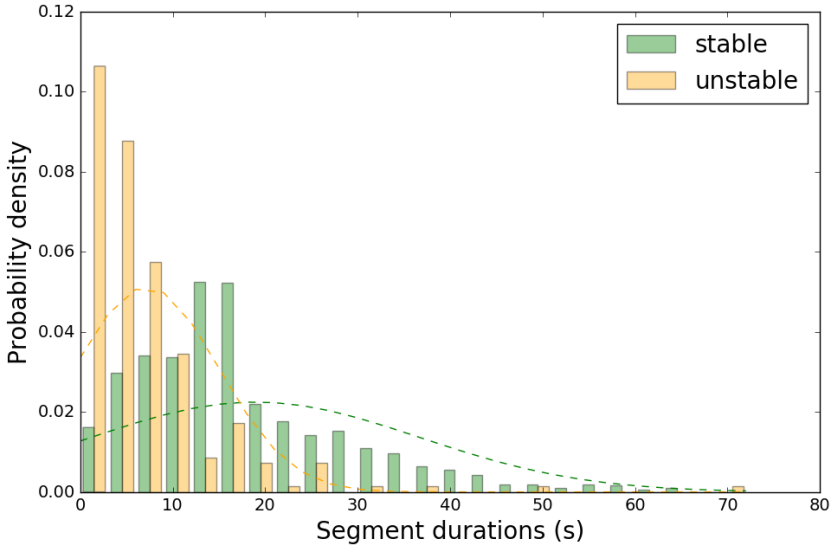


Figure 5.2: Histogram of segment durations for emotionally stable and unstable segments.

sitional sections. Figure 5.2 shows the histograms of segment lengths from the three annotators. We can see that the distribution is skewed to the right, 90% of the stable intervals are shorter than 35 seconds. Stable segments are much longer than unstable ones, on average.

5.4.2 Inter-annotator agreement on emotional segmentation

Segmentation tasks are not well-adapted to formal inter-annotator agreement calculation because of the alignment problem (boundary placement is usually imprecise). The standard way to evaluate agreement, common in the literature (Smith et al., 2011), is with the same procedure that is used to evaluate the result of a segmentation: by retrieving results with a certain tolerance window and then calculating F-measure like so:

$$F_1 = 2\,\frac{\text{precision}\cdot\text{recall}}{\text{precision}+\text{recall}}\,. \tag{5.1}$$

Table 5.1 shows the F-measure with tolerance windows of 0.5 and 3 seconds for all the pairs of annotators averaged, and for each pair separately. The agreement is lower than for the structural segmentation. We cannot compute the agreement on structural segmentation for the same 52 music pieces, because the structural annotations for them are single-keyed. We will compare emotional segmentation agreement to the agreement for other pieces in the SALAMI dataset (974 pairs of annotations). For large-large structural segmentation, the F-measure at 3 s was 0.77, and F-measure at 0.5 s was 0.69. For our annotations, the F-measure at 0.5 s window is much lower. This

is likely to be caused by the nature of the task. Though some emotional boundaries are rather abrupt, others are smeared by a transitional musical process necessary for an emotion to modulate from one state to another.

Evaluation metric	**Average**	A2→A1	A3→A2	A1→A3
Precision @ 0.5	**.49**	.47	.57	.44
Recall @ 0.5	**.48**	.45	.42	.57
F-measure @ 0.5	**.46**	.45	.46	.48
Precision @ 3	**.76**	.82	.62	.85
Recall @ 3	**.76**	.77	.84	.67
F-measure @ 3	**.73**	.78	.68	.73

Table 5.1: Inter-annotator boundary retrieval with the tolerance windows of 0.5 and 3 seconds.

Besides using retrieval with a tolerance window, we will also evaluate the agreement on the boundary placement with a standard agreement measure — Fleiss' kappa (Fleiss, 1971). This measure was developed for assessing reliability of agreement on categorical (in our case, binary) ratings on a number of items. In our case, the items are the short segments of music (we chose beats for this purpose) with binary labels whether a boundary is present within a beat or not. We used a beat-detection algorithm from Essentia to detect each beat onset. We evaluate the annotations using 2 conditions. The first condition is exact match (the boundary is indicated on the same beat by all the annotators). For that condition, the Fleiss' kappa is 0.25 ± 0.11, which indicates fair agreement. The second condition is using a window of three beats (a tolerance of 1 beat off), and the Fleiss' kappa for this condition is 0.64 ± 0.19, which indicates substantial agreement.

5.4.3 Boundary strength

For each of the segments we asked the annotators to estimate the segment's Valence and Arousal (if the segment was stable). We use these data to calculate how strong the boundary is (if the change in valence and arousal is big, the boundary is strong). The strength of the boundary is calculated by the formula:

$$\text{strength} = |a_{i+1} - a_i| + |v_{i+1} - v_i|, \tag{5.2}$$

where a_{i+1} and v_{i+1} are Arousal and Valence after the boundary, and a_i and v_i are arousal and valence before the boundary. Valence and Arousal were annotated on a scale from 0 to 10, which means that boundary strength varies from 0 to 20. Average boundary strength is 2.17 ± 3.02. For the unanimous boundaries, the average strength is 3.74 ± 3.32, and for the non-unanimous ones 2.99 ± 2.95.

The intra-class correlation coefficient (ICC) on the boundary strengths is 0.21, which indicates poor agreement. There are two causes why this could have happened: for a subjective task such as annotating Valence and Arousal only three annotators is clearly not enough, and the scaling problem that affected the dynamic annotations might still affect the per-segment ones.

5.4.4 Averaging the annotations

Evaluation metric	Valence→Arousal
Precision @ 0.5	.83
Recall @ 0.5	.74
F-measure @ 0.5	.76
Precision @ 3	.89
Recall @ 3	.80
F-measure @ 3	.83

Table 5.2: Retrieving arousal boundaries from valence boundaries with the tolerance windows of 0.5 and 3 seconds.

To create a ground truth from our annotations, we average the annotations using a 3 second window. We discard the boundaries that were marked only by one of the annotators and leave the boundaries which were marked by either two or three annotators, obtaining 533 boundaries. Of these boundaries, 475 indicate a change in arousal, and 405 indicate a change in valence. Table 5.2 shows how often the boundaries related to arousal and to valence coincide (how often valence changes when arousal changes). A decrease in F-measure with narrowing of the retrieval window shows that often the boundaries are situated close to each other (change in arousal is closely followed by a change in valence or vice versa).

5.4.5 Structural segmentation explaining emotional segmentation

In this section we investigate to what extent emotional segmentation can be explained by the structural segmentation. We compare the emotional boundary annotations to structural boundaries in the SALAMI dataset (version 1.2) for the same pieces. The SALAMI dataset contains hierarchical annotations on multiple levels — musical function (verse, chorus, etc.), lead instrument, and musical similarity on large and small scale. The function and large scale similarity annotations are 99% identical to each other for the 52 pieces we have in our dataset. Therefore, we will only include functions, small-scale and lead instrument change annotations in the comparison.

Table 5.3 shows the precision, recall and F-measure obtained when predicting emotional segmentation from structure. From the table we can see that only 66% of the emotional boundaries coincide with large section boundaries (functions). Small-scale similarity retrieval results in 90% recall and very low precision. This annotation contains too many boundaries (dissimilarly from emotional segmentation), and in any case these numbers are not very informative, because the segment sizes for small-scale segmentation are on average 4.78 ± 2.42, which, combined with a 3 second tolerance windows, allows to retrieve almost all the boundaries just by chance. A bit less than a half of the boundaries coincide with the lead instrument change.

From these results we can see that the emotional segmentation does not completely coincide with the structural. However, there is subjectivity both in emotional and structural annotations and we can not know whether the imperfect overlap happened because it actually exists or because of the subjectivity. When emotional segmentation is

retrieved from structural, the F-measure is on average 0.63, for emotional boundaries retrieved from each other the average F-measure is higher — 0.73, which leads us to conclude that some percentage of emotional boundaries is actually not explained by the structural boundaries.

Explanation of emotional segmentation

We asked one of the annotators to look through all the emotional boundaries in his annotations that did not coincide either with large scale nor small scale structural boundaries, and explain the reasons why the boundary was placed. Here is a list of all the reasons:

1. Modulation to another tonality.
2. Change in harmony/chords.
3. Change of vocal timbre or instrumental timbre.
4. End of emotionally stable section (e.g., start of crescendo towards the end of the structural segment).
5. Change in a non-lead instrument (drums, guitar).

The reasons listed are all related to some discontinuity, and none to repetition.

5.5 Evaluation of structural segmentation methods adapted to emotional segmentation

In this section we are going to evaluate methods that were proposed for structural and emotional segmentation, namely Convex NMF (Nieto & Jehan, 2013), Mood Tracking (Lu et al., 2006), the classic method by Foote (Foote, 2000) and Structural Features (Serra, Müller, Grosche, & Arcos, 2014). We implemented the Mood Tracking method as described in (D. Liu et al., 2003), and used an implementation[1] of the rest of the methods for our purposes, with parameters tuned for emotional segmentation as we describe below.

All of the methods are unsupervised and take as an input the time-series of features extracted from the audio. To compare the methods, we extract standard MFCC and

Evaluation metric	Functions			Small-scale			Instruments		
	A1	A2	A3	A1	A2	A3	A1	A2	A3
Precision @ 3	.70	.73	.56	.29	.31	.22	.48	.49	.39
Recall @ 3	.67	.66	.67	.90	.91	.87	.47	.44	.49
F-measure @ 3	.66	.67	.58	.41	.44	.33	.44	.44	.41

Table 5.3: Retrieving emotional segmentation from structural segmentation

[1] github.com/urinieto/SegmenterMIREX2014

HPCP beat-synchronized audio features using Essentia (Bogdanov et al., 2013). The music in our dataset is quite rhythmic, as the dataset consists mostly of rock, jazz and popular music. Beats are determined using the Essentia BeatTracker algorithm. All the music files have 44 100 Hz sampling rate and are converted to mono. To extract these low-level timbral and harmonic features we use a half-overlapping window of 100 ms. The features are smoothed with a median sliding window, normalized, and resampled according to detected beats (see Figure 5.3a). We use the same feature set to evaluate all the algorithms.

5.5.1 Summary of the evaluated methods

Foote

Foote's method (Foote, 2000) computes self-similarity matrix (SSM) using pairwise sample comparisons. Then, a short-time Gaussian checkerboard-shaped kernel is slid over the diagonal of the matrix, resulting in a novelty curve. The boundaries are detected by picking the peaks on the novelty curve. We experimented with different distance measures to compute the SSM and found that standardized euclidean distance gave the best results, which is computed between two vectors u and v as follows:

$$\sqrt{\sum \frac{(u_i - v_i)^2}{V[x_i]}}, \tag{5.3}$$

where V is the variance vector; $V[x_i]$ is the variance computed over all the i'th components of the points. We set the size of the checkerboard kernel to the size of the average emotionally stable segment — 20 seconds.

Convex NMF

The Convex Non-Negative Matrix Factorization method (Nieto & Jehan, 2013) (C-NMF) uses a convex variant of non-negative matrix factorization (NMF) in order to find clusters in the SSM. This algorithm focuses both on finding segments and grouping them by similarity. Here, we are only interested in the segmentation part. If an NMF of an input SSM matrix X is F, Convex NMF adds a constraint to the columns of the matrix $F = (\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_n)$ that the columns should become convex combinations of the features of X:

$$\mathbf{f}_j = \mathbf{x}_1 w_{1j} + ... + \mathbf{x}_p w_{pj} = X\mathbf{w}_j, \quad j \in [1:r], \tag{5.4}$$

where $\mathbf{x}_i$ is a column of matrix X, r is a rank of decomposition, and $w_{ij} \geq 0$, $\sum_i w_{ij} = 1$. This makes columns $\mathbf{f}_j$ interpretable as cluster centroids. We set the rank of decomposition to 4.

Mood Tracking

A method by Lu et al. (Lu et al., 2006) finds boundaries by comparing the audio features extracted from the two consecutive windows of 16 seconds and computing a

difference between them. A novelty curve is formed using an obtained difference feature, from which peaks are picked. The difference between the consecutive windows is computed using Divergence Shape:

$$D_{i|i+1} = \frac{1}{2}\mathrm{Tr}\left[(C_i - C_{i+1})\left(C_{i+1}^{-1} - C_i^{-1}\right)\right], \tag{5.5}$$

where C_i and C_{i+1} are the covariance matrices of features of windows i and $i+1$. Then, confidence of the boundary is computed:

$$\mathrm{Conf}_{i|i+1} = \exp\left(\frac{\left|D_{i|i+1} - D_{\mathrm{mean}}\right|}{D_{\mathrm{var}}}\right), \tag{5.6}$$

where D_{mean} and D_{var} are respectively the mean and variance of all divergence shapes for this song. From a list of boundary confidences the boundaries are retrieved by satisfying conditions of being a local maximum and exceeding a local adaptive threshold.

We implemented the method as it was described in (Lu et al., 2006), but it didn't work well in its original form on our data. The constraint of 16 seconds was too conservative and the adaptive threshold window was too narrow, which resulted in a low precision and F-measure. We describe an optimized version below. The optimized version performs on average about 10% better than the original method, and we only show the performance of the optimized version in Table 5.4.

Modified Mood Tracking method

The best results with Lu et al. method were obtained using a window of 4 seconds to compute the divergence shape measure. We smoothed the boundary confidence vector with a median filter before peak picking. To pick the peaks, we select a maximum in a neighbourhood of 10 beats in case it exceeds both of the two thresholds — a moving average and half of the global average.

The performance of the method improved a bit with these modifications (precision increased from 0.30 to 0.39), but the method still performed worse than other methods in our evaluation.

Structural Features

The Structural Features (SF) method (Serra et al., 2014) is both homogeneity and repetition based. It uses a variant of lag matrix to obtain structural features. The SF are differentiated to obtain a novelty curve, on which peak picking is performed. SF method calculates self-similarity between samples i and j as follows:

$$S_{i,j} = \Theta\left(\varepsilon_{i,j} - \|x_i - x_j\|\right), \tag{5.7}$$

where $\Theta(z)$ is a Heaviside step function, x_i is a feature time series transformed using delay coordinates, $\|z\|$ is a Euclidean norm, and ε is a threshold, which is set adaptively for each cell of matrix S. From matrix S structural features are then obtained using a lag-matrix, and computing the difference between successive structural features yields a novelty curve.

5.5.2 Evaluation results

	C-NMF	SF	MoodTrack	Foote
P@3	.53	.56	.39	.54
R@3	.65	.67	.55	.62
F@3	.56	.60	.44	.56

Table 5.4: Performance of C-NMF, Structural Features, MoodTracking and Foote's methods on emotional segmentation task.

Table 5.4 shows the results obtained in the evaluation. We use a tolerance window of 3 seconds for the evaluation. From the table we can see that the SF method consistently shows the best results in terms of F-measure. The method proposed in (Lu et al., 2006) consistently shows the worst results.

Segmentation methods are traditionally categorized into homogeneity, novelty and similarity based methods (Paulus et al., 2010). An emotional boundary is usually characterised by changes in loudness, timbre, harmony, instrumentation, etc. There is no straightforward connection between repetition and emotion. However, as far as most emotional boundaries are explained through structural segmentation, the same cues that are important for the structural segmentation must be important for emotional segmentation as well, including repetition. Structural Features method is the only method of the four that incorporates repetition-based cues, the rest of the methods are based on novelty and homogeneity-based cues.

We also tried to retrieve boundaries related to changes in valence and boundaries related to changes in arousal. The arousal-related boundaries were slightly easier to retrieve (F-measure of 0.56 as compared to 0.53 for valence).

5.6 Discussion

In this chapter we discussed the problems associated with dynamic MER and argued that these problems originate from the unnaturally low time resolutions that dynamic MER is usually dealing with. While static MER methods cannot deal with emotionally non-homogeneous music, dynamic MER methods approach this problem by taking the fragmentation to the extreme (the typical resolution of a dynamic MER method is 1 second). The output (per-second emotion prediction) produced by a dynamic MER method is not easily interpretable and useful.

We proposed to move to bigger time resolutions by representing music as a sequence of emotionally stable and unstable segments, and tracking these segments over time. We call this problem emotion-based segmentation.

We collected data on emotional segmentation of music; in total about 2000 emotional boundaries were annotated. In general, the annotators could agree rather well when identifying stable emotional segments, the inter-annotator F-measure was comparable to the one obtained for, supposedly less ambiguous, structural segmentation task, except at the very high precision level (0.5 s). In terms of F-measure emotional annotations are more similar to each other than to any of the structural segmentation

levels. That means that there exist some robust and important emotional boundaries which are not explained by structural segmentation. Approximately one third of the emotional boundaries did not coincide with the structural boundaries. According to some preliminary data, emotional change can occur within a structural section due to a modulation to a different tonality, start of emotionally unstable section, or change of harmony or timbre.

However, most of the emotional boundaries coincide with the structural boundaries, and the same methods are hence applicable to both tasks. About half of the emotional boundaries were accompanied by a lead instrument change.

We found that the average length of a stable emotional segment is approximately 20 seconds. This finding could be used to calculate a suitable length of musical excerpts to be employed for MEVD algorithms development and evaluation. Namely, such excerpts should be several times bigger than 20 seconds.

We evaluated different unsupervised segmentation algorithms on the task of emotional segmentation and found that local context based Mood Tracking method was least useful. This method only uses a very narrow local context to find the discontinuities in a feature matrix. The SF method performed best. This segmentation method is different from other methods by incorporation of repetition-based criteria along with homogeneity-based ones.

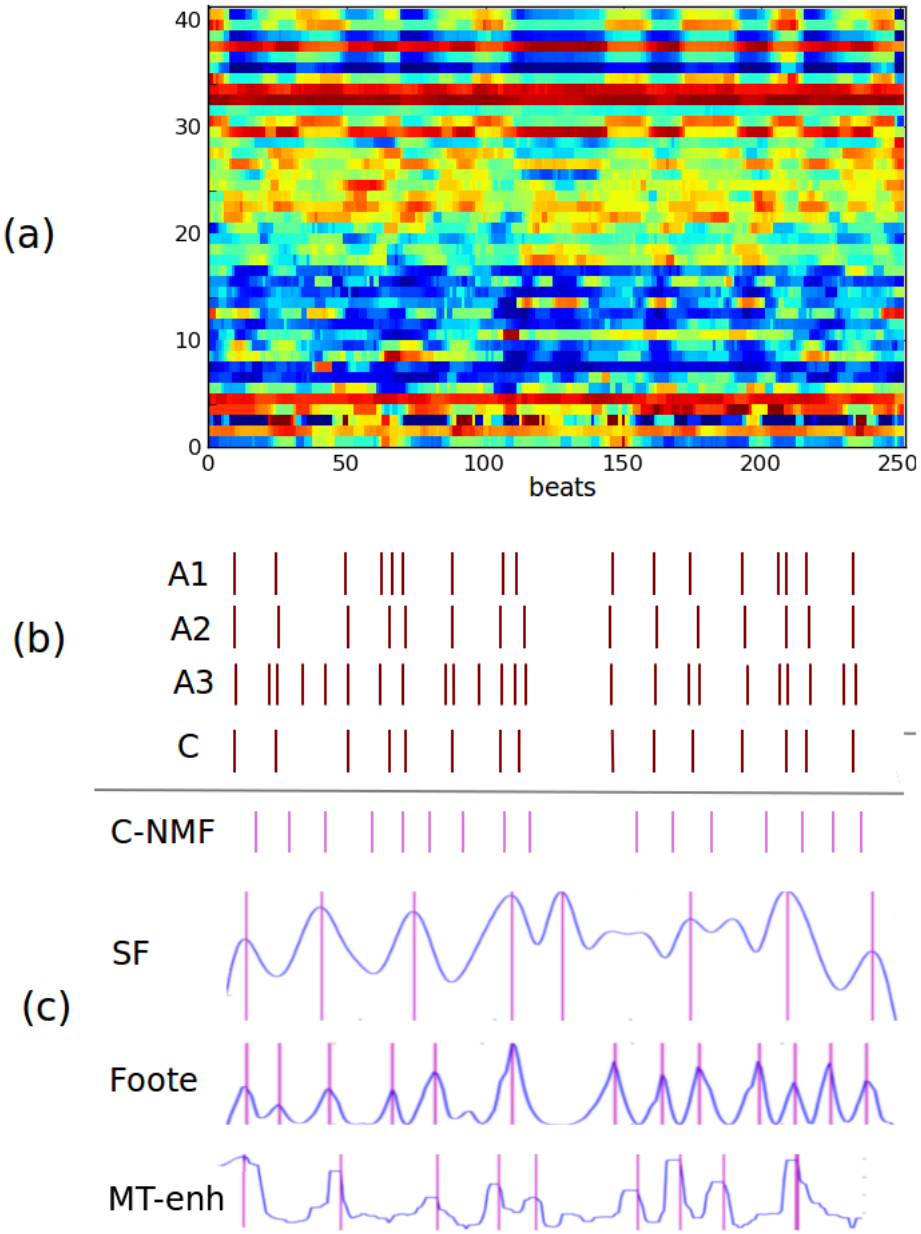


Figure 5.3: An illustration of the boundary detection process on the *Radetzky March* by J. Strauss Sr. a) Beat-synchronized features. b) Annotations. c) Novelty curves and detected boundaries for all the 4 evaluated methods.

CHAPTER 6

Supervised emotion-based segmentation

In this chapter we continue our work on emotion-based segmentation. We argue that for emotion-related tasks supervised approaches are more practical and propose an approach that uses two types of annotated ground truth information: emotional boundary locations and music excerpts annotated with Valence and Arousal. The approach consists of two steps: first, retrieval of the boundary candidates using a Convolutional Neural Network (CNN) and next, filtering the boundaries by calculating emotional boundary strength. This approach improves average emotion boundary retrieval precision from 0.67 to 0.69 and recall from 0.56 to 0.58, as compared to the best unsupervised approach (Structure Features method (Serra et al., 2014)).

6.1 Introduction

In Chapter 5 we defined the problem of emotion-based segmentation and showed that though emotional boundaries do not fully coincide with music structure, many of the boundaries do coincide and the same unsupervised approaches that work for structural segmentation can be applied to emotion-based segmentation as well. However, these approaches do not contain any emotion detection component, which limits their possibilities. Unsupervised segmentation methods detect discontinuity and repetitions in the feature matrix, where features describe the low-level timbral and harmonic properties of the sound. The start of a new structural segment is often marked by the start of a repetition, or timbral and harmonic discontinuity. The criteria that cause a listener to perceive a change in emotion are less well understood. A supervised approach would allow the algorithm to learn these criteria directly from the data.

The MIR field has relied on manually engineered features and algorithms for most of its history. When designing audio features, such as MFCC or chroma, or psychoacoustic loudness and roughness features, engineers rely on the knowledge of human

auditory system and harmonic sound organization. For instance, the knowledge about the logarithmic perception of the frequencies and loudness, and presence of harmonics in musical sound is incorporated into the algorithm.

Many MIR algorithms also rely on expert knowledge. For instance, the Structure Features approach to music segmentation exploits the knowledge about the repetitive structure of the music (Serra et al., 2014). Such unsupervised approaches have many advantages:

1. The operating principle of the algorithm is well understood, and it is easy to predict where the algorithm will succeed or fail.

2. No training data is required, the necessary knowledge has already been accumulated and analyzed by the experts of the field, and can be manually incorporated into the algorithm.

3. The algorithm designer has direct control of the parameters (such as sensitivity parameters).

However, there are also many drawbacks:

1. The phenomenon that we want to detect or classify may not be well understood (hence, there is not enough expert knowledge to translate into the algorithm). This is very often the case with emotion-related phenomena.

2. A manually engineered algorithm may not be sufficiently flexible. For instance, when applied to a different music culture or style, which operates on other principles, the algorithm has to be substantially adapted, or a new algorithm must be designed.

3. Depending on the complexity of the task, designing the algorithm may be more time and resource consuming than collecting the training data.

These drawbacks are eliminated by the supervised methods. Emotional segmentation task can benefit a lot from a machine learning approach, because the criteria for emotional boundary placement are not well understood yet, and these criteria might well differ per musical style and culture (and, perhaps, even per song, depending on its emotional variability).

For some of the MIR tasks the training data is ample: most of the available music has genre, composer and performer (artist) information from the record label or folksonomy tags. For some of the tasks, enough (for training a neural network) data has been accumulated through expert annotation (onset detection (Schlüter & Böck, 2014), chord recognition (Humphrey & Bello, 2012)). For emotion-based segmentation task, there is only one dataset, which we introduced in Chapter 5. The dataset consists of 52 songs with 533 boundaries (positive examples of boundaries). However, considering the number of the negative examples (sound excerpts where annotators did not find any boundaries), there is enough data to train a neural network. The boundary annotations do not contain information on the strength of the boundary (such data is actually available in the dataset, but there was not enough agreement in these data for it to be useful for training). This might compromise the performance of the network. Therefore, the next step that we propose is to use additional emotional annotations to

detect the strength of the boundary and filter the weak boundary candidates suggested by the CNN. This step improves the F-measure from 0.53 to 0.61.

This chapter is organized as follows. In section 6.2 we describe related research, mostly on Neural Networks in MIR. In section 6.3 we describe the feature extraction for training the NN (mel-spectrograms). In section 6.4 we describe the boundary retrieval with NN. In section 6.5 we describe a boundary filtering method using Valence and Arousal detection. Section 6.6 concludes the chapter.

6.2 Background

Music data is intricately organized both temporally and structurally. Extracting meaningful features that can describe the organization that is implicit in the signal is the main effort in MIR, and it is probably the most important obstacle that separates us from developing better MER systems, as we argued in chapter 3.

Currently, most of the audio features used in MIR are hand engineered. However, learning features directly from audio is a very attractive new area of research. Below we will describe the attempts in this regard.

The first approach to extract meaningful features from musical audio signal automatically was proposed by Pachet and Zils (2004) who built feature extractors using compositions of basic mathematical and signal processing functions, which were stacked together using genetic programming. Recently, neural networks were used for feature learning from video, speech and music. Neural networks can disentangle interacting factors and create meaningful high-level representations from large amounts of data. A variety of MIR problems were tackled with neural networks. A deep belief network (DBN) consisting of 6 layers of RBMs (Restricted Bolzman Machine) was applied to the chord detection problem (Zhou & Lerch, 2015). A combination of DBN for feature learning and SVM for classification was applied to the genre detection problem (Hamel & Eck, 2010). A convolutional neural network was used to learn a common representation for audio and MIDI features in Hamming space to match huge datasets (Raffel & Ellis, 2015). In (Lee, Largman, Pham, & Ng, 2009), (Wülfing & Riedmiller, 2012), (Nam, Herrera, Slaney, & Smith, 2012), (Dieleman & Schrauwen, 2013), (Yeh, Su, & Yang, 2013), (Sigtia & Dixon, 2014) unsupervised feature learning with deep neural nets has improved performance in phoneme detection, genre classification, auto-tagging and other audio-related tasks. In (Schmidt, Scott, & Kim, 2012), DBN was used to learn features explaining Valence and Arousal dimensions of emotion better than MFCCs. It is rather difficult to interpret the learned features. Indirectly, the relevance can be assessed by interpreting visualizations, or mapping the features to semantic tags. But though a neural network with its layered hierarchical representations of data is much more difficult to analyze, it has undoubtedly pushed the state-of-the-art results further in many fields.

Ullrich, Schlüter, and Grill (2014) and Grill and Schlüter (2015) applied Convolutional Neural network (CNN) to the structural segmentation problem. In (Ullrich et al., 2014), the CNN was trained using mel-spectrograms and structural boundary annotations, advancing the state-of-the-art performance (on the SALAMI dataset) in structural segmentation from 0.33 to 0.46 for tolerances of 0.5 seconds, and from 0.52 to 0.62 for tolerances of 3 seconds. In (Grill & Schlüter, 2015), self-similarity matrices

were added to the CNN training data. The CNN described in this chapter is based on approaches described in (Ullrich et al., 2014), (Grill & Schlüter, 2015) and (Schlüter & Böck, 2014) because our training data is very similar.

6.3 Feature extraction

The most attractive property of the CNN when applied to music is that it removes the necessity to extract audio features. However, though it is theoretically possible to apply it to a raw audio signal, it is normally impractical, and CNNs are applied to a processed form of audio — spectrograms or mel-spectrograms.

We train the network on mel-spectrograms, a perceptually meaningful representation of audio. Mel-scale is a perceptual scale of pitches judged by listeners to be equal in distance from one another. The extraction pipeline is as follows:

1. The magnitude spectrogram is extracted with a window of 2048 samples (this is equivalent to 46 ms for 44.1 kHz sampling rate) using a half overlapping window. The size of the window determines the frequency resolution (the bigger the window, the better the resolution). Better frequency resolution means less noise. With increasing the window size, we are losing the time resolution, but in our case this is not the problem, firstly because our window size is rather small, and secondly we are going to subsample the mel-spectrogram in the time dimension anyway.

2. We apply 50 mel-frequency filters logarithmically spaced from 80 Hz to 5 kHz. We limited the range of frequencies to the lower part of the spectrum, where the fundamental frequencies and first harmonics of most instruments and voice are situated. In this way we lose some timbral information in the upper harmonics, but reduce the size of the spectrogram.

3. The result is scaled logarithmically and each frequency band normalized to zero mean and unit variance. This is done to make it easier for the neural network to process.

4. We reduce the amount of data by subsampling the spectrogram with a window of 12 samples (we take the maximum from each 12 samples), resulting in ≈ 3.5 frames per second.

5. The spectrogram is padded with 16 seconds of pink noise near the beginning and the end of the audio file to allow predicting the boundaries near the beginning and the end. This is done because we are going to predict whether the boundary is exactly in the middle of the segment.

6. The spectrogram of the complete audio file is sliced into 22 second slices. This is a bit larger than the size of the average emotionally stable segment. In this way we are likely to obtain few emotional boundaries per segment.

7. We obtain 45 621 spectrograms (7647 positive examples).

We also tried to extract the spectrograms in a way that worked best for structural segmentation in (Ullrich et al., 2014). There, the mel-spectrograms were extracted in a similar way, but range of frequencies was 9 kHz larger (from 80 Hz to 16 kHz), there were 80 mel filters and the time context was 32 seconds. This setting worked best for boundary retrieval with a tolerance of 3 seconds. However, the mel-spectrograms with bigger resolution did not work for our data. For a small dataset such as ours, big resolution spectrograms probably contain too much information (and noise). For the structural segmentation task, the ground truth datasets are much bigger. Ullrich et al. (2014) used 733 songs for training and 487 for testing. We only have 52 songs in our dataset, this is why we decided to reduce the number of mel filters and only use frequencies in melodic range. We also tried training the network on the features that were used by unsupervised methods in the previous chapter (MFCCs and HPCPs), but it gave worse result than training on mel-spectrograms.

Figure 6.1 shows two different resolutions for the same timestamp of a song.

6.4 Boundary detection with the CNN

6.4.1 Annotations

We use the averaged annotations described in the previous chapter (533 emotional boundaries) as the ground truth. At least two of the three annotators agreed on each boundary, and the average time stamp was adopted as the ground truth boundary. The annotations are less precise than annotations for structural segmentation. Following (Ullrich et al., 2014) and (Grill & Schlüter, 2015), we use a window of 6 seconds to indicate the boundary. Everything 3 seconds to the right and to the left of the ground truth timestamp is considered a boundary. Frames to the right and to the left of the boundary receive partial weight, which is distributed using a Gaussian window (from 0 to 1). Figure 6.1 shows the input mel spectrograms with ground truth annotations (distributed using Gaussian window) on top of the boundary. The negative examples (the mel-spectrograms that do not have a boundary in the middle or in the 6 second neighbourhood from the middle) are annotated with a zero.

6.4.2 CNN

Convolutional Neural Network (CNN) is a powerful algorithm developed for image classification. CNNs were inspired by the studies of visual cortex of cats (Hubel & Wiesel, 1962) and monkeys (Hubel & Wiesel, 1968). This cortex contains a complex layered arrangement of cells, sensitive to small sub-regions of the visual field. The receptive fields of the cells are tiled to cover the entire visual field. These cells act as local filters, and are well-suited to exploit the strong local spatial correlations present in natural images. There are specific cell types to detect edge-like patterns, and another type with a larger receptive field, locally invariant to the exact position of the pattern.

CNNs are a variation of feed-forward artificial neural network (ANN), trained with back-propagation on the same principles as other ANNs. The key feature of the CNN are its convolutional layers, which contain convolution filters designed similarly to the biological neurons in the visual cortex, described above. The CNN network typically

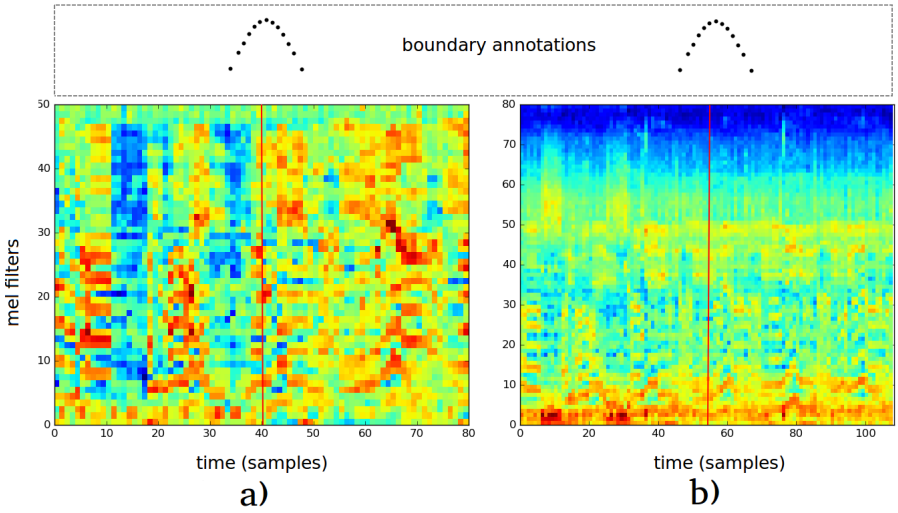


Figure 6.1: Input to the neural network: a frame centered on a boundary (red line) in a song RWC-MDB-G-2001-M08-03 (ID format of the RWC dataset (Goto et al., 2002)). a) Mel-spectrograms with 22 second time resolution (the one that gave a better result). b) Mel-spectrogram of 32 seconds.

consists of one or several convolutional and subsampling layers followed by one or several dense layers.

The input to a convolutional layer is a $m \times n \times r$ matrix (image), where m is the height, n is the width and r is the number of channels. In an RGB image, typically, $r = 3$. For magnitude spectrograms, there is only one channel (magnitude of a certain mel-frequency range at a certain time). The convolutional layer has k filters, the size of the filters is normally several times smaller than the width and height of the input image, and the depth can be the same or smaller than the number of channels. The filter is convolved with the image to produce feature maps, which are then subsampled over $n \times k$ regions. Then, an activation function (nonlinearity) is applied to each feature map. We use ReLU (Rectified Linear Unit) activation function.

The key property of the CNN is that it allows to process large data inputs (images) with few trainable parameters, and CNN allows to preserve the spatial layout of the input.

The architecture of our neural network is similar to (Ullrich et al., 2014) and is sketched on Figure 6.2. The first convolutional layer has 32 8×6 kernels, followed by a subsampling layer with max-pooling (3×6). The next convolutional layer has 64 6×3 layers, followed by another max-pooling layer. We add weight regularization in the convolutional layers and a dropout after the last max-pooling layer. The amount of regularization is set experimentally, we increase the regularization until the network does not overfit to the training set anymore. We use L2 regularization in both convolutional layers. This form of regularization penalizes the squared magnitude of all parameters directly in the objective. For each weight w we add a term $\frac{1}{2}\lambda w^2$ to the objective, where λ is regularization strength (0.01 in our case). The L2 regularization

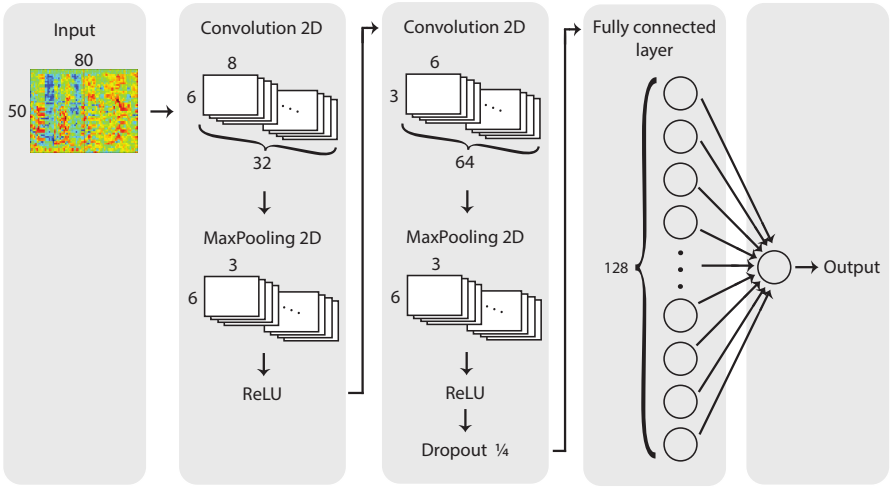


Figure 6.2: CNN architecture.

penalizes peaky weight vectors and makes network prefer diffuse weight vectors. This encourages the network to use all of its inputs instead of using some a lot. Lastly, the network has a dense 128 neuron layer and a 1 neuron output.

6.4.3 Peak picking

For each input sample, the network outputs a real number between 0 and 1: a score indicating the likelihood of the boundary being present in the sample. To find the peaks in this curve, we pick a maximum inside a moving window of 8 seconds, and mark it as a boundary if it is higher than a moving threshold. The threshold is necessary to filter the local peaks which are actually very low, but happen to be higher than their local context. The threshold is computed with the following formula:

$$T_i = x_i + \frac{2}{3}\overline{x}, \tag{6.1}$$

where T_i is a threshold for sample i, x_i is the output of the 10 second wide median filter, and $\overline{x}$ is the average of all the annotations.

Figure 6.3 shows the probability curve for the boundary shown on Figure 6.1 and some neighbourhood around that boundary.

6.4.4 Results

We achieved the best results on mel-spectrograms of size 50×80 with melodic range mel-filters which are described in section 6.3. The 80×108 spectrograms gave almost random level performance. This probably happened due to network not training because of high levels of noise-to-data ratio.

We don't have enough data to have a separate test set and therefore we use 10-fold cross validation. We divide our data into three sets — training, validation and

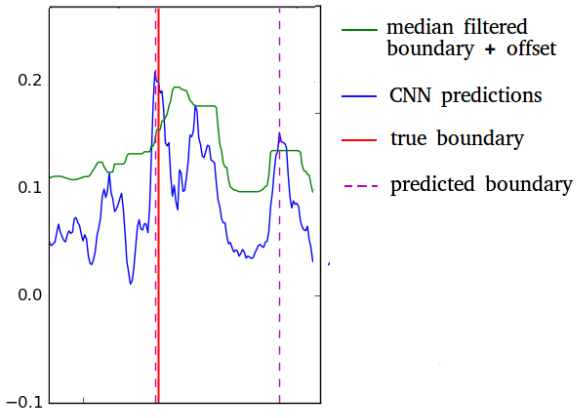


Figure 6.3: Peak picking.

test. A network is trained on the samples from 47 songs, the validation set is 10% of these data. After training the network predicts on the leave-out test set (5 songs). The process is repeated for the next 5 songs.

The network is trained using stochastic gradient descent with Mean Squared Error as a loss metric, in batches of 64 samples, for up to 100 epochs (early stopping is used, if the loss metric is not decreasing for 15 epochs on the validation set). The network is implemented in Theano[1].

We trained the network on the annotations where at least 2 annotators agreed on each boundary. Table 6.1 shows the evaluation on these annotations and also on all available annotations (including also those where the boundary was indicated only by 1 annotator). There are 31% more boundaries in the second case, on average 6.8 more boundaries per song. We also compare the predictions of the neural network with a random baseline, where we generate segment lengths with a Gaussian distribution centered on the average segment size (19 seconds).

The CNN can recall 73% of the boundaries, but the precision is unsatisfactory. When we evaluate on all the annotations, the F-measure does not increase, which means that the errors that the CNN makes (finding too many boundaries) are not explained by the weak boundaries that were indicated only by one of the annotators. The predictions are much better than random. The number of the retrieved boundaries (and the trade-off between precision and recall) can be changed by changing the parameters of the peak picking function. F-measure, however, is already at its highest.

6.5 Emotion change strength detection

The boundary predictions from CNN suffer from low precision: only half of the retrieved boundaries are actually relevant. A small percentage (5% more) can be explained by a presence of weak emotional boundary indicated only by one of the an-

[1] deeplearning.net/software/theano

notators, but 45% of the boundaries are excessive. This could be because the neural network learned to recognize some structures that often coincide with emotional boundary but do not indicate a change in emotion. The annotations that we use as the ground truth for training a neural network do not contain any information about the nature or the strength of the emotional boundaries. The network thus cannot know which emotion changed, and by how much. By supplementing this information, it could be possible to filter the boundary candidates suggested by the network and improve precision. In order to do that, we are going to apply a MER algorithm on the excerpts before and after every boundary, to detect whether there is a measurable change in emotion.

We will use a dataset of the song excerpts annotated with Valence and Arousal assembled by Witteveen (2015) from three sources (MediaEval Database for Emotional Analysis in Music, Soundtrack dataset and Saari dataset), as the training data. In total, there are 2700 sound excerpts in various music genres from 15 to 45 seconds long. The details about these data are in Table 6.2. The **DEAM** was already extensively described in Chapter 4. We use only the small 45 second excerpts and the static (per song) annotations. Of course, this dataset is controversial when applied to static emotion detection, because this music was supposed to contain emotional variation, but removing these data from the training set degraded the performance of the system, so we decided to retain the complete dataset. The **Saari** dataset was assembled by Saari and Eerola (2013) and consists of 596 music pieces in various popular music genres with social tags from last.fm and separately collected Valence and Arousal annotations. The **Soundtrack** dataset is described in (Eerola & Vuoskoski, 2011).

6.5.1 Features

We extract both low (MFCC, chroma, energy, dissonance and other spectral features) and high-level (scale, tempo, tonal stability) audio features using Essentia (Bogdanov et al., 2013). All the music files are converted to mono. To extract low-level timbral features we use a half-overlapping window of 100 ms, and a window of 3 seconds for high level features. The features are normalized to zero mean and unit variance.

6.5.2 MER training and evaluation

We train Support Vector Regression with Radial Basis Function kernel and grid-search optimized parameters (C and γ). Using 10-fold cross-validation on the training set (the three datasets described in Table 6.2), we obtain $R^2 = 0.61$, RMSE = 0.94 (the range

Algorithm	Annotations	Precision @ 3	Recall @ 3	F-measure @ 3
CNN	agreement of 2	.50	.73	.56
	all annotations	.55	.55	.53
Random	agreement of 2	.37	.40	.37
	all annotations	.41	.31	.35

Table 6.1: Performance of the CNN and a random baseline on emotional segmentation.

Dataset	Number of songs	Clip length	Number of annotators per song
DEAM	1744	45 s	10
Saari	596	15–30 s	29
Soundtrack	360	10–30 s	12

Table 6.2: Datasets used for MER.

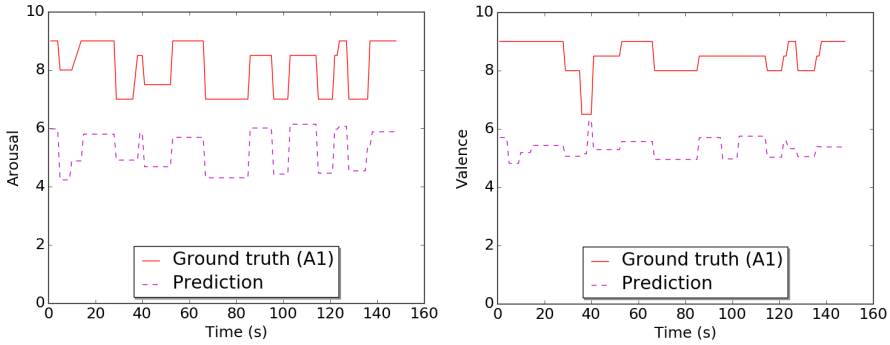


Figure 6.4: Predicting emotion of the segments.

of the annotations is between 0 and 10), Concordance Correlation Coefficient (CCC) = 0.77 for arousal and R^2 = 0.29, RMSE = 1.06, CCC = 0.44 for valence.

We evaluate the model on the Valence and Arousal annotations that accompany the emotional segmentation. Table 6.3 shows the averaged results of such evaluation for each annotator separately. We use the segment boundary annotations and predict Arousal and Valence values for each segment. We ignore that some of the segments' emotion is not stable and predict a stable emotion for those segments as well. Transitional segments are usually quite short, and few, so this should not influence the estimation of performance of the system anyhow significantly. Figure 6.4 illustrates this for one of the songs.

Metric	Emotion	A_1	A_2	A_3
R^2	arousal	0.36	0.26	0.35
	valence	0.05	0.04	0.05
RMSE	arousal	2.02	1.26	2.29
	valence	1.92	0.98	2.03
CCC	arousal	0.16	0.22	0.14
	valence	0.03	0.02	0.03

Table 6.3: Evaluation on the 52 songs. A_i — annotator i.

The results of prediction on the test data are satisfactory for arousal, but still need a lot of improvement for valence. However, here we are not interested in absolute values of valence and arousal, or in the direction of change from one segment to another. We

are only interested in a system to estimate the magnitude of that change. We can not evaluate whether the system estimates this magnitude correctly, because ground truth annotations of the magnitude have unsatisfactory consistency, as we saw in Section 5.4.3 of Chapter 5. However, indirectly we will evaluate the magnitude estimation by evaluating the method in the next section.

6.5.3 Boundary strength

To filter the boundaries predicted by the CNN, we need to predict how strong is the difference between Valence and Arousal before and after the boundary. However, CNN did not find 27% of the ground truth boundaries, which means we cannot be sure that no boundary occurs within any given predicted segment. Therefore we are going to predict Valence and Arousal in the near neighbourhood of the boundary and not in the segments before and after. Experimentally we set this neighbourhood to 8 seconds (or smaller if a boundary is predicted closer). The trained SVR model predicts Valence and Arousal for this local neighbourhood, and we retain a boundary if the difference in Valence or Arousal is bigger than a threshold (which is set to 0.4). The threshold is small, because the predictions display a strong regression to the mean. The retrieval results after filtering are as follows: Precision @ 3 = 0.69, Recall @ 3 = 0.58 and F-measure @ 3 = 0.61. The F-measure is therefore improved from 0.53 to 0.61. Applying the method to Structure Features segmentation actually degrades the performance. Random segmentation benefits slightly from increase in precision obtained by removal of extra boundaries, but of course still it is the worst method because the recall is still as low as it was.

6.6 Conclusion

In this chapter we proposed a doubly-informed system to predict emotional boundaries, achieving a slight improvement over the best unsupervised approach. The difference was not statistically significant. A Convolutional Neural Network was trained on only 52 songs with 533 annotated boundaries. The size of the mel-spectrograms had to be kept small for the network to be able to train, probably due to a small size of the dataset. Mel-spectrograms of such size contain very reduced amount of information, mostly changes in loudness in different mel bands, which indicate timbral changes. There is very limited harmonic information. Probably, the cues that the network was using for emotional boundary detection were very similar to the hand engineered discontinuity detection in unsupervised methods. However, the CNN approach has potential to learn more complicated dependencies, should more data be available for training. Then, more complex cues could be used by the network.

To supplement the missing information on emotional changes, we applied a MER method in a local neighbourhood of the boundary. The MER method we developed suffered from a typical problem of inability to reliably predict valence. Also, the performance on the test data (52 songs) was much worse than cross validated performance estimation on the train data. The test data excerpts were shorter, but mostly in pop and rock music style, which is the same as one of the training datasets (**Saari**). The problem with transferability of the models trained on one dataset and tested on different data was also described by (Witteveen, 2015). This problem can be caused both by

differences in music (signal processing and features problem) and differences in understanding of the Valence and Arousal dimensions by the annotators. Despite these problems, we could improve the F-measure from 0.53 to 0.61.

CHAPTER 7

Conclusion

Affective content analysis and emotion recognition are new, but rapidly growing fields, as can be seen by the proliferation of research and books published on these topics. On Figure 7.1 we see a search in Google Ngram Viewer on the keywords related to affective research. Google Ngram Viewer only has a corpus that extends until 2008, but the topic has probably become even more relevant since then, continuing the trend.

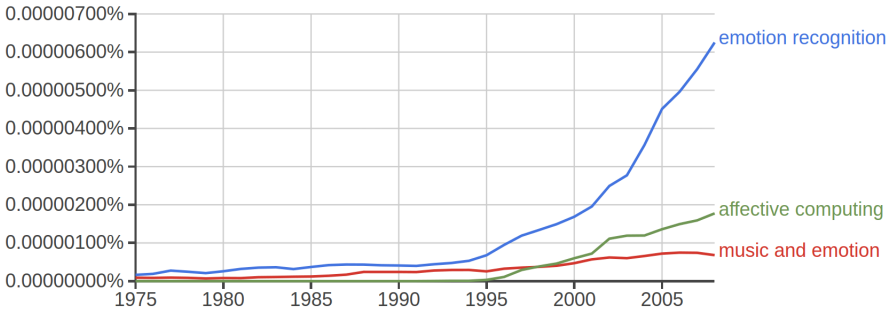


Figure 7.1: Frequencies of occurrence in online text corpora (5.2 million books) of terms related to this thesis, as displayed by Google Ngram Viewer.

In this thesis we dealt with a variety of tasks that were directed towards developing better Music Emotion Recognition algorithms. We handled three topics — modeling emotion induced by music, benchmarking MEVD (music emotion variation detection) algorithms, and emotion-based music segmentation. Each of the topics involved collecting a dataset (in total, as a result three public datasets were released):

1. The **Emotify** dataset is the first publicly available dataset of music annotated with induced emotion. Also, this is the biggest public dataset annotated using GEMS model.

2. The **DEAM** dataset has the largest number of continuous emotion annotations.

3. The **Emotional segmentation** dataset is unique in a sense that this is the first dataset for emotional segmentation task.

In the following we will summarize some general findings of this thesis, and suggest directions for future work.

Representations of musical emotion

Every MIR task — genre recognition, chord detection, structural segmentation — depends on human annotated data, and has to deal with the problem of subjectivity and cue application inconsistency in such data. But for music emotion recognition the situation is particularly bad. In fact, it is so bad that statistical measurement devices regularly "give up" on measuring agreement on emotional annotations. In this thesis we witnessed this two times, in Chapter 3, when 33 out of 400 songs had "out of bounds" negative Fleiss' kappa on GEMS ratings, and in Chapter 4 with a similar situation with negative Cronbach's α on continuous emotional annotations. However, the situation is definitely not hopeless. It is possible to obtain sufficiently consistent annotations of musical emotion given the right representation and sensible task demands.

From the data collected using the **Emotify** game we learned that:

1. Certain emotions are more universally understood in relation to music. Examples are joyful activation, calmness and power. Certain other emotions have much more variability in their interpretation. Examples are amazement, sadness, nostalgia.

2. The consistency of the annotations is influenced by whether music is liked or not.

3. The mood of the listener is an important factor in emotional response, other factors being his musical taste, and, to a much smaller extent, age.

From the data collected for the **Emotion in Music MediaEval Benchmark** we learned that:

1. Asking for the continuous ratings of musical emotion puts unreasonable task demands on the annotators, such as the necessity to rate on absolute scale, to evaluate response very frequently, and to react to changes fast.

2. Annotators require an initial orientation time of around 13 seconds.

3. Annotators tend to choose an arbitrary "zoom level" (sections, phrases, single notes).

From the dataset collected for **emotional segmentation** we learned that:

1. Listeners can agree on placement of emotional boundaries very well.
2. They cannot, however, agree on the magnitude of the change of emotion before and after the boundary with the same consistency.

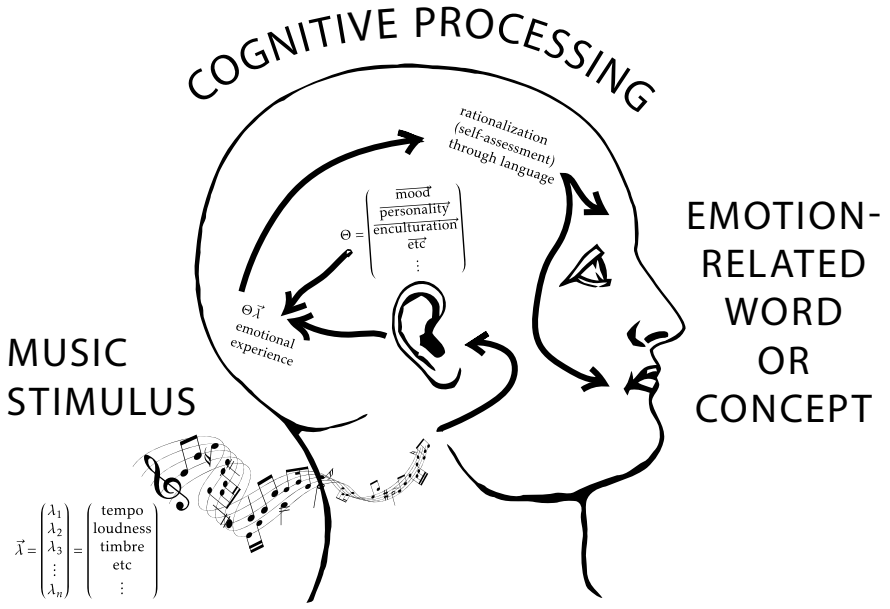


Figure 7.2: Cognitive processing of emotional stimulus. Illustration: human head is based on a drawing by F.E. Bilz (1894).

In the Introduction we said that the methodological basis of computational MER is similar to the Brunswick's lens model (that describes how listener matches templates with parameters β to identify emotion in music). We would like to introduce two more steps to this model, one step before/during template matching, and one step after that. First, if we are interested in induced emotion, factors related to listener's personality and situational factors are also important. Figure 7.2 shows how musical stimuli are processed with our additions. Music with parameters λ is heard by the listener. The listener has a transformation matrix Θ of parameters related to his cultural background, personality and current mood. This transformation matrix is applied to change both the influence of objective parameters λ, and also the emotion templates. For instance, if the listener tends to focus on beat more than on other aspects, the importance of beat will be increased. The next important addition is that the subjective emotional experience has to be translated to words (for instance, to submit a query to a MER system, or to annotate a song). During this verbalization stage the emotional experience is assessed in some way and the output is averaged to the nearest known emotional label. Both steps have some inconsistency of cue application.

Making personalized music predictions is therefore not only learning which templates the listener is using for certain emotions, but also taking into account the per-

sonality and mood-related factors that influence emotion induction. In the last stage of our music cognition process listeners map their internal sensations to words. Probably, some people are much better at it than others, and the internal sensations that different people call ‘tranquility’ or ‘tenderness’ can be quite different. It is important to realize that another source of inconsistency lies in these distinctions.

Audio features

In Chapter 3 we investigated which musical properties we need to account for to be able to predict induced emotion, and found that some of the strongest predictors are melodiousness, articulation, tonaleness and other descriptors based on human perception of tonality, harmony, rhythm. Extracting these sort of properties from musical signal is going to be very non-trivial. We proposed some interval and chord features, which could improve the performance of our comprehensive feature set. However, there still remained a big gap in performance between spectral features we are currently able to extract, and perceptual features which we would like to aim for. In terms of r, this gap was on average 0.15. With perceptual features, despite subjectivity (which is especially strong for induced emotion) it was still possible to achieve good predictions using meaningful perceptual features.

There is much less attention paid to predicting induced emotion than it deserves. Induced music emotion recognition system is probably more useful for the majority of music listeners than the one that can predict perceived emotion. Hopefully, the dataset that we released and findings from this thesis will help future research in this area.

In Chapter 4 we evaluated feature sets and algorithms for MEVD through benchmarking. Taking larger temporal context into account proved to be very important.

Music emotion variation detection

Through benchmarking MEVD methods we gradually arrived at a conclusion that current mainstream approach to tracking emotion over time takes both annotation and features to unreasonably high time resolution (less than a second) (though this is not what was meant by design). This results in predicting dynamic changes in loudness and timbre, which might be actually means of expressing emotion on a higher level rather than emotion changes themselves. We propose to approach MEVD by segmenting music into segments of stable and unstable emotion. Then, we propose the first supervised method for emotional boundary detection.

Song-level MER methods normally need to make sure that the excerpts for which they are predicting an emotion are emotionally homogeneous, at least at the ground truth collection stage. Emotional boundary detection method that we suggested would help to do this.

Future work

To improve induced emotion modeling, we suggest that using physiological signals, EEG or brain imaging techniques to augment self-assessment could give an additional

verification step to ground truth collection, and should be used more in MIR research. This will complement the emotion self-assessment stage (last stage in our music processing model on Figure 7.2), which is difficult for many people and can lead to more inconsistency.

An emotional segmentation task has been introduced in this thesis. This task is very novel and needs more research. The data used in Chapters 5 and 6 only contained popular music. This genre usually has a fixed repetitive structure. We don't know how would emotional segmentation work on through-composed pieces. Also, we only addressed emotionally stable excerpts and not emotionally unstable ones. Probably, the methods for detecting short-time dynamic changes such as described in Chapter 4 could find application there.

Another obvious direction for future work is cognitively motivated audio feature development. Most of the audio features so far developed by the MIR community are computed from very short parts of the spectrum, and do not take into account human cognitive music processing, musical expectation, structure. This approach is of course inadequate and can not yield satisfactory results. Wiggins (2009) stresses this point in his work where he maintains that any MIR research not based on human cognition is pointless:

> ...any system that deals with Music effectively is de facto a cognitive model (even if a "black box"), because Music is fundamentally cognitive; and by the same token, only cognitive models are likely to succeed in processing Music in a human-like way. To treat Music in a way which is not human-like is meaningless, because Music is defined by humans.

When designing new features manually, development of these cognitive audio features is a very complicated task. However, with the emergence of such powerful models capable of learning the temporal context as LSTM-RNN (Long–Short Term Memory Recurrent Neural Networks) or CNN (convolutional neural network) that can learn spatial organization of sound as presented on spectrogram, or perhaps a combination of the two, and with more training data becoming available, we can hope that processing music in a more meaningful way will be feasible. In fact, this is the direction of research that the author of this thesis is going to pursue.

MER field is still very young and there are many open questions left, and, hopefully, results, datasets and findings from this thesis will help to advance the field further.

Bibliography

Ahn, L. von, & Dabbish, L. (2004). Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 319–3265).

Aljanaki, A., Soleymani, M., & Yang, Y.-H. (2014). Emotion in Music Task at MediaEval 2014. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*

Aljanaki, A., Wiering, F., & Veltkamp, R. (2015). MediaEval 2015: A Segmentation-based Approach to Continuous Emotion Tracking. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Aljanaki, A., Wiering, F., & Veltkamp, R. (2016). Studying emotion induced by music through a crowdsourcing game. *Information Processing and Management*, *52*(1), 115–128.

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2013). MIRUtrecht Participation in MediaEval 2013: Emotion in Music Task. In *Working Notes Proceedings of the MediaEval 2013 Workshop.*

Aljanaki, A., Wiering, F., & Veltkamp, R. C. (2014). Computational modeling of induced emotion using GEMS. In *Proceedings of the 15th International Society for Music Information Retrieval Conference* (pp. 373–378).

Asmus, E. P. (2009). The measurement of musical expression. In *Suncoast Music Education Research Symposium.*

Aucouturier, J.-J., & Bigand, E. (2012). Mel Cepstrum and Ann Ova: The Difficult Dialog Between MIR and Music Cognition. In *Proceedings of the 13th International Society for Music Information Retrieval Conference* (pp. 397–402).

Bachorik, J. P., Bangert, M., Loui, P., Larke, K., Berger, J., Rowe, R., et al. (2009). Emotion in motion: Investigating the time-course of emotional judgments of musical stimuli. *Music Perception*, *26*(4), 355–364.

Backlund, P., Engstrom, H., Hammar, C., Johannesson, M., & Lebram, M. (2007). Sidh – a Game Based Firefighter Training Simulation. In *11th International Conference on Information Visualization* (pp. 899–907).

Balen, J. M. H. van. (2016). Audio Description and corpus analysis of popular music (Doctoral dissertation, Utrecht University). *CPI Koninklijke Wöhrmann*.

Balkwill, L.-L., Thompson, W., & Matsunaga, R. (2004). Recognition of emotion in Japanese, North Indian and Western music by Japanese listeners. *Japanese Journal of Psychological Research*, *46*, 337–349.

Balkwill, L.-L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception*, *17*, 43–64.

Baltes, F. R., Avram, J., Miclea, M., & Miu, A. C. (2011). Emotions induced by operatic music: Psychophysiological effects of music, plot, and acting: A scientist's tribute to Maria Callas. *Brain and Cognition*, *76*(1), 146–157.

Barrington, L., O'Malley, D., Turnbull, D., & Lanckriet, G. (2009). User-centered design of a social game to tag music. In *Proceedings of the ACM SIGKDD Workshop on Human Computation* (pp. 7–10).

Barthet, M., Fazekas, G., & Sandler, M. (2012). Multidisciplinary Perspectives on Music Emotion Recognition: Implications for Content and Context-Based Models. In *Int'l Symp. Computer Music Modelling & Retrieval* (pp. 492–507).

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In *Proceedings of the 12th International Society for Music Information Retrieval Conference* (pp. 591–596).

Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., & Dacquet, A. (2005). Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. *Cognition and Emotion*, *19*, 1113–11139.

Bittner, R., Salamon, J., Tierney, M., Mauch, M., Cannam, C., & Bello, J. P. (2014). MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research. In *Proceedings of the International Society for Music Information Retrieval Conference.*

Bogdanov, D., Wack, N., Gomez, E., Gulati, S., Herrera, P., & Mayor, O. (2013). ESSENTIA: an Audio Analysis Library for Music Information Retrieval. In *International Society for Music Information Retrieval Conference* (pp. 493–498).

Cabrera, D. (1999). PsySound: A computer program for psychoacoustical analysis. *Proc. Australian Acoustical Society Conf.*, 47–54.

Cai, K., Yang, W., Cheng, Y., Yang, D., & Chen, X. (2015). PKU-AIPL Solution for MediaEval 2015 Emotion in Music Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Cantor, J. R., & Zillmann, D. (1973). The effect of affective state and emotional arousal on music appreciation. *The Journal of General Psychology*, *89*(1), 97–108.

Chen, Y.-A., Wang, J.-C., Yang, Y.-H., & Chen, H. H. (2015). The AMG1608 dataset for music emotion recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 693–697).

Cheng, H.-T., Yang, Y.-H., Lin, Y.-C., Liao, I.-B., & Chen, H. H. (2008). Automatic chord recognition for music classification and retrieval. In *IEEE International Conference on Multimedia and Expo* (pp. 1505–1508).

Chmulik, M., Guoth, I., Malik, M., & Jarina, R. (2015). UNIZA System for the "Emotion in Music" task at MediaEval 2015. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Choppina, S., Trost, W., Dondainea, T., Milleta, B., Drapiera, D., Verina, M., et al. (2016). Alteration of complex negative emotions induced by music in euthymic patients with bipolar disorder. *Journal of Affective Disorders*, *191*, 15–23.

Collier, G. L. (2007). Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, *35*(1), 110–131.

Costa, M., Fine, P., & Bitti, P. E. R. (2004). Interval Distributions, Mode, and Tonal Strength of Melodies as Predictors of Perceived Emotion. *Music Perception*, *22*(1), 1–14.

Coutinho, E., & Cangelosi, A. (2011). Musical emotions: Predicting second-by-second subjective feelings of emotion from low-level psychoacoustic features and physiological measurements. *Emotion*, *11*(4), 921–937.

Coutinho, E., & Scherer, K. (2012). Towards a brief domain-specific self-report scale for the rapid assessment of musically induced emotions. In *12th International Conference of Music Perception and Cognition (ICMPC12).*

Coutinho, E., Trigeorgis, G., Zafeiriou, S., & Schuller, B. (2015). Automatically Estimating Emotion in Music with Deep Long-Short Term Memory Recurrent Neural Networks. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Coutinho, E., Weninger, F., Schuller, B., & Scherer, K. R. (2014). The Munich LSTM-RNN Approach to the MediaEval 2014 Emotion in Music Task. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334.

Deng, J., & Leung, C. (2015). Dynamic Time Warping for Music Retrieval Using Time Series Modeling of Musical Emotions. *IEEE Transactions on Affective Computing*, *PP*(99).

Deponti, D., Maggiorini, D., & Palazzi, C. (2009). DroidGlove: An android-based application for wrist rehabilitation. In *International Conference on Ultra Modern Telecommunications Workshops* (pp. 1–7).

Dibben, N. J. (2004). The role of peripheral feedback in emotional experience with music. *Music Perception*, *22*(1), 79–115.

Dieleman, S., & Schrauwen, B. (2013). Multiscale approaches to music audio feature learning. In *Proceedings of the 14th International Society for Music Information Retrieval.*

Eerola, T. (2011). Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of Datasets Spanning Classical, Film, Pop and Mixed Genres. *Journal of New Music Research*, *40*(4), 349–366.

Eerola, T. (2014). Modelling Emotions in Music: Advances in Conceptual, Contextual and Validity Issues. In *AES International Conference.*

Eerola, T., Friberg, A., & Bresin, R. (2013). Emotional expression in music: contribution, linearity, and additivity of primary musical cues. *Frontiers in Psychology*, *4*(487).

Eerola, T., & Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychology of Music*, *39*(1), 18–49.

Ekman, P. (2005). Basic Emotions. In *Handbook of Cognition and Emotion* (pp. 45–60). John Wiley & Sons, Ltd.

Elfenbein, H., & Ambady, N. (2002). On the universality and cultural specificity of emotion recognition: A meta-analysis. *Psychological Bulletin*, *128*, 203–235.

Eyben, F., Weninger, F., Gross, F., & Schuller, B. (2013). Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Pro-*

ceedings of the 21st ACM International Conference on Multimedia (pp. 835–838). New York, NY, USA: ACM.

Fan, Y., & Xu, M. (2014). MediaEval 2014: THU-HCSIL Approach to Emotion in Music Task using Multi-level Regression. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*

Farnsworth, P. (1958). *The Social Psychology of Music*. The Dryden Press.

Fitzgerald, D. (2010). Harmonic/percussive separation using median filtering. In *13th International Conference on Digital Audio Effects.*

Fleiss, J. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 378–382.

Fontaine, J. R. J., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The World of Emotions is not Two-Dimensional. *Psychological Science*, *18*(12), 1050–1057.

Foote, J. (2000). Automatic Audio Segmentation Using a Measure Of Audio Novelty. In *Proceedings of the IEEE International Conference of Multimedia and Expo* (pp. 452–455).

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., et al. (2009). Universal recognition of three basic emotions in music. *Current Biology*, *19*(7), 573–576.

Futrelle, J., & Downie, J. S. (2002). Interdisciplinary communities and research issues in music information retrieval. In *Proceedings of the 3rs International Society for Music Information Retrieval Conference* (pp. 215–221).

Gabrielsson, A. (1973). Adjective ratings and dimension analysis of auditory rhythm patterns. *Scandinavian Journal of Psychology*(14), 244–260.

Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different? *Musicae scientiae*, *5*(1), 123–147.

Gabrielsson, A. (2011). *Strong Experiences with Music: Music is much more than just music*. Oxford University Press.

Gabrielsson, A., & Juslin, P. (2002). Handbook of Affective Sciences. In (chap. Emotional expression in Music). Oxford University Press.

Gabrielsson, A., & Lindström, E. (2011). Handbook of Music and Emotion: Theory, Research, Applications. In (chap. The role of structure in the musical expression of emotions). Oxford University Press.

Gaggi, O., Galiazzo, G., Palazzi, C., Facoetti, A., & Franceschini, S. (2012). A Serious Game for Predicting the Risk of Developmental Dyslexia in Pre-Readers Children. In *21st International Conference on Computer Communications and Networks* (pp. 1–5).

Geringer, J. M., Madsen, C. K., & Gregory, D. (2004). A Fifteen-Year History of the Continuous Response Digital Interface: Issues Relating to Validity and Reliability. *Bulletin of the Council for Research in Music Education*, *160*, 1–15.

Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R. (2002). RWC Music Database: Popular, Classical, and Jazz Music Databases. In *Proceedings of the 3rd International Conference on Music Information Retrieval* (pp. 287–288).

Gregory, D. (1989). Using computers to measure continuous music responses. *Psychomusicology: A Journal of Research in Music Cognition*, *8*(2), 127–134.

Grill, T., & Schlüter, J. (2015). Music boundary detection using neural networks on combined features and two-level annotations. In *Proceedings of the 16th*

International Society for Music Information Retrieval.

Guan, D., Chen, X., & Yang, D. (2012). Music Emotion Regression Based on Multimodal Features. In *Symposium on Computer Music Multidisciplinary Research* (pp. 70–77).

Gupta, R., & Narayanan, S. (2015). Predicting Affect in Music Using Regression Methods on Low Level Features. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Hamel, P., & Eck, D. (2010). Learning features from music audio with Deep Belief networks. In *Proceedings of the 9th International Society for Music Information Retrieval Conference.*

Harte, C. A., & Sandler, M. B. (2006). Detecting harmonic change in musical audio. In *Proceedings of Audio and Music Computing for Multimedia Workshop.*

Hevner, K. (1936). Experimental studies of the elements of expression in music. *American Journal of Psychology*, *48*, 246–268.

Hodges, D. A. (2011). Handbook of Music and Emotion: Theory, Research, Applications. In (chap. Psycho-physiological measures). Oxford University Press.

Hu, X., & Downie, J. (2010a). Improving mood classification in music digital libraries by combining lyrics and audio. In *Proceedings of the Joint Conference on Digital Libraries.*

Hu, X., & Downie, J. (2010b). When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of the 11th international society for music information retrieval conference.*

Hu, X., Downie, J. S., Laurier, C., Bay, M., & Ehmann, A. F. (2008). The 2007 MIREX Audio Mood Classification Task: Lessons Learned. In *Proceedings of International Society for Music Information Retrieval* (pp. 462–467).

Hubel, D., & Wiesel, T. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. *Journal of Physiology*, *195*, 215–243.

Humphrey, E. J., & Bello, J. (2012). Rethinking Automatic Chord Recognition with Convolutional Neural Networks. In *2012 11th International Conference on Machine Learning and Applications* (Vol. 2, pp. 357–362).

Hunter, P. G., Schellenberg, E. G., & Schimmack, U. (2008). Mixed affective responses to music with conflicting cues. *Cognition & Emotion*, *22*(2), 327–352.

Huq, A., Bello, J. P., & Rowe, R. (2010). Automated Music Emotion Recognition: A Systematic Evaluation. *Journal of New Music Research*, *39*(3), 227–244.

Ilie, G., & Thompson, W. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, *23*(4), 319–329.

Imbrasaite, V., Baltrušaitis, T., & Robinson, P. (2013). Emotion tracking in music using continuous conditional random fields and relative feature representation. In *2013 IEEE International Conference on Multimedia and Expo Workshops* (pp. 1–6).

Imbrasaite, V., Baltrušaitis, T., & Robinson, P. (2014). CCNF for Continuous Emotion Tracking in Music: Comparison with CCRF and Relative Feature Representation. In *Multimedia Affective Computing, IEEE International Conference on Multimedia and Expo.*

Inskip, C., Macfarlane, A., & Rafferty, P. (2012). Towards the disintermediation of creative music search: analysing queries to determine important facets. *International Journal on Digital Libraries*, *12*(2), 137–147.

Jaimovich, J. (2013). Emotion Recognition from Physiological Indicators for Musical Applications (Doctoral dissertation, Queen's University Belfast). *Queen's University Belfast*.

Jaimovich, J., Coghlan, N., & Knapp, R. (2012). Emotion in Motion: A Study of Music and Affective Response. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval* (pp. 29–44).

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance*, *26*, 1797–1813.

Juslin, P. N. (2013, Sep). From everyday emotions to aesthetic emotions: Towards a unified theory of musical emotions. *Physics of Life Reviews*, *10*(3), 235–266.

Juslin, P. N., & Laukka, P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, *33*(3), 217–238.

Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: Listener, music, and situation. *Emotion*, *8*(5), 668–683.

Juslin, P. N., & Lindström, E. (2010). Musical expression of emotions: modeling listeners' judgements of composed and performed features. *Music Analysis*, *29*, 334–364.

Juslin, P. N., & Sloboda, J. (2011). *Handbook of Music and Emotion: Theory, Research, Applications*. Oxford University Press.

Juslin, P. N., & Västfjäll, D. (2008). Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and Brain Sciences*, *31*, 559–575.

Kaelen, M., Barrett, F., Roseman, L., Lorenz, R., Family, N., Bolstridge, M., et al. (2015). LSD enhances the emotional response to music. *Psychopharmacology*, *232*(19), 3607–3614.

Kim, Y. E., Schmidt, E., & Emelle, L. (2008). Moodswings: A collaborative game for music mood label collection. In *Proceedings of the 9th International Society for Music Information Retrieval Conference* (pp. 231–236).

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., et al. (2010). Music Emotion Recognition: A State of the Art Review. In *Proceedings of the 11th International Society for Music Information Retrieval Conference.*

Kivy, P. (1990). *Music alone: Reflections on a purely musical experience*. Cornell University Press.

Kivy, P. (1993). *The Fine Art of Repetition: Essays in the Philosophy of Music*. Cambridge University Press.

Kleinen, G. (1968). Experimentelle Studien zum musikalischen Ausdruck [Experimental studies on musical expression] (Doctoral dissertation, Universität Hamburg). *Hamburg, Germany*.

Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., et al. (2012). DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computin*, *3*(1), 18–31.

Konečni, V. (2008). Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts*, *2*(2), 115–129.

Korhonen, M., Clausi, D., & Jernigan, M. (2006). Modeling emotional content of music using system identification. *IEEE Transactions on Systems, Man, and Cybernetics*, *36*(3), 588–599.

Krumhansl, C. L. (1997). An exploratory study of musical emotions and psychophysiology. *Canadian Journal of Experimental Psychology*, *51*(4), 336–52.

Kumar, N., Gupta, R., Guha, T., Vaz, C., Segbroeck, M. van, Kim, J., et al. (2014). Affective Feature Design and predicting continuous affective dimensions from music. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*

Lahdelma, I., & Eerola, T. (2016). Single chords convey distinct emotional qualities to both naïve and expert listeners. *Psychology of Music*, *44*(1), 37–54.

Lartillot, O., & Toiviainen, P. (2007). A Matlab Toolbox for Musical Feature Extraction From Audio. *Conf. Digital Audio Effects*.

Laurier, C., & Herrera, P. (2007). Audio Music Mood Classification using support vector machine. In *MIREX task on audio mood classification.*

Laurier, C., Lartillot, O., Eerola, T., & Toiviainen, P. (2009). Exploring Relationships between Audio Features and Emotion in Music. In *Proceedings of the 7th Triennal Conference of European Society for Cognitive Sciences of Music* (pp. 260–264).

Laurier, C., Sordo, M., Serra, J., & Herrera, P. (2009). Music mood representation from social tags. In *Proceedings of the 10th International Society for Music Information Retrieval Conference.*

Law, E. L. M., Ahn, L. von, Dannenberg, R. B., & Crawford, M. (2007). TagATune: A game for music and sound annotation. In *Proceedings of the 8th International Conference on Music Information Retreival* (pp. 361–364).

Lee, H., Largman, Y., Pham, P., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in neural information processing systems.*

Li, T., & Ogihara, M. (2003). Detecting emotion in music. In *Proceedings of the International Conference on Music Information Retrieval* (pp. 239–240).

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics (International Biometric Society)*, *45*(1), 255–268.

Lin, Y.-C., Yang, Y.-H., & Chen, H. (2011). Exploiting Online Music Tags for Music Emotion Classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *7S*(1).

Lindström, E., Juslin, P., Bresin, R., & Williamon, A. (2003). Expressivity comes from within your soul: a questionnaire study of music students perspectives on expressivity. *Research Studies in Music Education*, *20*, 23–47.

Liu, D., Lu, L., & Zhang, H.-J. (2003). Automatic Mood Detection from Acoustic music data. In *Proceedings of the 4th International Conference on Music Information Retrieval* (pp. 81–87).

Liu, Y., Liu, Y., & Gu, Z. (2015). Affective Feature Extraction for Music Emotion Prediction. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Lu, L., Liu, D., & Zhang, H. (2006). Automatic Mood Detection and Tracking of Music Audio Signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(1), 5–18.

Lykartsis, A., Pysiewicz, A., Coler, H. von, & Lepa, S. (2013). The Emotionality of Sonic Events: Testing the Geneva Emotional Music Scale (GEMS) for Popular and Electroacoustic Music. In *Proceedings of the 3rd International Conference on Music and Emotion.*

Mandel, M., & Ellis, D. (2008). A web-based game for collecting music metadata. *Journal of New Music Research*, *37*(2), 151–165.

Markov, K., Iwata, M., & Matsui, T. (2013). Music Emotion Recognition using Gaussian Processes. In *Working Notes Proceedings of the MediaEval 2013 Workshop.*

Markov, K., & Matsui, T. (2014). Dynamic Music Emotion Recognition Using State-Space Models. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*

Mckay, C., & Fujinaga, I. (2004). Automatic genre classification using large high-level musical feature sets. In *International Conference on Music Information Retrieval* (pp. 525–530).

McKeown, G. J., & Sneddon, I. (2014). Modeling continuous self-report measures of perceived emotion using generalized additive mixed models. *Psychological Methods*, *19*(1), 155–174.

Mehrabian, A., & Russell, J. A. (1974). *An Approach to Environmental Psychology*. MIT Press.

Müller, M. (2015). Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications. In (chap. Music Structure Analysis). Springer.

Nam, J., Herrera, J., Slaney, M., & Smith, J. O. (2012). Learning sparse feature representations for music annotation and retrieval. In *Proceedings of the 13th International Society for Music Information Retrieval Conference.*

Nicolaou, M. A., Gunes, H., & Pantic, M. (2011). Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence–Arousal Space. *IEEE Transactions on Affective Computing*, *2*(2), 92–105.

Nielzen, S., & Cesarec, Z. (1969). On the perception of emotional meaning in music. *Psychology of Music*(9), 17–31.

Nieto, O., & Jehan, T. (2013). Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In *Proceedings of the 38th IEEE International Conference on Acoustics Speech and Signal Processing* (pp. 236–240).

Pachet, F., & Zils, A. (2004). Automatic extraction of music descriptors from acoustic signals. In *Proceedings of the 3rd International Society for Music Information Retrieval Conference.*

Panda, R., & Paiva, R. P. (2011). Using Support Vector Machines for Automatic Mood Tracking in Audio Music. In *130th Audio Engineering Society Convention.*

Patra, B. G., Maitra, P., Das, D., & Bandyopadhyay, S. (2015). MediaEval 2015: Music Emotion Recognition based on Feed-Forward Neural Network. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Paulus, J., Müller, M., & Klapuri, A. (2010). Audio-based music structure analysis. In *Proceedings of the 11th International Society for Music Information Retrieval Conference* (pp. 625–636).

Payne, E. (1961). Emotion in Music and in music appreciation. *Music Review*, *22*, 39–50.

Pearce, M. T., & Halpern, A. (2015). Age-related patterns in emotions evoked by music. *Psychology of Aesthetics, Creativity, and the Arts*, *9*(3), 248–253.

Pellegrini, T., & Barriere, V. (2015). Time-continuous Estimation of Emotion in Music with Recurrent Neural Networks. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Peltola, H. R., & Eerola, T. (2016). Fifty shades of blue: Classification of music-evoked sadness. *Musicae Scientiae*, *20*(1), 84–102.

Plutchik, R. (1980). *A general psychoevolutionary theory of emotion*. New York: Academic press.

Plutchik, R. (2001). The nature of emotions. *American Scientist*, *89*(344), 319–329.

Pratt, C. (1952). *Music as a language of emotion*. U.S. Government Printing Office.

Raffel, C., & Ellis, D. P. W. (2015). Large-Scale Content-Based Matching of MIDI and Audio Files. In *Proceedings of the 16th International Society for Music Information Retrieval Conference.*

Ratan, R., & Ritterfeld, U. (2009). *Classifying serious games. Serious games: Mechanisms and effects*. Routledge.

Revelle, W., & Rocklin, T. (1979). Very Simple Structure – alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*(4), 403–414.

Rickard, N. S. (2004). Intense emotional responses to music: A test of the physiological arousal hypothesis. *Psychology of Music*, *32*(4), 371–388.

Robinson, C. R. (1988). Differentiated modes of choral performance evaluation using traditional procedures and a Continuous Response Digital Interface device (Doctoral dissertation, Florida State University). *Florida State University*.

Russell, J. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178.

Saari, P., & Eerola, T. (2013). Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, *26*(10), 2548–2560.

Schellenberg, E. G., Corrigall, K. A., Ladinig, O., & Huron, D. (2012). Changing the Tune: Listeners Like Music that Expresses a Contrasting Emotion. *Frontiers in psychology*, *3*(574).

Scherer, K. (2004). Which emotions can be induced by music? What are the underlying mechanisms? And how can we measure them? *The Journal of New Music Research*, *33*(3), 239–251.

Schlüter, J., & Böck, S. (2014). Improved musical onset detection with convolutional neural networks. In *Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing.*

Schmidt, E. M., & Kim, Y. E. (2010). Prediction of Time-Varying Musical Mood Distributions Using Kalman Filtering. In *9th ICMLA* (pp. 655–660).

Schmidt, E. M., & Kim, Y. E. (2011). Modeling musical emotion dynamics with conditional random fields. In *Proceedings of the 12th International Society for Music Information Retrieval Conference.*

Schmidt, E. M., Scott, J., & Kim, Y. E. (2012). Feature learning in dynamic environments: modeling the acoustic structure of music emotion. In *Proceedings of the 13th International Society for Music Information Retrieval Conference.*

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010, March). Feature selection for content-based, time-varying musical emotion regression. In *Proc. ACM Int. Conf. Multimedia Information Retrieval.* Philadelphia, PA. (2010-06-01 11:40:24 -0400)

Schubert, E. (1996). Continuous response to music using a two dimensional emotion space. In *Proceedings of the 4th International Conference of Music Perception and Cognition* (pp. 263–268).

Schubert, E. (1999). Measuring Emotion Continuously: Validity and Reliability of the Two-Dimensional Emotion-Space. *Australian Journal of Psychology*, *51*(3), 154–165.

Schubert, E. (2003). Update of the Hevner adjective checklist. *Perceptual Motor Skills*, *96*, 1117–1122.

Schubert, E. (2004). Modeling Perceived Emotion with Continuous Musical Features. *Music Perception: An Interdisciplinary Journal*, *21*(4), 561–585.

Schubert, E. (2007). The influence of emotion, locus of emotion and familiarity upon preference in music. *Psychology of Music*, *35*(3), 499–515.

Schubert, E. (2013). Reliability Issues Regarding the Beginning, Middle and End of Continuous Emotion Ratings to Music. *Psychology of Music*, *41*(3), 350–371.

Schubert, E., & Dunsmuir, W. (1999). Music, mind, and science. In (pp. 298–352).

Schubert, E., Ferguson, S., Farrar, N., Taylor, D., & McPherson, G. E. (2012). Continuous Response to Music using Discrete Emotion Faces. In *Proceedings of the 9th international symposium on computer music modeling and retrieval.*

Schuller, B., Dorfner, J., & Rigoll, G. (2010). Determination of Nonprototypical Valence and Arousal in Popular Music: Features and Performances. *EURASIP Journal on Audio, Speech, and Music Processing, Special Issue on Scalable Audio-Content Analysis*, 735–854.

Serra, J., Müller, M., Grosche, P., & Arcos, J. L. (2014). Unsupervised Music Structure Annotation by Time Series Structure Features and Segment Similarity. *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*.

Sigtia, S., & Dixon, S. (2014). Improved music feature learning with deep neural networks. In *Proceedings of the 38th International Conference on Acoustics, Speech, and Signal Processing.*

Skowronek, J., McKinney, M., & Par, S. van de. (2006). Ground-truth for automatic music mood classification. In *Proceedings of the 5th International Society for Music Information Retrieval Conference* (pp. 395–396).

Smith, J. B. L., Burgoyne, J. A., Fujinaga, I., Roure, D. D., & Downie, J. S. (2011). Design and creation of a large-scale database of structural annotations. In *Proceedings of the International Society for Music Information Retrieval Conference* (pp. 555–560).

Soleymani, M., Aljanaki, A., Yang, Y.-H., Caro, M., Eyben, F., Markov, K., et al. (2014). Emotional Analysis of Music: a comparison of methods. In *Proceedings of the ACM International Conference on Multimedia.*

Soleymani, M., Caro, M. N., Schmidt, E. M., & Yang, Y.-H. (2013). The MediaEval 2013 Brave New Task: Emotion in Music. In *Working Notes Proceedings of the MediaEval 2013 Workshop.*

Soleymani, M., & Larson, M. (2010). Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus. In *Workshop on*

Crowdsourcing for Search Evaluation, SIGIR 2010. Geneva, Switzerland.

Soleymani, M., Larson, M., Pun, T., & Hanjalic, A. (2014). Corpus development for affective video indexing. *IEEE Trans. Multimedia*, *16*(4), 1075–1089.

Sollberge, B., Rebe, R., & Eckstein, D. (2003). Musical Chords as Affective Priming Context in a Word-Evaluation Task. *Music Perception*, *20*(3), 263–282.

Song, Y., Dixon, S., & Pearce, M. (2012). Evaluation of Musical Features for Emotion Classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference.*

Speck, J. A., Schmidt, E. M., Morton, B. G., & Kim, Y. E. (2011). A Comparative Study of Collaborative vs. Traditional Musical Mood Annotation. In *Proceedings of the 12th International Society for Music Information Retrieval Conference.*

Stern, D. (1985). *The Interpersonal World of the Infant: A View from Psychoanalysis and Development Psychology*. Karnac Books.

Sturm, B. L. (2013). Classification accuracy is not enough: On the evaluation of music genre recognition systems. *Journal of Intelligent Information Systems*, *41*(3), 371–406.

Sturm, B. L. (2014). A Simple Method to Determine if a Music Information Retrieval System is a Horse. *IEEE Transactions on Multimedia*, *16*(6), 1636–1644.

Thayer, R. E. (1989). *The Biopsychology of Mood and Arousal*. New York: Oxford University Press.

Thompson, W. F., Graham, P., & Russo, F. (2005). Seeing music performance: Visual influences on perception and experience. *Semiotica*, *156*, 177–201.

Torres-Eliard, K., Labbe, C., & Grandjean, D. (2011). Towards a Dynamic Approach to the Study of Emotions Expressed by Music. In *Proceedings 4th International ICST Conference on Intelligent Technologies for Interactive Entertainment* (pp. 252–259).

Trochidis, K., Delbe, C., & Bigand, E. (2011). Investigation of the relationship between audio features and induced emotions in Contemporary Western music. In *Sound and Music Computing Conference.*

Turnbull, D., & Lanckriet, G. (2007). A Supervised Approach for Detecting Boundaries in Music Using Difference Features and Boosting. In *Proceedings of the 5th International Conference on Music Information Retrieval* (pp. 42–49).

Ullrich, K., Schlüter, J., & Grill, T. (2014). Boundary detection in music structure analysis using convolutional neural networks. In *Proceedings of the 15th International Society for Music Information Retrieval Conference.*

Vuoskoski, J., & Eerola, T. (2013). A review of music and emotion studies: Approaches, emotion models, and stimuli. *Music Perception*, *30*(3), 307–340.

Vuoskoski, J. K., & Eerola, T. (2010). Domain-specific or not? The applicability of different emotion models in the assessment of music-induced emotions. In *Proceedings of the 10th International Conference on Music Perception and Cognition* (pp. 196–199).

Vuoskoski, J. K., & Eerola, T. (2011). Measuring music-induced emotion: A comparison of emotion models, personality biases, and intensity of experiences. *Musicae Scientiae*, *15*(2), 159–173.

Wager, T., Barrett, L., Bliss-Moreau, E., Lindquist, K., Duncan, S., Kober, H., et al. (2008). Handbook of emotions (3rd edn). In (pp. 249–267). New York: Guilford

Press.

Wang, J. C., Lee, Y. S., Chin, Y. H., Chen, Y. R., & Hsieh, W. C. (2015). Hierarchical Dirichlet Process Mixture Model for Music Emotion Recognition. *IEEE Transactions on Affective Computing*, *6*(3), 261–271.

Wang, J.-C., Yang, Y.-H., Wang, H.-M., & Jeng, S.-K. (2012). The Acoustic Emotion Gaussians Model for Emotion-based Music Annotation and Retrieval. In *Proc. ACM Multimedia* (pp. 89–98).

Wang, J.-C., Yang, Y.-H., Wang, H.-M., & Jeng, S.-K. (2015). Modeling the Affective Content of Music with a Gaussian Mixture Model. *IEEE Transactions on Affective Computing*, *6*(1), 56–68.

Wang, S., & Ji, Q. (2015). Video affective content analysis: a survey of state of the art methods. *IEEE Transactions on Affective Computing*, *PP*(99), 1.

Weber, M., Krismayer, T., Wöß, J., Aigmüller, L., & Birnzain, P. (2015). MediaEval 2015: JKU-Tinnitus Approach to Emotion in Music Task. In *Working Notes Proceedings of the MediaEval 2015 Workshop.*

Wedin, L. (1969). Dimension analysis of emotional expression in music. *Swedish Journal of Musicology*(51), 119–140.

Wedin, L. (1972a). Evaluation of a three-dimensional model of emotional expression in music. *Reports from the Psychological Laboratories*(349).

Wedin, L. (1972b). A Multidimensional Study of Perceptual-Emotional Qualities in Music. *Scandinavian Journal of Psychology*, *13*(1), 241–257.

Weninger, F., Eyben, F., & Schuller, B. (2013). The TUM Approach to the MediaEval Music Emotion Task Using Generic Affective Audio Features. In *Working Notes Proceedings of the MediaEval 2013 Workshop.*

Weninger, F., Eyben, F., Schuller, B. W., Mortillaro, M., & Scherer, K. R. (2013). On the Acoustics of Emotion in Audio: What Speech, Music and Sound have in Common. *Frontiers in Emotion Science*, *4*(Article ID 292), 1–12.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems* (pp. 2035–2043).

Wiggins, G. (2009). Semantic Gap?? Schemantic Schmap!! Methodological Considerations in the Scientific Study of Music. In *Proceedings of IEEE AdMIRe* (pp. 1–7).

Witteveen, J. (2015). Predicting Relevance of Emotion Tags (Master's thesis, Utrecht University). *Utrecht University*.

Wu, B., Zhong, E., Horner, A., & Yang, Q. (2014). Music Emotion Recognition by Multi-label Multi-layer Multi-instance Multi-view Learning. In *Proceedings of the ACM International Conference on Multimedia* (pp. 117–126).

Wülfing, J., & Riedmiller, M. (2012). Unsupervised learning of local features for music classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference.*

Xiao, Z., Dellandrea, E., Dou, W., & Chen, L. (2008). What is the best segment duration for music mood analysis. In *Proceedings of the IEEE International Workshop on Content-Based Multimedia Indexing* (pp. 17–24).

Xu, M., Li, X., Xianyu, H., Tian, J., Meng, F., & Chen, W. (2015). Multi-Scale Approaches to the MediaEval 2015 "Emotion in Music" Task. In *Working Notes*

Proceedings of the MediaEval 2015 Workshop.

Xue, H., Xue, L., & Su, F. (2015). Multimodal Music Mood Classification by Fusion of Audio and Lyrics. In *Multimedia modeling* (pp. 26–37).

Yang, B., & Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. *Signal Processing*, *90*(5), 1415–1423.

Yang, W., Cai, K., Wu, B., Wang, Y., Chen, X., Yang, D., et al. (2014). Beatsens Solution for MediaEval 2014 Emotion in Music Task. In *Working Notes Proceedings of the MediaEval 2014 Workshop.*

Yang, Y.-H., & Chen, H. H. (2011a). *Music Emotion Recognition*. Boca Raton, Florida: CRC Press.

Yang, Y.-H., & Chen, H. H. (2011b). Ranking-Based Emotion Recognition for Music Organization and Retrieval. *IEEE Transactions on Audio, Speech and Language Processing*, *19*(4), 762–774.

Yang, Y.-H., & Chen, H. H. (2012). Machine Recognition of Music Emotion: A Review. *ACM Transactions on Intelligent Systems and Technology*, *3*(3), 1–30.

Yang, Y.-H., Lin, Y.-C., Su, Y.-F., & Chen, H. H. (2008). A Regression Approach to Music Emotion Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 448–457.

Yang, Y.-H., Liu, C.-C., & Chen, H. H. (2006). Music Emotion Classification: A Fuzzy Approach. In *Proceedings of the 14th Annual ACM International Conference on Multimedia* (pp. 81–84).

Yeh, C.-C. M., Su, L., & Yang, Y.-H. (2013). Dual-layer bag-of-frames model for music genre classification. In *Proceedings of the 37th International Conference on Acoustics, Speech, and Signal Processing.*

Zagrodski, M. (2013). Influence of Musical Context on the Perception of Emotional Expression of Music. In *Proceedings of the 3rd International Conference on Music & Emotion.*

Zentner, M., & Eerola, T. (2011a). Handbook of Music and Emotion: Theory, Research, Applications. In (chap. Music Education: the role of affect). Oxford University Press.

Zentner, M., & Eerola, T. (2011b). Handbook of Music and Emotion: Theory, Research, Applications. In (chap. Self-report measures and models). Oxford University Press.

Zentner, M., Grandjean, D., & Scherer, K. R. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8*(4), 494–521.

Zhou, X., & Lerch, A. (2015). Chord detection using Deep Learning. In *Proceedings of the 16th International Society for Music Information Retrieval Conference.*

Samenvatting in het Nederlands

Het beschrijven van audiomuziek door middel van emoties is een subjectieve taak, zodat we afhankelijk zijn van gegevens van menselijke beoordelaars. De kwaliteit daarvan is cruciaal voor het trainen van computerprogrammas. Het benoemen van muziekemotie op een natuurlijke manier is belangrijk voor de kwaliteit van trainingsgegevens en voor het ontwikkelen van intutieve aanbevelings-systemen voor muziek. Dit onderzoek draagt daar op drie manieren aan bij.

Ten eerste modelleren we welke muziekaspecten emotie opwekt. Daarvoor is de game Emotify ontwikkeld die gegevens verzamelt over de emotie van de spelers bij afgespeelde muziek. Met die game zijn trainingsgegevens verzameld van hoge kwaliteit. Op basis daarvan hebben we een computerprogramma ontwikkeld dat emotie relateert aan kenmerken van de muziek. Het blijkt dat het beter werkt om nieuwe kenmerken mee te nemen dan om al gebruikte kenmerken slimmer te verweken. De harmonie in de muziek is daar een goede kandidaat voor.

Vervolgens maken we een standaardtest voor Muziek Emotie Variatie Detectie computerprogrammas. We voeren een systematische evaluatie uit van computerprogrammas en de muziekkenmerken. Wat het beste werkt is om aparte kenmerken te nemen voor de twee dimensies waardering en opwinding, die in de psychologie veel gebruikt worden.

Daarop voortbouwend ontwikkelen we een nieuwe manier van Muziek Emotie Variatie Detectie door de muziek als een opeenvolging van emotioneel stabiele segmenten en onstabiele overgangs-segmenten.

Een beter begrip van hoe muziek emotie opwekt heeft de potentie om bij te dragen aan muziekspelers met betere aanbevelingsfunctie, en kan een rol spelen in het monitoren van mentaal welzijn.

Curriculum Vitae

Anna Aljanaki was born on the 8th of August 1987 in Feodosia, Krim (USSR at the time). In 2005, she went to study computer science at the University of Tartu. In 2011, she graduated with an MSc degree. Her master thesis concerned a Music Information Retrieval topic, automatic key detection. Since then Anna was interested in continuing this sort of research. MIR is a cross-section of her interests: music and computer science. Anna plays several music instruments and sings and is generally very interested in music, and in music processing both from computational and cognitive side.

In 2012, Anna Aljanaki started her PhD research at the Department of Information and Computing Sciences in Utrecht University. This dissertation is the result of the work she carried out during her four years in Utrecht.