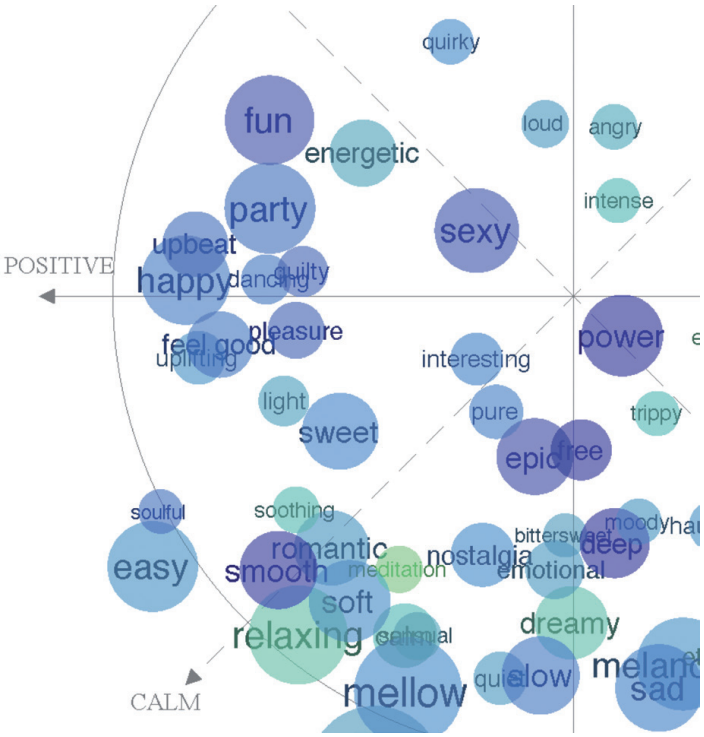
JYVÄSKYLÄ STUDIES IN HUMANITIES
243

Pasi Saari

Music Mood Annotation Using Semantic Computing and Machine Learning



JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN HUMANITIES 243

Pasi Saari

Music Mood Annotation Using Semantic Computing and Machine Learning

Esitetään Jyväskylän yliopiston humanistisen tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Historica-rakennuksen salissa H320
tammikuun 31. päivänä 2015 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Humanities of the University of Jyväskylä,
in building Historica, hall H320, on January 31, 2015 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Music Mood Annotation Using Semantic Computing and Machine Learning

JYVÄSKYLÄ STUDIES IN HUMANITIES 243

Pasi Saari

Music Mood Annotation Using Semantic Computing and Machine Learning

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Editors
Tuomas Eerola
Department of Music, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-6074-2
ISBN 978-951-39-6074-2 (PDF)
ISSN 1459-4331

ISBN 978-951-39-6073-5 (nid.)
ISSN 1459-4323

Jyväskylä University Printing House, Jyväskylä 2014

ABSTRACT

Saari, Pasi
Music Mood Annotation Using Semantic Computing and Machine Learning
Jyväskylä: University of Jyväskylä, 2015, 58 p.(+included articles)
(Jyväskylä Studies in Humanities
ISSN 1456-5390; 243)
ISBN 978-951-39-6073-5 (nid.)
ISBN 978-951-39-6074-2 (PDF)
Finnish summary
Diss.

The main appeal of music lies in its capability to express moods, so mood-based music discovery and management is highly beneficial. Online music services enable access to wide music catalogues, and social and editorial tagging produces large amounts of semantic information on music mood. However, tag data are not as reliable at representing moods expressed by music as self-report data obtained from listening experiments. The primary aim of the present work was to examine computational methods for enhancing the existing mood tag data and to enable automatic annotation of music according to the expressed mood. Semantic computing based on large-scale tag data aims to improve the accuracy of tags at representing moods, and a novel technique called Affective Circumplex Transformation (ACT) was proposed for this purpose. Improvements to the generalizability of audio-based mood annotation performance were sought using audio feature selection and a proposed technique termed as Semantic Layer Projection (SLP) that efficiently incorporates large-scale tag data. Moreover, a genre-adaptive technique was proposed to take into account genre-specific aspects of music mood in audio-based annotation. Performance evaluation of the techniques was carried out using social and editorial tags, listener ratings, and large corpora representing popular and production music. ACT led to clear improvements in the accuracy as opposed to raw tag data and conventional semantic analysis techniques. Moreover, ACT models could be generalized across tag types and different music corpora. Audio-based annotation results showed that exploiting tags and semantic computing using SLP can lead to similar or even higher performance than tag-based mood inference. Adapting both the semantic mood models and audio-based models to different genres led to further improvements, especially in terms of the valence dimension.

Keywords: music mood annotation, music emotion recognition, social tags, editorial tags, circumplex model, feature selection, genre-adaptive, semantic computing, audio feature extraction

Author

Pasi Saari
Department of Music
University of Jyväskylä, Finland

Supervisors

Professor Tuomas Eerola
Department of Music
Durham University, United Kingdom

Doctor Olivier Lartillot
Department for Architecture, Design and Media Technology
Aalborg University, Denmark

Reviewers

Professor Dan Ellis
Department of Electrical Engineering
Columbia University, USA

Associate Professor George Tzanetakis
Department of Computer Science
University of Victoria, Canada

Opponent

Professor Björn W. Schuller
Chair of Complex Systems Engineering
University of Passau, Germany
Senior Lecturer at Department of Computing
Imperial College London, United Kingdom

ACKNOWLEDGMENTS

During the time working on this thesis, I was fortunate to be part of the Finnish Centre of Excellence in Interdisciplinary Music Research at the University of Jyväskylä. The team has been a great example of a group of innovative, hard-working and talented people who join together on a common mission to build up an internationally renowned research center. I think this is something that is way above of what one could expect to suddenly emerge from a small town somewhere in the north. The results speak for themselves, and I feel humbled to have contributed even a little to that story through this work.

First and foremost, I am grateful to my supervisor Professor Tuomas Eerola for having the confidence in me, for acting as a warm and caring mentor, and for being a perfect collaborator – both inspiring and inspired. I felt it was thoroughly satisfying how these multiple roles switched organically and supported my work. Thanks for supporting me in times of trouble, for celebrating some academic victories with me, and for retaining the encouraging attitude during the more independent phases of my work. My sincere gratitude goes to Doctor Olivier Lartillot, my second supervisor, for first of all accepting my application to the doctoral student position at the CoE and for being a pleasant office mate. I benefited from your expertise a lot. Thanks for introducing me to MIRtoolbox and for allowing me to contribute my small add-ons as well. Many thanks to Professor Petri Toiviainen for being a true role model as a research leader, for being a dependable resort of advice, and for having the confidence in my work. Also thanks for agreeing to be the custos at my defence.

I want to express my gratitude to my external readers Professor Dan Ellis and Associate Professor George Tzanetakis. I feel I was privileged to receive the attention from two such high-level experts whose work in MIR receive my utmost admiration. Thanks for all the positive and inspiring feedback, and for suggestions that I did my best to properly acknowledge in the final version of the thesis summary. Thank you Professor Björn Schuller for taking the time and lending your insights as my opponent at my defence.

My warm thanks go to the people at the Department of Music at the University of Jyväskylä. Thank you Markku Pöyhönen for so well making sure that everything goes smoothly in the backstage of the research arenas and for your devotion to make the lives of everyone around you more enjoyable. Thank you Suvi Saarikallio and Geoff Luck for teaching me various topics important for this dissertation during my undergraduate studies in the MMT, and for being nice colleagues thereafter. Thanks to my fellow, present or former, doctoral students, including Birgitta Burger, Anemone Van Zijl, Martín Hartmann, Henna Peltola, Marc Thompson, Rafael Ferrer, Jonna Vuoskoski, Vinoo Alluri, and Iballa Burunat, for the friendship and the many lunch, coffee, and other breaks. Thanks to Mikko Leimu, Mikko Myllykoski, Tommi Himberg, Jaakko Erkkilä, Esa Ala-Ruona, Jörg Fachner, and Marko Punkanen for being integral to forming a fine research and teaching environment. Finally, thanks to Riitta Rautio and Erkki

Huovinen for handling many practical issues regarding the completion of this thesis.

During this work I had the opportunity to pay a research visit to the Centre for Digital Music at Queen Mary University of London. The three months spent there played a pivotal role in the preparation of many of the publications included in this thesis. My gratitude goes especially to György Fazekas and Mathieu Barthet for their fruitful collaboration, for including me in the activities of their research projects, and for their sincere kindness. Thank you Mark Plumbley and Mark Sandler for supporting my stay. I also want to express special thanks to Yading Song, Katerina Kosta, and Ioana Dalca for making my stay much more pleasant. During the later phase of the thesis preparation, I was fortunate to be hired as a Principal Researcher at Nokia Technologies. This appointment gave me useful insights into how MIR research can be applied in the industry to solve real-world problems on a large scale. I want to thank my colleagues at Nokia, especially Jussi Leppänen, Arto Lehtiniemi, Antti Eronen, and Ville-Veikko Mattila for the supportive and inspiring attitude, and for introducing me to the intriguing world of professional IPR work.

Unquestionably, my biggest gratitude goes to my wife Anna Pehkoranta. Your scholarly endeavors first stirred up in me the desire to pursue a research career. Most importantly, your love is the engine under my hood and my main source of inspiration. Although my research journey has several times taken me to distant places, my home remains wherever you are. I also want to thank my parents and family for their love and support over the years, and my late grandmother Leena Turunen for exemplifying a thoroughly humane approach for dealing with life's challenges. I dedicate this work to her memory.

Durham, UK, 7 January 2015

Pasi Saari

CONTENTS

LIST OF PUBLICATIONS

List of the publications (reprinted after the introductory part) included in this thesis. The first author was the primary contributor to the data collection, experimental design, implementation, analysis, and paper writing of all of the publications.

I Pasi Saari, Tuomas Eerola & Olivier Lartillot. Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music. *IEEE Transactions on Audio, Speech, and Language Processing*, 19 (6), 1802-1812, 2011.

II Pasi Saari & Tuomas Eerola. Semantic Computing of Moods based on Tags in Social Media of Music. *IEEE Transactions on Knowledge and Data Engineering*, 26 (10), 2548-2560, 2014.

III Pasi Saari, Mathieu Barthet, György Fazekas, Tuomas Eerola & Mark Sandler. Semantic Models of Mood Expressed by Music: Comparison Between Crowd-sourced and Curated Editorial Annotations. *In IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, 2013.

IV Pasi Saari, Tuomas Eerola, György Fazekas & Mark Sandler. Using Semantic Layer Projection for Enhancing Music Mood Prediction with Audio Features. *In Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013)*, 722-728, 2013.

V Pasi Saari, Tuomas Eerola, György Fazekas, Mathieu Barthet, Olivier Lartillot & Mark Sandler. The Role of Audio and Tags in Music Mood Prediction: A Study Using Semantic Layer Projection. *In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR)*, 201-206, 2013.

VI Pasi Saari, György Fazekas, Tuomas Eerola, Mathieu Barthet, Olivier Lartillot & Mark Sandler. Genre-adaptive Semantic Computing Enhances Audio-based Music Mood Prediction. *submitted*.

1 INTRODUCTION

Music listening has been revolutionized in recent times with the shift to the digital age and the introduction of online music services. Online digital music catalogues include millions of tracks, and music streaming enables listeners to access, virtually, all the music ever recorded. Moreover, music listening is increasingly linked to social networking, which produces a vast amount of contextual information that is of interest to both the music industry and researchers. To facilitate access to large-scale music data, the interdisciplinary research field of Music Information Retrieval (MIR) has promoted the development of techniques and applications for automatic music annotation. Advances in this field have been made by combining areas as diverse as musicology, signal processing, data mining, machine learning, social sciences, psychology, and cognitive sciences.

The ability of music to express and evoke moods is one of the key reasons why it appeals to people (Juslin & Sloboda, 2009), and an extensive body of research in music psychology and affective sciences is found to be dedicated to seeking the understanding of the phenomena related to music-mediated moods. It is therefore important to support mood-based music discovery in modern-day music listening. A number of online music services incorporate mood information as the mode of access to music[1]. The development of techniques that facilitate the organizing and retrieving of music by mood has also been one of the central topics in MIR research.

Music mood annotation that relies on audio content is one of the key technologies enabling mood-based music discovery. Audio-based music mood annotation or music emotion recognition (Lu, Liu, & Zhang, 2006; Yang, Lin, Su, & Chen, 2008) refers to the automated recognition and annotation of the mood associated with music items, such as tracks or clips, represented by digital audio. A system extracts computational features from audio tracks and applies machine learning to train models that map the features to mood, based on a set of human-labeled music items (Kim et al., 2010; Barthet, Fazekas, & Sandler, 2012). Manual mood annotations are therefore considered the ground truth, and the way moods

[1] These currently include Gracenote (`http://www.gracenote.com/`), Musicovery (`http://musicovery.com/`) and Allmusic (`http://www.allmusic.com/`).

are represented typically conforms to scientifically justified dimensional (Russell, 1980; Thayer, 1989) or categorical (Ekman, 1992; Hevner, 1936) emotion models. The ground truth gathered in controlled or semi-controlled listening experiments and aggregated from multiple participants is arguably the most reliable semantic data attainable for representing moods in music (Zentner & Eerola, 2010). However, the laborious task of obtaining these types of listener ratings restrict the data set sizes, affecting the generalizability of the obtained models to the large music catalogues typical of today's music listening.

The present work studies three approaches to increase the generalizability and thereby the performance of music mood annotation. First, advances can be made by focusing on improving the robustness of machine learning techniques to better exploit the limited data available for model training. This approach requires finding a balance to robustly exploit the training data while not blindly relying on it. Blind reliance would lead to problems related to overfitting (Hawkins, 2004) and hence decrease the generalizability of the obtained models. The present work addresses this problem by applying audio feature selection to optimize the input feature subset employed for annotation.

The second approach is to include a large amount of annotated training data that might be less reliable than listener ratings. Free-form and unstructured social tags provided by user communities in online music services (Lamere, 2008) and editorial tags for music catalogues offer large-scale semantic data useful for model training. Tags have been exploited to index music according to semantic concepts including mood and genre, mainly for the purposes of music recommendation (Eck, Lamere, Bertin-Mahieux, & Green, 2007; Kaminskas & Ricci, 2012). However, exploiting such data demands attention to the reliability of the data itself and applicable data analysis techniques. The reliability of tags at representing the mood of music items has not been evaluated systematically with reliable listener ratings. Previous studies have shown that semantic computing based on social tags associated with music can yield mood representations resembling those suggested by emotion research (Levy & Sandler, 2007; Laurier, Sordo, Serra, & Herrera, 2009). Moreover, semantic analysis has yielded positive results for music recommendations (Symeonidis, Ruxanda, Nanopoulos, & Manolopoulos, 2008). The advantages of semantic computing at inferring the mood of music tracks and its utility in music mood indexing have not been substantiated with reliable ground truth that conforms to emotion models. MIR research has widely applied audio-based techniques to predict tags associated to music items, alleviating the so-called cold-start problem, i.e., the problem of recommending items not yet tagged or seen by users (Lam, Vu, Le, & Duong, 2008). Audio-based semantic annotation or auto-tagging can, therefore, complement item-based collaborative filtering in music recommendation (Sarwar, Karypis, Konstan, & Riedl, 2001). Auto-tagging has been successful at annotating music with various concepts including mood, genre, instrumentation, and usage (Turnbull, Barrington, Torres, & Lanckriet, 2008; Bertin-Mahieux, Eck, Maillet., & Lamere, 2008; Bischoff et al., 2009). Furthermore, semantic computing applied in conjunction with audio-based annotation has yielded positive results for mu-

sic auto-tagging (Levy & Sandler, 2009; Law, Settles, & Mitchell, 2010). However, despite the benefit of semantic computing for improving auto-tagging and determining mood representations from tag data (Levy & Sandler, 2007; Laurier et al., 2009), previous studies have not sufficiently examined the benefit of semantic computing for audio-based music mood annotation.

Finally, the third approach tackles the annotation of large and heterogeneous music catalogues by adapting audio-based models to music genres. Genres are the most typical categories used to organize music catalogues in libraries and music stores (Scaringella, Zoia, & Mlynek, 2006). Large music catalogues comprise a large variety of musical genres, so robust audio-based music mood annotation that can perform well across diverse musical materials is required. Previous studies have shown that music genre is related to mood at various levels. Different genres represent different moods (Hu & Downie, 2007), and the patterns in which audio features relate to moods differ between genres (Eerola, 2011). Moreover, exploiting music genre in mood annotation has been shown to improve annotation accuracy (Lin, Yang, & Chen, 2009, 2011). However, previous studies have not shown whether tag-based semantic models of music mood benefit from adapting to different genres.

The present work is organized as follows. Chapter 2 provides a theoretical background to the definition and conceptualization of music mood. Chapter 3 reviews previous studies on music mood annotation, concentrating on self-reports, semantic analysis of tags, and audio-based approaches. Chapters 4 and 5 introduce the aims of the study and review the materials, methods, and results of the included research articles (Studies I-VI). Finally, Chapter 6 presents implications and conclusions of the study.

2 MUSIC MOOD

2.1 Definition

Defining the theoretically distinguishable but often interchangeably applied concepts of mood, emotion, affect, and feeling is a challenging and non-trivial problem (Russell & Barrett, 1999; Scherer, 2005). A commonly accepted view is that emotion comprises a set of components: cognitive appraisal, bodily reactions, action tendencies, motor expressions, and subjective feelings (Scherer, 2005). Notably, per this view, the term "feeling" is often regarded synonymous to emotion, while it is actually one of the components of emotion. Moods, on the other hand, have been described as diffuse and enduring affect states (Scherer, 2005) that have a lower intensity than emotions, last longer than emotions, and have no clear object (Juslin, 2013a). However, no clear consensus exists in the academic literature as to under which criteria mood and emotion can be distinguished (Beedie, Terry, & Lane, 2005). "Affect" is considered an umbrella term encompassing both emotion and mood as well as preferences (Juslin & Västfjäll, 2008). On the other hand, Russell and Barrett (1999) distinguished between prototypical emotional episodes concerned with, or directed at, specific objects – persons, conditions or events – and "core affects" that are the most elementary affective feelings not necessarily directed to anything. Russell and Barrett (1999) also defined mood as a prolonged core affect.

Emotions involved in music listening differ from everyday emotions (Juslin, 2013a). Affective states evoked by music listening are related to utilitarian emotions that are close to those experienced in the everyday context such as happiness and sadness as well as to aesthetic or music-specific emotions that are induced by the appreciation of intrinsic qualities of a music piece (Scherer, 2004; Juslin, 2013a), such as wonder, admiration, ecstasy, and solemnity. Emotions involved in music listening could also be mixtures of several emotions such as happiness and sadness (Gabrielsson, 2010). When studying emotions involved in music listening, a distinction must be made between emotions evoked by music and emotions expressed by music. Similar emotions may be induced and ex-

pressed by music, but self-reports of induced and expressed emotions related to a music piece are not necessarily the same (Gabrielsson, 2002). Similar to induced moods, assessment of expressed moods relies on the unique impressions of individual listeners (Juslin, 2013b). The present work, however, focuses on emotions expressed by music.

For music information retrieval purposes, this research has mainly focused on the emotions expressed by music, also termed as "music mood" (Lu et al., 2006). In MIR research, the most frequently used term is "mood", while "emotion" and "affect" have also been used to a similar extent[1]. However, terminological choices do not necessarily reflect the theoretical distinctions, and as in general emotion research (Beedie et al., 2005), the terms have been applied interchangeably. Taking the duration criterion as an example, "emotion" has been used to describe both the static emotion of an entire music clip (Yang et al., 2008) and the time-varying emotion within a clip (Schmidt, Turnbull, & Kim, 2010). Similarly, the term "mood" has been used for referring to both the time-varying and clip-level moods (Lu et al., 2006).

Although the MIR view of music mood focuses on the properties of a music piece, music mood is ultimately composed of the unique impressions of individual listeners (Juslin, 2013b). The moods a listener perceives in music tracks are influenced by multiple factors: the musical structure (Gabrielsson & Lindström, 2001), listener attributes such as individual mood and personality (Vuoskoski & Eerola, 2011), and listening context (Scherer & Zentner, 2001). Although the influence of listener-related and contextual factors limits the degree to which music mood can be derived from the musical structure, the influence of musical structure transcends the cultures of the listeners to some degree (Balkwill & Thompson, 1999; Fritz et al., 2009) and is not highly dependent on the listeners' musical expertise (Bigand, Vieillard, Madurell, Marozeau, & Dacquet, 2005).

2.2 Models

Although several different approaches have been applied in music research to model emotions in music, two main types have been used frequently: the categorical and dimensional models. Over 70% of the articles on music and emotion published between 1988 and 2008 apply either of these models (Eerola & Vuoskoski, 2013). MIR research has adopted these emotion models either directly or with modifications to represent music moods.

2.2.1 Categorical Model

The categorical model is closely related to the theory of basic emotions (Ekman, 1992) postulating that all emotions can be derived from a limited set of innate

[1] This can be confirmed by searching for these terms in paper titles in the proceedings of the ISMIR in 2000-2012 (See `http://www.ismir.net`)

and universal basic emotions such as anger, fear, sadness, happiness, surprise, and disgust. The distinct and identifiable characteristics of these six emotions are clearly indicated by the analysis of facial expressions and autonomic nervous system patterns (Ekman, 1999).

The set of emotion categories have typically been modified to better account for the emotions that are relevant to music. Either slight modifications or completely different categorizations have been proposed. Vieillard et al. (2008) used the categories happy, sad, scary, and peaceful to represent emotions expressed by musical excerpts, whereas Eerola and Vuoskoski (2011) replaced disgust with tenderness in the typical set of basic emotions and omitted surprise owing to low agreement between listeners. Categorizations specifically developed to represent emotions related to music include a set of adjectives organized into eight groups (Hevner, 1936) and the updated version featuring nine groups (Schubert, 2003). An attractive property of these two models is that the groups form a circular structure in an emotion plane, providing a link between the categorical and dimensional models. Nevertheless, evidence also speaks in favor of the basic emotion categorization of music, since anger, fear, sadness, happiness, and tenderness are among the emotions most frequently expressed by music according to listeners (Juslin, 2013b).

2.2.2 Dimensional Model

The dimensional model is based on the assumption that affective states arise from a few common neurophysiological systems (Plutchik, 1980). In 1897, Wundt (1897) proposed a dimensional model that included the dimensions of pleasure–displeasure, arousal–calmness, and tension–relaxation. Since then, different formulations of the model have been proposed, comprising typically two or three dimensions. A notable commonality between most of the models is the inclusion of some form of the arousal component (the intensity of emotion or activity) and valence component (corresponding to the pleasure–displeasure or positive–negative affect) (Scherer, 2000). The most well-known example of the two-dimensional model is the circumplex model of emotion (Russell, 1980) which suggests that all emotions can be identified and distinguished by their placement in the dimensions of valence and arousal and that the terms form a circular structure at the perimeters of the space. This structure was derived using factor analyses of the ratings of similarities between emotion-related terms. The circumplex model thus implies that emotions close to one another in the space are similar, and conversely, emotions at the opposite ends of the space can be considered to be bipolar. Scherer (1984) performed factor analysis similar to Russell (1980) on similarity judgment of emotion terms and provided evidence disputing the circular structure of emotions, claiming that the entire two-dimensional space may be filled when a larger number of emotion terms are evaluated. Nevertheless, the analyses supported fairly well the two-dimensional valence-arousal representation although different labels of the underlying dimensions of the factor model could have also been justified.

A variant of the two-dimensional model widely used in emotion research is the multidimensional model of activation proposed by Thayer (1989). This model divides the affect space into bipolar dimensions of energetic arousal and tense arousal. However, this model has been seen as being compatible with the circumplex model if the two dimensions are rotated by 45° and represented by a mixture of valence and arousal (Yik, Russell, & Barrett, 1999). Schimmack and Grob (2000), on the other hand, claimed that valence, energy, and tension cannot be reduced to two dimensions and that more than three dimensions are required to sufficiently account for the structure of emotions. Bradley and Lang (1994) suggested using dominance instead of tension as the third dimension to measure a person's affective reactions to a wide variety of stimuli. This model of valence, arousal, and dominance has been applied to collect affective ratings for large numbers of words in English language (Bradley & Lang, 1999; Warriner & Brysbaert, 2013).

Zentner, Grandjean, and Scherer (2008) argued that the dominant dimensional and categorical models cannot explain music-specific emotions and proposed a model consisting of nine music-specific dimensions (wonder, transcendence, tenderness, nostalgia, peacefulness, power, joyful activation, tension, and sadness) and three higher-level dimensions (sublimity, vitality, and unease). Nevertheless, valence and arousal are the two dimensions frequently employed to study music and emotion (Eerola & Vuoskoski, 2013). The circumplex model, in particular, has favorable qualities for music research, since it lends itself to the study of a variety of affect states, represented by points in the emotion space, in relation to the underlying dimensions. Analysis of self-reports of the expressed emotion in music has also shown that music represented by valence and arousal can be robustly mapped to the basic emotions, indicating high correspondence between the categorical and dimensional model (Eerola & Vuoskoski, 2011).

3 MUSIC MOOD ANNOTATION

A wealth of research has been dedicated to automated music mood annotation, also frequently termed as Music Emotion Recognition (MER) (Kim et al., 2010; Yang & Chen, 2012; Barthet et al., 2012). This section reviews the relevant background literature on the topic.

3.1 Self-reports

MIR research on music mood annotation essentially relies on self-report data. Self-reports on music mood can be divided into data gathered from listening tests taken by participants specifically for research purposes, editorial annotations organizing commercial and production music catalogues, and social tags crowd-sourced from web communities. Scattered semantic information about music may also be obtained from diverse sources such as microblogs (Schedl, Hauger, & Urbano, 2013), web searches (Knees, Pampalk, & Widmer, 2004), and lyrics (Hu, Downie, & Ehmann, 2009), but exploiting these sources is out of the scope of the present work.

3.1.1 Listener Ratings

Research-oriented data refers to data collected from self-reports of music mood, derived from emotion studies in psychology, conforming in general to the categorical or dimensional models of emotion (Zentner & Eerola, 2010). Data are typically obtained from participants who are asked to listen to music pieces and report the related emotions, either perceived or induced. This approach is supported by patterns observed in psychophysiological (Gomez & Danuser, 2004) and neurological (Juslin, Harmat, & Eerola, 2013) data. Self-report methods include the use of ordinal Likert scales and adjective checklists. Since the assessment of music mood relies on unique impressions of individual listeners (Juslin, 2013b), a more objective measure can be achieved by aggregating self-reports ob-

tained from several individuals. It can be argued that the mood expressed by a music item can be reliably assessed in a listening test involving a large pool of annotators.

Focusing on the perceived emotion, Eerola and Vuoskoski (2011) conducted two listening experiments where participants were asked to rate clips of film soundtracks according to arousal, valence, tension, and six basic emotions. For the first experiment, 12 expert musicologists selected 360 excerpts between 10 and 30 seconds long and evaluated the excerpts using seven-point Likert scales. The ratings were analyzed, and a subset of 110 excerpts that clearly expressed distinct emotions were selected for further analysis. In the second experiment, 116 university students were asked to rate the excerpts on nine-point Likert scales. The inter-rated consistency was sufficient enough to represent the excerpts as the mean across participants. The mean ratings from the second experiment were subsequently used for audio-based modeling by Eerola, Lartillot, and Toiviainen (2009). Study I in the present work, on the other hand, transforms the rating data to discrete basic emotion categories. In another study, Yang et al. (2008) collected ratings of valence and arousal from multiple participants for 25-second-long clips from 195 tracks comprising a range of musical styles and cultures (Western, Chinese, and Japanese). In all, 253 volunteers participated in the experiment, and each participant rated 10 excerpts randomly drawn from the data set on 11-point Likert scales. Each clip was then represented by the average rating across participants in further audio-based modeling. Hu, Downie, Laurier, Bay, and Ehmann (2008) employed a web survey to collect categorical self-report data, forming a ground-truth set for the Music Information Research Evaluation eXchange Audio Mood Classification task (MIREX AMC)[1]. A data set of 600 music clips was sampled from a large production music catalogue so that it was distributed evenly across five mood adjective clusters according to pre-assigned mood labels by catalogue curators. The online listening experiment involved a mini-training on exemplar songs to articulate to the participants what each mood cluster meant. Moreover, participants were instructed to ignore lyrics in the evaluations to facilitate audio-based mood inference. The listeners agreed on 68% of the labels, and, thus, the authors recommended to exclude clips that do not receive agreement to reduce ambiguity in future evaluation data sets.

Turnbull et al. (2008) established the CAL500 data set by recruiting participants to listen to music tracks and evaluate them according to 135 semantic concepts related to mood, genre, instrumentation, solo, usage, and vocals. The data set consists of 500 tracks by unique artists and covers Western popular music. In particular, participants were asked to rate 18 mood-related words on a three-point bipolar scale. For each track, annotations were collected from at least three participants, and the labels obtained from each participant were aggregated. Mood words that were rated using the bipolar scale were transformed to unipolar labels by separating each scale into two unipolar tags indicative of negative and positive associations. The CAL500 is useful for benchmarking auto-tagging algorithms since it comprises strongly labeled tags, i.e., the absence of a tag means

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

that the tag is not relevant although the subjectivity of moods makes this assumption problematic.

The laboriousness of collecting self-reports tends to limit the number of music pieces that can be assessed. This is a serious limitation to audio-based modeling of music mood where model training is typically performed with majority of the data, resulting in only a few excerpts being available for model evaluation. In order to build larger data sets, previous studies have resorted to recruiting few human evaluators (Trohidis, Tsoumakas, Kalliris, & Vlahavas, 2008; Schuller, Hage, Schuller, & Rigoll, 2010) or even a single labeler (Li & Ogihara, 2003; Wieczorkowska, Synak, & Zbigniew, 2006). Nevertheless, the gains to the amount of data usually come at the expense of reliability, generalizability, and accuracy. When subjective annotations are gathered from few participants, the responses may not be generalizable to larger populations, and human error may play a significant role in the aggregated data.

3.1.2 Editorial Tags

Several commercial music catalogues are actively annotated by music experts using semantic tags. Although the details of the annotation process are considered classified information and the data are proprietary, industry collaborations and public availability have enabled MIR research to exploit certain parts of the data sets.

The most well-known study that gathered editorial tags was conducted by Pandora's Music Genome Project [2], mainly for the purpose of personalized radio playlist generation. Expert musicologists are trained to annotate music tracks according to several hundreds of "musically objective" tags, mostly related to genre, sound, musical structure, and instrumentation. It is estimated that as many as one million tracks were annotated in the first 14 years of the project. Tingle, Kim, and Turnbull (2010) harvested a 10,000-track subset of this data to be shared with MIR researchers. However, because of the objective nature of the evaluated tags, moods were not represented in the data. In contrast, mood data in the Allmusic.com web service [3] constituted a major part of editorial tags submitted for music albums along with album reviews and genre tags. The mood tag vocabulary in the service comprises more than 300 unique tags, which is why the produced data are frequently exploited in MIR research on mood (B. Han, Rho, Jun, & Hwang, 2009; Lin et al., 2009). However, the album-level tags do not represent very well the music mood at the track-level.

Production music available from various stock media houses is also frequently tagged by editors according to mood. Moods are considered an important search criteria for this type of music, usually composed to be included in productions such as films, commercials, and radio broadcasts. For example, the I Like Music (ILM) catalogue, aggregating several production music libraries,

[2] http://www.pandora.com/about/mgp

[3] http://www.allmusic.com

refers to production music as "mood music"[4]. Another aggregated production music catalogue was exploited by Hu et al. (2008) to sample tracks according to the mood content for the MIREX AMC data set. An alternative to collecting tag data by conducting typical listening tests or hiring music experts is the so-called games-with-a-purpose recommended and developed by (Law & Von Ahn, 2009; Mandel & Ellis, 2008). Although these games involve mechanisms for curating the annotation process, the obtained tag data are not strongly labeled, and mood-related tags represent only a small minority of all the tags in the existing data sets. However, Kim, Schmidt, and Emelle (2008) developed an online game focusing on mood, involving participants to collaboratively assign tracks to coordinates on the valence-arousal plane while curating each other's assignments.

Editorial tags provide semantic data about large collections of music. However, exploiting these data in the research on music mood is problematic because of the proprietary nature of these editorial data and because a major proportion of the data are not necessarily related to mood. Moreover, the efficiency of editorial tags in the evaluation of ground truth for music mood modeling is questionable since editorial tags may neither describe mood accurately at the track level nor represent the agreement of a large group of listeners.

3.1.3 Social Tags

Social tags can be defined as free-form textual labels or phrases collaboratively applied to particular resources by users in online services. Social tagging produces a vast amount of information on many websites, such as Del.icio.us (web bookmarks), Flickr (photos), and Pinterest (images, videos, etc.)[5]. The most well-known site supporting social tagging of music is Last.fm[6] that exploits user-generated data for music discovery and for generating personalized recommendations. Last.fm users apply tags to music tracks, albums, and artists and register their music listening via specified desktop and mobile applications on the website profile. Last.fm tags have been of interest to MIR research since a significant proportion of the tag data is accessible through a developer API, allowing the study of music listening behavior to be conducted at an unprecedentedly large scale. On looking at the data from the Last.fm website and data used in the present work and in other research papers, e.g., Laurier et al. (2009), it can be seen that the Last.fm tag data comprises tens of millions of tracks and millions of unique tags applied by millions of users. The majority of Last.fm tags are descriptors of the type of music content, referring typically to genres (Bischoff, Firan, Nejdl, & Paiu, 2008) as well as to moods, locales, and instrumentations that are well represented in the data as well. Moods account for an estimated 5% of the most prevalent tags (Lamere, 2008).

There are a number of incentives and motivations for users to apply social

[4] C.f. `http://media.ilikemusic.com/ilm-media/faqs.php`

[5] Del.icio.us: `http://www.delicious.com`; Flickr: `http://www.flickr.com`; Pinterest: `http://pinterest.com`.

[6] `http://www.last.fm/`

tags to music, such as task organization, social signaling, opinion expression, and social contribution (Ames & Naaman, 2007). For instance, tags such as "Female", "Electronic", or "Check out" may be applied for future personal retrieval and discovery, and tags such as "Awesome" or "Seen live" may be used to express one's musical taste and opinions about music. Moreover, tags such as "Not metal" may be applied simply to contribute to group knowledge (Lamere, 2008). Although social tags are essentially applied for personal use, and, as such, are subjective by nature, aggregating data from a large number of users helps create valuable information, commonly referred to as a "folksonomy" (Sinclair & Cardew-Hall, 2008). The free-form nature of social tags makes the data a rich source of information. Generated by a large user base and for millions of tracks, Last.fm provides aggregated track-level tag data represented by "counts" of the most popular tags associated to a track.

Because of the free-form nature of social tags, research focusing on social tags needs to deal with several issues that reduce the reliability of the data. Polysemy and synonymy and user error are frequent issues in studies on social tags (Golder & Huberman, 2006). Polysemy refers to the phenomenon where one word can have more than one meaning. For example, the tag "Free" may express that a particular track can be heard for free or may simply relate to the band "Free", and the tag "Blue" may refer to the blues genre, color of the album cover, mood of the listener, or mood of the track. Also, synonymous words (such as "R&B", "Rhythm and blues", "R 'n' B" and "Happy", "Happiness") as well as typographical errors that are regularly found in tag data pose major challenges. The distribution of social tags is highly uneven and many tracks are sparsely tagged or untagged (Levy & Sandler, 2009). The absence of a tag for a track does not therefore necessarily indicate that a tag is not required for the track. On the other hand, tag data are highly biased toward popular tags such as "Rock" and "Pop", which may be overrepresented. Tag scarcity is related to the well-known problem termed as cold-start: as new tracks are created, they remain untagged until discovered, but these untagged tracks are less likely to be discovered. Moreover, the reliability of tag data is reduced by malicious tagging behavior, especially toward artists disliked by the user community (Lamere, 2008). Listeners are also more likely to tag tracks which they like strongly (Marlin, Zemel, Roweis, & Slaney, 2007), and tag data, in general, may be biased toward tastes and opinions of, probably, the young and affluent users who are not representative of the general population (Lamere, 2008).

Despite the issues described above, social tagging provides a rich and extensive source of information about music mood unattainable through listening tests or editorial tagging although exploiting social tags involves a trade-off between quality and quantity (Mandel & Ellis, 2008). MIR research has exploited social tags, for instance, for audio-based tag generation (Eck et al., 2007), determining music similarities (Schedl & Knees, 2009), and making music recommendations (Nanopoulos, Rafailidis, Symeonidis, & Manolopoulos, 2010). Track-level Last.fm tags are also provided for a subset of the Million Song data set (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011) which is, at present, the largest pub-

licly available music data set useful for benchmarking MIR algorithms. Social tags also serve as a promising avenue for MIR research on emotion. MIR research has focused on inferring representations of mood (Levy & Sandler, 2007; Laurier et al., 2009) and training audio-based mood annotation models based on social tags (Sordo, Laurier, & Celma, 2007), but these analyses have been explorative rather than systematic. Recently, basic emotion categories derived from social tags were found to predict the perceived and induced emotion of music tracks measured on the arousal-valence quadrants, but the accuracy was not above the chance level for all emotions (Song, Dixon, Pearce, & Halpern, 2013). In order to better benefit from the vast source of information from social tags about music mood, further studies and new techniques are needed. Appropriate assessment of the accuracy with which social tags represent music mood still requires reliable ground truth, preferably gathered from listening tests.

3.2 Semantic Analysis of Tags

The large number of unique tags in social tag data are correlated to a varying degree. For example, the tags "Alternative" and "Alternative metal" are frequently associated to the same track, whereas "Alternative Metal" and "Classical" are not. Based on this information, the tag "Alternative" may be associated with a track tagged as "Alternative Metal" with some certainty, even if the former has not been applied by users. In contrast, it is highly unlikely that the tag "Classical" applies to a track tagged as "Alternative Metal". The process of inferring the semantic relationships between tags can be automated by semantic analysis of tag co-occurrences using a large collection of tracks or other tagged resources. In particular, semantic analysis provides a means for tackling many of the problems related to social tags.

3.2.1 Techniques

Latent Semantic Analysis (LSA) (Deerwester, Dumais, Furnas, & Landauer, 1990) is a technique widely used to infer semantic information from tag data. LSA mitigates the problems arising with the use of social tags, such as synonymy, user error, and data scarcity, and also increases the robustness of searching and retrieval in large data collections (Levy & Sandler, 2008). To enable computational analysis, tag data are first transformed into the Vector Space Model (VSM) (Salton, Wong, & Yang, 1975), representing associations between documents and tags in a matrix form. Strong dominance of popular tags in the VSM representation is typically countered by normalizing the VSM by Term Frequency–Inverse Document Frequency (TF-IDF) scoring. This reduces the weight given to popularly applied tags and conversely increases the weight of tags associated to fewer tracks. The TF-IDF matrix is then mapped to a lower-dimensional space using low-rank approximation. In LSA, the low-rank approximation is computed by Singular Value

Decomposition (SVD), but other techniques such as Nonnegative Matrix Factorization (NMF) (Seung & Lee, 2001) and Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2001) have been used for this purpose as well. The obtained approximation represents both tracks and tags as mixtures of dimensions related to semantic concepts or topics. Semantic similarities between documents or tags in the low-rank space may be computed using, for example, the Cosine distance measure.

While LSA is the dominant approach followed for semantic analysis, various other techniques have been applied for the analysis of the semantics of tag data. Variants of NMF have been exploited to enhance image tagging (Zhou, Cheung, Qiu, & Xue, 2011), and PLSA has been used for collaborative tagging of websites (Wetzker, Umbrath, & Said, 2009). Peter, Shivapratap, Divya, and Soman (2009) compared the performance of SVD and NMF as a means for low-rank approximation of the LSA matrix in a bibliographic metadata retrieval task but found no significant differences between the two. On the other hand, NMF outperformed SVD and PLSA in the classification of text documents into mood categories (Calvo & Mac Kim, 2012). Garcia-Silva, Corcho, Alani, and Gomez-Perez (2012) reviewed several research papers inferring semantic relationships between tags using clustering techniques, ontology-based techniques, e.g., via Wikipedia[7] searches, and hybrid approaches. Also, user-centered representations have been employed in semantic analysis; these retain the associations between tags and the users that applied them in a tripartite form (term-document-user). For instance, Heymann, Ramage, and Garcia-Molina (2008) included the user as a factor for predicting tags to websites using the Support Vector Machine (SVM) classifier, while (Peng, Zeng, & Huang, 2008) proposed a technique for user-centered collaborative tagging of websites, research papers, and movies. On the other hand, Song, Zhang, and Giles (2011) claimed that the tripartite representation increases the problems related to tag scarcity and resorted to the traditional VSM representation, proposing a graph-based approach to recommend tags to academic papers and websites.

In the MIR domain, Levy and Sandler (2008) employed LSA and PLSA for topic modeling of Last.fm tags and compared these techniques in genre and artist retrieval. Genre ground truth for the artists was obtained from several sources including the Allmusic editorial tags. PLSA outperformed LSA with various model dimensionalities and showed a performance comparable to that obtained using the original VSM. This indicates that semantic analysis that maps tags to low-dimensional semantic space still robustly accounts for relevant information in the higher-dimensional tag data. The topics learnt with PLSA were highly dominated by genres, accounting for 60 topics in a 90-topic model. Only one topic, related to sadness or melancholy, represented the mood unambiguously. A subsequent study (Levy & Sandler, 2009) using the same data showed that including audio features describing the musical content as "tags" in PLSA improved the retrieval performance over the original PLSA model, especially for sparsely tagged tracks. Other previous studies include topic modeling of game-based tags using

[7] http://www.wikipedia.org/

Latent Dirichlet Allocation (LDA) combined with audio-based topic prediction (Law et al., 2010) and modeling the tripartite representation of Last.fm tags using an SVD-based tensor-reduction algorithm to provide user-specific music recommendations (Symeonidis et al., 2008).

3.2.2 Uncovering Structured Representations of Music Mood

Several studies in MIR have employed semantic analysis of tags to determine structured representations of music mood. The analyses of tag similarities emerging from co-occurrence data have yielded semantic spaces resembling both the dimensional and categorical models of emotion. The majority of the studies have either analyzed the relationships between album-level editorial mood labels (Lin et al., 2009; Hu & Downie, 2007) or crawled Last.fm tags for a large set of tracks and analyzed the sub-space of mood-related tags (Levy & Sandler, 2007; Laurier et al., 2009).

Levy and Sandler (2007) retrieved Last.fm tags for 5,000 tracks and selected 57 tags that were related to moods. Applying LSA with 40 dimensions and mapping tags to a 2-dimensional space using the Self-organizing Map (SOM) yielded a space that showed a relationship with the circumplex model of affect. Levy (2012) employed Multidimensional Scaling (MDS) instead of SOM for the same purpose. Laurier et al. (2009) analyzed Last.fm tags associated with over 60,000 tracks and assessed how well the mood representations emerging from the tag data corresponded with expert knowledge and emotion models. First, a reference vocabulary of 120 mood words was created by aggregating word lists from several sources including studies in psychology, music, and emotion, and MIR. The vocabulary was then matched with the tag data, and 80 words were retained. A dimensional representation of mood was inferred using the LSA and SOM, again resulting in a space that resembled the circumplex model. Clustering of tags using the Expectation Maximization (EM) algorithm yielded categories characterized by anger, sadness, tenderness, and happiness, which corresponded well the basic emotion model. Moreover, tag similarities found with the LSA were strongly linked to Hevner's adjective groups (Hevner, 1936) as well as MIREX AMC mood clusters (Hu et al., 2008).

Also, editorial tags, especially those from Allmusic, have been used to create semantic representations of mood. The five mood clusters used in the MIREX AMC were created by clustering Allmusic mood tags associated with albums and songs[8] (Hu & Downie, 2007). Robust mood categories were obtained by subjecting track- and album-level tags to agglomerative hierarchical clustering separately and retaining only those mood tags that were consistently clustered together. Also, Lin et al. (2009) used clustering to infer mood categories from Allmusic tags. This time, tags associated with over 6,000 albums were subjected to spectral clustering, yielding 12 mood categories. However, the goodness of fit of

[8] Mood tags for songs were obtained from "Top lists" associated with each mood. However, these lists are likely to be created by the web service by propagating album-level mood tags according to the "song picks" from each album

the categorization was not evaluated.

The analysis of tag co-occurrences has also enabled investigation of the relationship between mood and genre. Hu and Downie (2007) assessed the statistical significance of associations between mood and genre tag pairs using Allmusic data. The associations for less than 10% of the pairs were statistically significant, so the authors concluded that genre and mood provide different and, to a large degree, independent modes of access to music. The result was further corroborated by the analysis of Last.fm tags. Levy (2012) used other means of statistical analysis for analyzing Last.fm tags and concluded that moods characterized genres to some extent.

Deriving semantic mood spaces from tags can be seen as an alternative to the factor analysis of direct similarity in judgments of emotion terms (Russell, 1980; Scherer, 1984) or self-reports of emotion term co-occurrences (Zentner et al., 2008) on which several of the original emotion models are based. Exploiting large-scale social tag data provides arguably a high ecological validity for this type of analysis, since tag folksonomies emerge in an "organic" fashion (Levy, 2012). However, this approach faces inevitable challenges related to the unreliability of data. Although attempts have been made to evaluate the external validity of the semantic mood spaces by comparing them to emotion models, further studies are needed to assess how robustly music items can be represented in these spaces. In particular, previous studies have neither projected music tracks to the semantic mood spaces nor evaluated the results with reliable data from listening tests.

3.3 Audio-based Annotation

An audio-based music mood annotation system receives a digital audio signal as an input and maps the signal to a specified representation of mood, such as categories or dimensions. This is typically achieved by means of audio feature extraction and machine learning.

3.3.1 Audio Feature Extraction

Audio feature extraction aims to automatically present a complex digital audio signal in the form of a feature as suitable as an input to a machine learning model. MIR studies have applied several approaches to audio feature extraction for audio-based annotation. These can be generally divided into those attempting to model structural characteristics relevant to the target concept (cf. Eerola et al. (2009) for moods), those motivated or inspired by other music-related tasks such as beat tracking (Goto, 2001) or other fields of research such as signal processing and speech perception (e.g., MFCCs (Logan, 2000)), and those applying unsupervised machine learning to infer feature representations (e.g., feature learning using deep neural networks (Lee, Pham, Largman, & Ng, 2009)). In practice, many

methods combine the first two approaches to produce large numbers of features, thus relying on the machine learning stage to identify relevant patterns between the features and the target (Yang et al., 2008; Tzanetakis & Cook, 2002).

The influence of musical structure on the perceived mood has been established already before the dawn of the digital age and the emergence of MIR by analyzing the correlations between listener ratings of the expressed mood and structural characteristics of music or structurally manipulated music pieces and passages (Gabrielsson & Lindström, 2001). For example, Hevner (1937) conducted experiments with several versions of the same piano pieces manipulated in terms of various structural factors and concluded that listener ratings of the expressed mood are affected by musical structure. The strongest effect was found when changes were made to musical tempo and mode, followed by pitch level, harmony, and rhythm. Other structural characteristics found relevant to mood perception, reviewed at length by Gabrielsson and Lindström (2001) include articulation (staccato/legato), harmony (consonant/dissonant), loudness, melodic range, timbre (e.g., soft/sharp), tonality (tonal/atonal/chromatic), and musical form (e.g., complexity, repetition). These types of features have been modeled computationally with various techniques, typically involving spectrogram computation, i.e., the extraction of time-frequency representation of audio, pitch detection (Tolonen & Karjalainen, 2000), chromagram extraction (Pauws, 2004), and onset detection (Bello et al., 2005).

Features developed in the other research fields are also regularly employed in music mood annotation. One of the most frequently adopted features from research on speech recognition are the Mel-Frequency Cepstral Coefficients (MFCCs) (Logan, 2000). MFCCs are computed by grouping and smoothing the magnitude spectrum according to the perceptually motivated Mel-frequency scale and decorrelating the resulting values using the Discrete Cosine Transform (DFT). MFCCs have particularly been employed to model musical timbre (Casey et al., 2008). Various other features relevant to sound perception have been applied to music mood prediction, such as spectral shape features (e.g., centroid, flatness, roll-off, and spread) and spectral dissonance (Yang et al., 2008; Barthet et al., 2012).

The third approach to audio feature extraction, not examined in the present work, employs machine learning to automatically produce features useful for further mapping to the target concepts. This can be achieved by using deep neural networks (Lee et al., 2009) or sparse coding (Blumensath & Davies, 2006) or via simple means of feature clustering (Dieleman & Schrauwen, 2013). The potential of these approaches has been demonstrated for various music-related tasks including auto-tagging (Hamel & Eck, 2010) and recognition of time-varying mood (Schmidt & Kim, 2011).

In audio feature extraction, the signal is usually first cut into short overlapping time frames, and audio features are extracted from each frame. The frame length is determined by the feature type. For example, low-level spectral features are computed from short frames of 23–50 ms, whereas tempo is typically computed from 1–3-second-long frames. Various approaches have been

used to compute the representation of features that are presented as input to a machine-learning model. Clip- or song-level features can be computed by calculating the statistical means and covariances or standard deviations over multiple frames (Mandel & Ellis, 2005). This is the so-called "bag-of-frames" approach (Aucouturier, Defreville, & Pachet, 2007). A variant of this approach was used by Ness, Theocharis, Tzanetakis, and Martins (2009), who computed the running mean and standard deviation over 1-second-long texture frames and represented full songs by computing another set of mean values and standard deviations over the texture frames. Models have also been trained on the original frame-based features, leaving the song-level aggregation for the prediction stage (Turnbull et al., 2008; Coviello, Chan, & Lanckriet, 2011; Miotto & Lanckriet, 2012).

3.3.2 Machine Learning

Machine learning research has a long history, and techniques developed in the field have been applied to almost all domains, ranging from bioinformatics to computer visions (J. Han & Kamber, 2001). The first applications of machine learning to audio-based music annotation included classification of musical instruments (Wold, Blum, Keislar, & Wheaton, 1996) and genres (Tzanetakis & Cook, 2002). Later, the first techniques applied to moods were found to be strongly grounded on similar approaches (Li & Ogihara, 2003). This is reasonable since the same audio features may be used for different tasks, and shifting the focus from genres to mood categories is technically equivalent to replacing the category labels irrelevant to a machine-learning model.

3.3.2.1 Exploiting Listener Ratings

Classification of music tracks into categorical representation of moods has been the dominant approach used in MIR. Li and Ogihara (2003) extracted 30 features related to timbre, rhythm, and pitch and trained the binary SVM models to classify 499 music clips into 13 hand-labeled mood categories. Training and testing the model on 50%–50% splits of the data resulted in accuracies ranging from 50% to 80%, depending on the category. Wieczorkowska, Synak, Lewis, and Ras (2005) analyzed the same data set but used a rather compact grouping of moods into six clusters. The results obtained with the binary (k-NN) classifiers trained on 29 features related to loudness and timbre yielded classification accuracies between 64% and 96%. Feng, Zhuang, and Pan (2003) obtained relatively high recall (0.66) and precision (0.67) values for four basic emotions by employing the binary MLP models trained on tempo and articulation features representing 223 popular music tracks. On the other hand, Lu et al. (2006) developed a hierarchical approach based on the GMM to classify music into four quadrants in Thayer's emotion model. A first-level GMM trained on intensity features was applied to classify tracks into low vs. high arousal, and based on the output, a second-level GMM, trained on timbre and rhythm features, classified tracks into negative and positive valence. The system was evaluated using a set of 800 classical music clips and

it achieved a classification accuracy of 85%. Schmidt et al. (2010) performed the classification using the SVM also into the quadrants of the valence-arousal space. Ground truth associated with 15-second-long clips of 240 popular music tracks was obtained by averaging second-by-second valence-arousal ratings across the clips. On comparing different audio feature sets, the highest accuracy achieved was found to be 50% by combining spectral contrast and MFCC features. Several studies have argued that one mood category is not always enough to represent the mood of a music piece, and, thus, music mood classification may better be formulated as a multi-label classification problem (Wieczorkowska et al., 2006; Trohidis et al., 2008).

Although it is tempting to draw conclusions regarding the overall performance of music mood classification, it is not reasonable to compare systems based on the performance observed using different data sets, ground truth, and evaluation metrics. By analyzing the results of the MIREX AMC contest, a more systematic comparison across systems may be performed and the general picture of the state-of-the-art performances may be obtained. Submitted systems are evaluated with 3-fold cross-validation of a data set of 600 popular music clips divided into 5 mood categories. The year-by-year results show that the mean accuracy has increased from 52% to 60%, with the best accuracy increase ranging from 62% to 70%. However, the increasing accuracy may be attributed partly to the learning from trial-and-error each year, as systems may be specifically developed to perform well on this particular data set rather than on mood classification in general. To date, all high-performing systems in the MIREX AMC have employed SVM, except the year 2007, when the GMM-based system yielded the highest performance.

Common regression techniques employed to predict emotion dimensions include the MLR, PLS regression, and SVR. Typically, models have been trained to predict the valence and arousal for music pieces by training separate regression models for both dimensions. MacDorman and Ho (2007) employed the MLR to predict the valence and arousal ratings for 6-second-long excerpts of 100 songs. The prediction error using features related to spectrum, periodicity, fluctuation, and MFCC was 0.17 z-score for valence and 0.12 z-score for arousal, corresponding to $R^2 > 0.95$, i.e., the proportion of variance in the ratings was explained by the predictions. This prediction rate is extremely high, considering that the reported inter-rater reliability was more than four times higher than that in other studies and that other studies have reported drastically lower prediction rates. For example, Yang et al. (2008) employed the SVR trained on 114 features to predict listener ratings of valence and arousal for 195 popular Western, Chinese, and Japanese songs and obtained $R^2 = 0.28$ for valence and $R^2 = 0.58$ for arousal. Moreover, Korhonen, Clausi, and Jernigan (2006), with regard to performance, obtained R^2 of 0.22 and 0.78 at predicting time-varying valence and arousal, respectively, in 6 classical music pieces using system identification, whereas Yang, Su, Lin, and Chen (2007) employed the SVR and obtained similar rates of 0.17 and 0.80, respectively. It is notable that prediction rates for valence have been consistently lower than those for arousal. Yang and Chen (2012) argued that this is

partly because valence perception is more subjective and elusive in terms of musical characteristics. On the other hand, some studies have also reported higher rates for valence. Eerola et al. (2009) obtained a peak R^2 performance of 0.72 for valence and of 0.85 for arousal in 110 film soundtrack excerpts using PLS regression. Regression models were trained also for other mood scales including tension, happiness, and sadness, yielding a performance between 0.58 and 0.79. The modeling was based on a theoretically selected set of audio features related to timbre, harmony, register, rhythm, articulation, and structure. Regression has been employed also for predicting time-varying valence-arousal positions of music clips (Schmidt et al., 2010; Schmidt, Scott, & Kim, 2012).

Eerola (2011) examined the genre-specificity of moods in music by applying Random Forest regression to predict arousal-valence ratings in various data sets representing classical, film, popular music, and mixed genres; 39 features were extracted from audio tracks, and models were evaluated both within genres and across genres. With regard to valence, for the model performance using tracks from genres used for training (proper cross-validation was applied) the $R^2 = 0.31$ (classical), 0.58 (film music), 0.25 (pop music), and 0.52 (mixed genres). In contrast, the models using genres not used for model training yielded dramatically lower performance rates. Prediction of arousal showed similar patterns although the rates were higher and the performance did not suffer as much across genres. The results indicated that different features and model parameterizations are optimal for different genres in music mood prediction. On the other hand, (Schuller et al., 2010) showed that relying solely on genre information can yield mood prediction performance comparable to that obtained using audio features or lyrics.

Owing to the laboriousness of obtaining listener ratings of music mood, the number of music items available for model training is limited. Model training with limited amount of data easily leads to problems related to overfitting (Jensen & Cohen, 2000), as evidenced by the high prediction performance with training data and the low performance with new data unseen during training. The tendency of a model to overfit increases with a large number of input features (Kohavi & John, 1997). Dimension reduction of the input feature space has therefore been proposed as a way to improve the generalizability of the models. Dimension reduction has been used in audio-based music mood annotation with success (Yang et al., 2008; Eerola et al., 2009; Schuller et al., 2010). Cunningham (2008) distinguished between two forms of dimension reduction: feature selection that selects an appropriate feature subset for model training and feature transformation that maps the original features to a smaller number of latent dimensions. A well-known feature selection technique is the wrapper selection technique that cross-validates a large number of candidate feature subsets using the learning algorithm corresponding to that employed for training the final model (Kohavi & John, 1997). This technique has been employed in MIR for audio-based recognition of moods (Yang, Liu, & Chen, 2006) and genres (Yaslan & Cataltepe, 2006). Wrapper selection is effective since it inherently takes into account the biases of the learning algorithm in order to select the subset with the highest estimated prediction accuracy (John, Kohavi, & Pfleger, 1994). However,

Kohavi and John (1997) showed that wrapper selection itself causes model overfitting owing to the likelihood of the candidate subsets yielding high predictive accuracy, irrespective of the effectiveness of the subsets. It was therefore recommended that evaluation and reporting of wrapper selection be based on an outer loop of cross-validation that ensures that the test data remains unseen by the selection process. Reunanen (2007) argued that the outer loop of cross-validation does not alleviate the problem of the choice of the optimal subset still being, probably, biased, yielding a suboptimal result in terms of generalizability. A proposed meta-algorithm called cross-indexing was found to be able to eliminate this bias and outperform the outer loop of cross-validation in a series of experiments.

3.3.2.2 Exploiting Tags

Exploiting large-scale tag data in model training can help prevent problems related to using limited listener ratings. Audio-based generation of semantic tags for music items, often termed as auto-tagging, has been a major topic in MIR. Although the majority of auto-tagging studies consider tags in general and do not concentrate on mood, equivalent techniques can be applied for multi-label mood classification. A straightforward approach to building auto-tagging systems is to train separate classification models for each tag. This has been done, for example, with social tags (Eck et al., 2007), CAL500 tags (Turnbull et al., 2008; Hoffman, Blei, & Cook, 2009) and game-based tags (Mandel & Ellis, 2008).

Improvements have been made to auto-tagging performance by exploiting relationships between tags. Several studies have employed two-stage approaches, where in the first stage, models are trained separately for each tag, and in the second stage, the model for each tag is trained with the outputs obtained in the first stage (Bertin-Mahieux et al., 2008; Ness et al., 2009; Miotto & Lanckriet, 2012). For example, a second-stage model may improve the performance for a tag "Happy" by combining first-stage models corresponding to "Happy", "Sad", and "Energetic". While these techniques exploit correlations observed in auto-tags, correlations observed in the original tag data have also been exploited. For example, Bertin-Mahieux et al. (2008) reweighed the first-stage auto-tag outputs relative to observed tag co-occurrences. On the basis of empirically observed tag correlations, Yang, Lin, Lee, and Chen (2009) transformed binary tag data to ordinal relevance scores and trained regression models for each tag. Auto-tags were then modified by a second-stage discriminative model, again based on tag correlations.

Techniques that merge audio-based models with other sources of information have also been implemented (Knees, Pohle, Schedl, & Widmer, 2007; Bu et al., 2010; Levy & Sandler, 2009; Miotto & Orio, 2012). Combining different sources has yielded positive results. For example, Turnbull, Barrington, Lanckriet, and Yazdani (2009) predicted CAL500 tags with a combination of auto-tags, social tags, and web-mined text and found statistically significant improvements as compared to using any one of the information sources alone. Notably, an algorithm relying solely on MFCCs outperformed social tags with respect to accuracy.

This rather surprising result may be attributed to the scarcity of social tag data.

Apart from mapping audio features directly to tags, audio features can also be mapped to latent concepts inferred by semantic analysis. Law et al. (2010) employed LDA topic modeling and inferred a small number of topics from game-based tags. Audio-based models first predicted binary topic labels for each track, and then, the LDA model transformed the predicted topics to labels corresponding to the original tags. This approach improved the performance more than the approach using the modeling tags alone.

Music auto-tagging studies concentrating on moods have mostly relied on editorial mood tags from Allmusic (B. Han et al., 2009; Lin et al., 2009, 2011). B. Han et al. (2009) manually organized 11 Allmusic mood tags on an emotion plane derived from Thayer's model and placed tracks on the plane according to the associated mood tags. SVR yielded high performance at predicting the mood positions from audio tracks. Lin et al. (2009) examined the benefit of using genre information in the classification of tracks into 12 mood clusters automatically derived from the Allmusic tags. A two-level mood classification approach was employed, where a genre classifier was applied first, followed by a mood classifier trained on tracks corresponding to the predicted genre. As compared to using a single-level mood classifier, the use of the two-level scheme improved the performance, especially for moods that were the most challenging for the single-level model. Furthermore, consistent improvement was observed by combining several genre-specific mood classifiers according to predicted probabilities of song genres. Similar results were obtained by Lin et al. (2011) using a larger set of mood tags (183 in total) and replacing the genre classifier with a web-crawler of genre tags.

As discussed previously in 3.2.2, semantic analysis of tags has been successful at inferring structured representations of music mood. However, audio-based means to map tracks to such representations has not been studied in the past although progress has been made in that direction. For instance, Wang, Yang, Chang, Wang, and Jeng (2012) showed that both editorial and social mood tags can be mapped to valence-arousal representation on the basis of audio features. However, mapping of tags to the emotion space was not completely based on tags, since it relied on track-level listener ratings of arousal and valence. Moreover, the technique was evaluated only at the tag level, by comparing the mappings to emotion models. Tags provide a large amount of data on music mood, and audio-based techniques combined with semantic analysis enable mood prediction without the need for listener ratings in model training. However, past accounts of audio-based mood prediction exploiting tag data have only evaluated the proposed techniques on tags itself, which do not necessarily represent a reliable ground truth.

4 AIMS OF THE STUDIES

The primary aim of the present work is to facilitate the annotation of large and heterogeneous music collections in terms of mood for the benefit of modern-day music listening. Therefore, the work investigates computational means that incorporate information inferred from large-scale semantic tag data and audio features. Empirical studies aim to predict listener ratings of moods expressed by music by using machine learning based on audio features (I, IV–VI) and by semantic analysis based on social (II–VI) and editorial (III, V) tags. Audio-based machine learning models are trained either with small data sets to directly map audio features to listener ratings (I, IV) or with large data sets by exploiting tag data (IV–VI). The employed information sources and projections between them are presented schematically in Fig. 1.

Previous studies on music mood annotation have frequently employed machine learning to predict listener ratings of music mood based on audio features. However, the laboriousness of collecting reliable listener ratings tends to drastically limit the amount of data available for model training. Model training with a small amount of data, coupled with the use of a large number of input features, increases the risk of overfitting, thereby limiting the generalizability of the

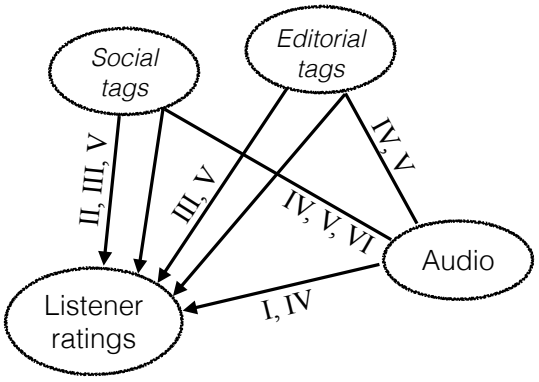


FIGURE 1 An overview of the sources of information employed in the studies (I – VI).

model to the new data. Study I aims to investigate whether the generalizability of audio-based models trained with a small amount of data can be improved by using the wrapper selection process (Kohavi & John, 1997) that prefers models trained with few robust features over those trained with large feature sets. The focus of this study is to optimize the feature subset size employed by a learning algorithm. Three methods applicable for this purpose are compared: outer loop of cross validation (Kohavi & John, 1997), cross-indexing (Reunanen, 2007), and the proposed modified version of cross-indexing. The study also aims to contribute to the understanding of the audio feature combinations related to emotion perception in music.

Studies II–VI examine the robustness of large-scale social and editorial tag data in music mood annotation. Studies II and III aim to improve and evaluate music mood inference solely on the basis of tags associated with music tracks. Tags have been exploited in music mood annotation in previous studies (Lin et al., 2011), but the reliability of such data has not been assessed. Moreover, structured representations of music mood have been identified from social and editorial tags using semantic analysis (Levy & Sandler, 2007; Laurier et al., 2009), but it has not been shown whether the proposed representations facilitate the annotation of music items. Study II aims to fill these gaps in the knowledge by proposing a novel technique called Affective Circumplex Transformation (ACT) that infers the emotion from tags and projects tracks to representations based on associated tags. ACT is evaluated using listener ratings of music mood collected for 600 tracks from different genres and compared to raw tags and other semantic analysis techniques. The study also aims to investigate whether the dimensional or categorical representation better describes the structure of music mood emerging from tag data. Study III aims to evaluate the performance of ACT at representing the mood based on tags obtained from two different sources: *1)* crowd-sourced tags available from Last.fm, and *2)* curated editorial annotations used in a production music catalogue. Moreover, the study seeks to assess the width of the gap between the semantic representations of mood from these two data sources by applying semantic models across the corpora. ACT is applied for semantic modeling, and the evaluation is conducted using listener ratings from Study II and another set of ratings collected for 205 production music tracks.

The rest of the studies (IV–VI) aim to demonstrate how large-scale tag data can be exploited to improve audio-based music mood prediction. Previous studies have proposed audio-based techniques for the generation of semantic tags (Turnbull et al., 2008; Eck et al., 2007), but these techniques have either been developed specifically to predict moods or evaluated using reliable listener rating data. Moreover, the benefit of incorporating semantic analysis of tags into audio-based mood prediction has not been assessed. Study IV proposes and evaluates a novel technique termed as Semantic Layer Projection (SLP) that maps audio features to a semantic mood space and then maps these obtained estimates to listener ratings. SLP performance is compared to conventional methods that map the audio directly to listener ratings. On the basis of the evaluations in Studies II and III, the semantic mood space is inferred using ACT, exploiting social tags

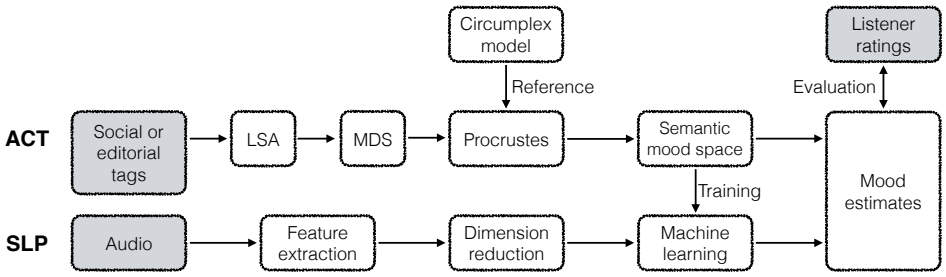


FIGURE 2 An overview of ACT and SLP processing stages and the sources of information (highlighted in gray).

as the source of semantic information. While Study IV exploits listener ratings in the training for the second stage of mapping from the semantic space, Study V evaluates a variant of SLP that does not require listener ratings in the model training. Study V also evaluates SLP using both social tags and editorial tags and compares the tags and audio tracks as model inputs for novel tracks. Fig. 2 outlines the processing stages and sources of information involved in the ACT and the SLP techniques. The techniques are briefly explained in the next section and detailed in the included studies.

Previous studies suggest that using different audio features may be an optimal technique for predicting valence and arousal in different genres (Eerola, 2011), and adapting audio-based auto-tagging models to genres is beneficial (Lin et al., 2009, 2011). Even using genre information directly as input for mood prediction has been found to be an effective strategy (Schuller et al., 2010), indicating the fact that the relevance of moods differs between genres (Hu & Downie, 2007). However, previous studies have not tested whether adapting semantic mood models and subsequent audio-based models to different genres would be beneficial for mood prediction; moreover, genre information has not been shown to improve the prediction of listener ratings of emotion dimensions. Study VI aims to fill these gaps in the knowledge by adapting both the semantic and audio-based models to different genres.

5 MATERIALS, METHODS, AND RESULTS

This chapter presents a summary of the materials, methods, and findings of Studies I–VI (Please refer to the original publications for details). All studies evaluated mood annotation performance primarily using listener ratings collected from pools of participants. The focus of all ratings was on the expressed mood. Study I took listener ratings and audio materials from a previous study by Eerola and Vuoskoski (2011), whereas Studies II–VI used listener ratings collected in the present work (Studies II and III). All audio features in Studies I and IV–VI were extracted using MIRtoolbox (Lartillot & Toiviainen, 2007).

5.1 Improving the Generalizability using Feature Selection (I)

Study I used wrapper selection as a means for feature subset selection and compared different methods to optimize the feature subset size for different classifiers. Three methods were compared: the outer loop of cross-validation (Kohavi & John, 1997), cross-indexing (Reunanen, 2007), and a proposed modified version of cross-indexing that introduces a parameter that inherently favors subsets with a smaller number of features in subset size optimization.

The experiments were run on two data sets of film soundtrack excerpts taken from listening tests conducted by Eerola and Vuoskoski (2011): *1)* a smaller set of excerpts rated by 116 non-musicians was used as the primary set for feature selection, and *2)* a larger set rated by 12 expert musicologists was used for external validation of the feature selection performance. The listener ratings, originally given in terms of basic emotions on seven-point Likert scales, were transformed to classes according to the highest-rated emotion for each excerpt. Subsets of the data sets were selected in order to avoid emotionally ambiguous excerpts and obtain an equal number of excerpts for each class. Moreover, anger and fear were grouped as one category owing to their high correlation. The resulting primary and validation sets consisted of 64 and 160 excerpts, respectively. Audio features related to dynamics, rhythm, pitch, harmony, timbre, and structure were

extracted in a frame-based manner, and each excerpt was finally represented by a vector of 66 features corresponding to the mean and standard deviation of each feature over the excerpt lengths.

The experiment was run using three well-known classifiers: the Naive Bayes, k-NN, and SVM. The hypothesis was that the performance of k-NN and Naive Bayes would be significantly improved by wrapper selection since they are highly influenced by redundancy and irrelevance of features, whereas SVM would provide an example of a more sophisticated classifier less prone to overfitting. Two search methods were compared for feature selection: forward selection (FS) and backward elimination (BE) that involve a stepwise addition (FS) or removal (BE) of one feature at a time, starting from the empty (FS) or full (BE) feature set. The wrapper selection process was run with each classifier – a search method combination – until the full or empty feature set was reached, depending on the search method. For the wrapper selection, the primary set was split into a number of 50% training–50% test folds. The selection process was run on each training fold, and the performance of each successive subset was evaluated on the corresponding test folds and the validation set. The optimal subset size was then estimated based on the test fold performances using the three optimization methods. Finally, the methods were compared by examining how well they could predict the optimal subset size and classification rate for the validation set.

The estimates obtained using the modified cross-indexing estimates corresponded fairly well with the optimal subset sizes and classification rates observed in the validation set and outperformed the original cross-indexing. Moreover, the results showed that the outer loop of cross-validation provides inaccurate estimates of the optimal subset size and performance of the classifiers. In general, smaller subsets (with less than ten features) led to better generalization of the models than larger subsets. This pattern was insufficiently uncovered using the outer loop of cross validation and the original cross-indexing algorithm. When taking into account the classification rates and the simplicity of the optimized models, the k-NN with BE yielded the most promising results: 56.5% classification accuracy with only 4 features. The most useful feature subsets for k-NN included the majorness of mode and key clarity in combination with dynamical, rhythmical, and structural features.

5.2 Incorporating Tag Data as Information Sources (II–III)

Studies II and III examined how accurately moods expressed by music can be predicted on the basis of social and editorial tags. In particular, the studies investigated whether semantic analysis based on tag data to infer structured representations of mood can improve tag-based mood prediction. Study II focused on social tag data and crawled a substantially large data set of track-level tags using the Last.fm API. In order to build a data set balanced in terms of moods as well as genres, vocabularies consisting of 568 mood terms and 864 genre terms were

used as queries for data crawling. These vocabularies were gathered in a similar manner as in the study by (Laurier et al., 2009) by aggregating term lists from different research papers and online sources. The final set comprised 1,338,463 tracks and 924,230 unique tags. After a series of data cleaning procedures, mood-related data included 259,593 tracks and 357 terms. Partitions of the data set were formed similarly for genres, instruments, opinions, and locations.

LSA (Deerwester et al., 1990) followed by MDS was applied to the mood data to map tags to a three-dimensional space reflecting the semantic similarities between tags. The choice of the dimensionality was based on previous studies on emotion modeling suggesting that emotions would be explained by two to three underlying dimensions (Russell, 1980; Schimmack & Grob, 2000; Zentner et al., 2008). Tracks were then projected to the space on the basis of associated tags. To examine whether the dimensional or categorical representation of mood provides a better fit to the data, the clusterability of tracks in the mood space was analyzed using the Hopkins' index (Hopkins & Skellam, 1954). For comparison, the same analysis was conducted also with the data related to genres, instruments, opinions, and locations. The Hopkins' index for moods remained at a range of 0.6–0.7, supporting the fact that tracks are distributed continuously rather than categorically in the mood space. In contrast, the index for the other concepts was consistently higher, i.e., 0.7–0.95. In particular, this finding supported the common practice of assigning songs categorically to genres.

The abovementioned typical LSA and MDS procedures may not be adequate for characterizing moods since the obtained dimensions do not explicitly represent the dimensional model of emotion. Therefore, a novel technique, ACT, based on Russell's circumplex model of emotion (Russell, 1980) was proposed to conceptualize the dimensions of the MDS mood space. This technique involved a mapping process to conform the MDS space to the circumplex model on the basis of reference positions for mood terms on the valence-arousal dimensions obtained from previous studies by Russell (1980) and Scherer (1984). This was achieved using classical Procrustes analysis (Gower & Dijksterhuis, 2004). Finally, music items could be projected to the resulting space based on the associated tags. In particular, the track positions along the first and second dimensions in the space represent valence and arousal, respectively. A listening experiment was conducted to evaluate ACT performance at track-level mood prediction. An evaluation set of 600 tracks, not overlapping with the already analyzed data, was obtained from Last.fm. This set was sampled in a balanced manner to cover the semantic mood space as well as the six main genres: electronic, folk, jazz, metal, pop, and rock. The listening experiment involved a total of 59 participants who were asked to rate 15-second-long clips of the tracks on 9-point Likert scales. The three core affect dimensions of valence (negative–positive), arousal (calm–energetic), tension (relaxed–tense) were rated in terms of bipolar scales, and seven mood terms (atmospheric, happy, dark, sad, angry, sensual, and sentimental) were rated on unipolar scales. For further analysis, the ratings were aggregated by calculating the mean value for ratings provided by all participants. The ACT's ability to predict the listener ratings based on tag data associated with

the evaluation set was then evaluated, and for comparison, several conventional tag-based techniques including SVD, NMF, PLSA, as well as the raw tag data were employed.

ACT outperformed the baseline techniques consistently, regardless of the number of dimensions used in the LSA stage. Moreover, ACT and all semantic analysis techniques outperformed the raw tag data, supporting the use of semantic analysis in tag-based music mood prediction. The median correlation between the ACT estimates and listener ratings was 0.58 for valence and 0.64 for arousal. For the other mood scales, the overall performance level of all examined techniques was slightly lower. Further analysis showed that ACT performance is robust even if there is a scarcity of track-level mood tags. In general, the results suggested that significant performance improvements can be made by representing the moods of music tracks in an interpretable and robust fashion based on semantic computing of social tags and research on emotion modeling.

Study III examined the performance of ACT across the corpora of curated editorial tags associated with production music tracks and the data from Study II. A corpus of 226,344 production music tracks was extracted from I Like Music's (ILM) collection that aggregates 29 individual production music catalogues; 288 mood terms could be identified from the ILM corpus. ACT models were then trained separately with social and editorial tags, and these models were applied to predict listener ratings in two evaluation sets: one collected in Study I, and the other collected for a subset of the ILM corpus. For the listening experiment, a set of 205 tracks was sampled from the ILM corpus, again in a balanced manner, to sufficiently cover different moods and genres. The ratings were collected from 46 participants for the 3 core affect dimensions and additionally for 3 bipolar scales related to dominance (submissive/dominant), romance (cold/romantic), and humor (serious/funny). The evaluations however focused on the core affects.

ACT models trained with editorial tags outperformed the models trained with social tags at predicting the listener ratings from Study II. This result was partly expected since curated editorial tags are considered more reliable than crowd-sourced social tags but, nevertheless, surprising owing to the difference between the musical material in the two corpora. Unsurprisingly, when tested with the listener ratings related to production music, ACT models trained with editorial tags again outperformed those trained with social tags. However, the performance of the models trained with social tags did not suffer when evaluated with production music tracks, except in the case of arousal. In general, these results showed that semantic models of moods obtained using ACT could be generalized across tag types and musical material.

5.3 Exploiting Tag Data for Audio-based Annotation (IV–VI)

Studies IV–VI focused on improving audio-based music mood annotation by exploiting large-scale tag data. Study IV sought to demonstrate whether an audio-

based model exploiting a large set of audio and tags improves the prediction of listener ratings more than a model trained directly to map audio features to the ratings. To achieve this aim, tag data and listener ratings were taken from Study II, and an experiment was conducted to predict the listener ratings of valence and arousal. A novel technique, i.e., SLP, derived from ACT was proposed for the tag-based prediction. SLP involves mapping audio features (audio level) to a semantic mood space inferred using ACT (semantic layer) first and then mapping the semantic mood space to listener ratings (perceptual level). The idea behind using a semantic layer rather than raw tags as an intermediate representation was to make use of the benefit of the high performance of ACT at representing the moods of music tracks.

The ACT model was trained in the same manner as in Study II to create the semantic layer. However, the semantic layer in this study was represented by ten dimensions rather than three dimensions. A subset of 9,662 tracks was then sampled from the full training corpus for the audio-based modeling, and 15–30-second-long preview audio clips were obtained from Last.fm. A total of 128 audio features were extracted to represent the clips in a similar manner as in Study I. Regression models were then trained to map audio features separately to each semantic layer dimension using the PLS regression, bearing in mind its effectiveness in previous studies (Eerola et al., 2009). The resulting mappings were then applied to the evaluation set of 600 tracks. Within the evaluation set, linear regression models were trained to map the track estimates on the semantic layer to the listener ratings. Two inputs for the linear regression were compared: (1) track estimates along all dimensions together, and (2) track estimates along separate dimensions corresponding to valence (1st dimension) and arousal (2nd dimension). The former input would result in exploitation of tag data merely as a means of dimension reduction of audio features, whereas the latter would produce estimates of the listener ratings without adapting to the rating data (only the overall range of the estimates would need to be adjusted). SLP was compared to two baseline techniques that mapped audio features directly to the listener ratings: PLS and SVR. These techniques were chosen since they were found to be successful in previous MIR studies, e.g. (Yang et al., 2008; Eerola et al., 2009; B. Han et al., 2009).

To obtain the final performance, 2-fold cross-validation was run 50 times within the evaluation set. With regard to valence, SLP using all semantic layer dimensions produced, by far, the highest performance of $R^2 = 0.34$, outperforming the other SLP variant ($R^2 = 0.25$) as well as the baseline techniques ($R^2 = 0.15$ for PLS and $R^2 = 0.25$ for SVR). This showed the efficiency of exploiting tag data in audio-based modeling. A notable result was that the SLP variant that modeled the ratings directly with the semantic layer dimensions outperformed (although slightly) the techniques that carried out complex model training to adapt to the ratings. The overall performance level was higher for arousal than for valence, consistent with the findings of previous studies (Yang & Chen, 2012). SLP (all dimensions) again showed the highest performance ($R^2 = 0.78$), but this time outperformed SVR only slightly. The lowest performance was obtained by SLP

with a single semantic layer dimension ($R^2 = 0.75$). However, this performance was still promising considering that no adaptation to listener ratings was performed[1]. Finally, SLP performance using different subsets of audio features was examined to determine the type of features that would be the most useful ones for modeling valence and arousal. Harmony-related features were found to be the most useful ones for modeling valence ($R^2 = 0.19$), supporting the findings from Study I, whereas timbral features representing characteristics of the audio spectrogram were the most useful ones for arousal ($R^2 = 0.687$). In general, the results proved the usefulness of exploiting tag data in audio-based music mood prediction and highlighted the difficulty in modeling the valence dimension in music.

Study V carried out further exploration with the SLP technique. In comparison to Study IV, in this study, the ILM data set and listener ratings from Study III were employed, and performance was examined for all rated mood scales. The SLP variant 2) from Study IV that produces mood estimates without the need to adapt to listener ratings was employed since it provided promising results and could be directly compared to tag-based prediction using ACT. Instead of cross-validation, the R^2 performance was obtained by computing the squared correlations between the estimates and the ratings. Another difference in Study IV was that SLP involved a three-dimensional rather than ten-dimensional semantic layer. This way both of these techniques would employ the same semantic mood representation and thus enable fair performance comparison. Two experiments were carried out: (1) comparison of SLP to baseline techniques that map audio features to each tag separately, and (2) comparison of audio-based prediction using SLP and tag-based prediction using ACT. In the first experiment, SLP outperformed the baseline techniques by a clear margin using both Last.fm and ILM data, showing the efficiency of mapping audio features to the semantic layer rather than to the tags. Concerning the Last.fm data, using the three-dimensional rather than ten-dimensional semantic layer in SLP improved the performance for valence ($R^2 = 0.32$), but decreased the performance for arousal ($R^2 = 0.71$). The other mood scales that were strongly associated with valence, e.g., happy, sad, and dark, were the most difficult to predict. The prediction rate for valence was considerably higher ($R^2 = 0.49$) with the ILM data than with the Last.fm data, probably because the aim of production music is to provide clear structural cues conveying positive or negative emotion.

The second experiment allowed direct comparison of prediction performance achieved when using audio and tags as information sources for novel tracks. Audio-based prediction using SLP consistently outperformed tag-based prediction using ACT; ACT outperformed SLP only for valence and happy with the Last.fm data. The most radical performance difference was observed for arousal, for which the audio-based prediction improved tag-based prediction from $R^2 = 0.42$ (Last.fm) and 0.50 (ILM) to over 0.70. Also, the performance of the combination of ACT and SLP was examined. The performance showed notable improve-

[1] The performance is actually surprisingly high as compared to the tag-based prediction of arousal with ACT in Study II. This difference was further examined in Study V.

ments, especially for valence. However, giving more weight to SLP estimates led to higher results for all moods, suggesting that audio and tags can be used effectively as complementary information sources of music mood.

The rather surprising results related to the performance difference in SLP and ACT lead to two obvious questions: First, how can the tag data that provides direct human-labeled semantic information be inferior to audio features? Second, since SLP models were originally trained to map audio to ACT dimensions, how could they yield higher prediction performance than the actual dimensions? The scarcity of tag data coupled with their inherent unreliability may provide answers to both of these questions. Although ACT alleviates the problems related to tag scarcity, tracks are still mapped to the semantic space based only on a few tags, causing local inconsistencies. In contrast, mapping audio features to ACT dimensions using SLP may tap into more global patterns and provide a way to "smooth out" these inconsistencies. Nevertheless, these results have positive implications for music mood annotation since they indicate that human-generated labels are not necessarily required for efficient mood inference of novel tracks.

Study VI examined whether the audio-based prediction exploiting tags could be improved by taking into account the genre of music tracks. Although audio-based prediction using SLP was found to be efficient in Studies IV and V, the results showed that there was room for improvement, especially for the valence dimension. A novel genre-adaptive technique called ACT+SLPwg was employed in the study: First, a number of genre-specific SLP models were trained separately using tracks associated with different genres, and these models were applied to novel tracks according to the associated genres. By allowing the semantic mood space to adapt to different genres, the technique took genre-adaptivity beyond the previous approaches presented by Lin et al. (2009) and Lin et al. (2011). This technique was compared to the general non-genre-adaptive SLP techniques and several variants of SLP that exploited genre information. Performance evaluation was carried out using Last.fm social tag data, comprising 100 mood and genre tags, and the listener ratings from Study II associated with the Last.fm evaluation set.

Genre information of tracks was represented by a weighted combination of genre clusters inferred from the social tag data. A genre clustering survey was conducted to evaluate different clustering techniques for this task and to determine the number of genre clusters needed to represent the tag data. In the survey, participants were asked to arrange the 100 genre tags into a number of groups using an online interface. The ability of three conventional clustering techniques, in terms of the Mirkin metric (Mirkin, 1996), to produce the human-generated clusters automatically based on tag data was compared: K-means, Agglomerative hierarchical clustering, and Spectral clustering. Based on the evaluation, the K-means technique was chosen for further mood prediction analyses. The survey did not yield a clear optimal number of genre clusters. However, K-means with six clusters produced a result resembling the main genres in the evaluation set. Therefore, the subsequent mood prediction analyses were primarily conducted using this clustering.

In the first mood prediction experiment, various general models not exploiting genre information were considered. First, tag-based mood prediction was performed using ACT with three alternative mood term reference configurations. This was done to obtain a tag-based performance reference and to optimize the mappings from mood tags to the semantic mood space for the subsequent audio-based analyses. In this evaluation, a reference configuration consisting of a subset of mood terms from Study II (happy, calm, sad, and angry) outperformed the original configuration. Furthermore, mood space inferred directly from human-labeled normative data obtained from Warriner and Brysbaert (2013) did not yield convincing results. This supported the exploiting of music-specific tag data in forming the semantic mood space rather than using a mood space that describes affective connotations of mood words in general. In the audio-based mood prediction analysis, SLP yielded a prediction performance of $R^2 = 0.36, 0.73$, and 0.49 for the core affects valence, arousal, and tension, respectively. The performance level was found to be more favorable than that of stacked SVM classifiers that had been found to be efficient in previous studies on music auto-tagging (Ness et al., 2009).

The second mood prediction experiment compared ACT+SLPwg to different genre-adaptive techniques and to the general models not exploiting genre information. The genre-adaptive techniques exploited genre either directly as input features to SLP or trained a collection of mood prediction models within different genres. Moreover, genre information for novel tracks was either derived from tag data or predicted from audio. ACT+SLPwg showed the highest performance using both the tag- and audio-based genres. In particular, ACT+SLP showed more improvements than general SLP and outperformed the technique that did not involve genre-adaptive modeling of the semantic mood space. The highest performing technique overall was ACT+SLPwg using audio-based genres, yielding a statistically significant improvement as opposed to general SLP for all core affects ($R^2 = 0.43, 0.74$, and 0.52 for valence, arousal, and tension, respectively). Further analysis showed that the performance of ACT+SLPwg was not sensitive to the number of genre clusters.

6 CONCLUSIONS

The ability of music to express moods is one of the main reasons why people are attracted to music listening. Annotation of music in terms of the expressed mood is therefore useful for various purposes, such as music catalogue organization, music recommendation, mood regulation, and research on cognitive disorders, to name a few. The present work aimed to improve techniques to annotate music according to the expressed mood by employing semantic computing based on social and editorial tag data and machine learning based on audio features.

Social and editorial tags are abundantly and increasingly available for millions of music tracks, and a considerable proportion of the tag data is related to mood. However, the findings of Study II suggested that tag-based annotation is not adequate as such to represent moods of music tracks. As a solution, semantic computing applied to tag data to infer a mood representation, resembling the well-known dimensional model of emotion, led to significant improvements to the annotation accuracy. The proposed ACT technique performed well at representing moods of music tracks based on social tags when compared to other semantic analysis techniques (Study II). Moreover, ACT models were generalizable across tag data types (social vs. editorial tags) and musical corpora (popular vs. production music) (Study III). These results have positive implications for the use of various sources of semantic information about music to represent mood. Use of computational means to improve the accuracy of these annotations is arguably the most economical approach to improve the quality of the data since additional manual labeling is not required in this case. In the present work, only tag data were studied, but similar approaches could be applied to other data as well, such as microblogs, lyrics, and web-mined text.

Owing to the sheer pace of the emergence of new music, all of the world's music is not, and will never be, labeled by editors or tagged by online communities in terms of mood. The benefit of using audio-based annotation is that it does not rely on human-generated labels of new music items, so these audio-based techniques are more applicable to music in general. In previous studies, performance improvements have been made by applying audio feature selection as a pre-processing step prior to training the final predictive models. However,

according to Study I, carrying out automatic feature selection using typical cross-validation procedures leads to suboptimal results, and the obtained models cannot be generalized to unknown data. Clear improvements in this approach were achieved using the proposed cross-indexing procedure that optimized the feature subset size by favoring small subsets that simplified the training data. Concerning the examined models employing audio features, such as an input to music mood classification, cross-indexing increased the accuracy of the optimal subset size and classification performance estimates. Although cross-indexing was applied to optimize the feature subset size, in principle, it can be applied to any model parameter optimization task. For example, the number of components in the PLS regression models was optimized using cross-indexing in Studies III and IV. Examining the influence of cross-indexing on prediction performance was, however, out of the scope of these studies.

While Study I exploited a limited amount of listener ratings as the ground truth for model training, Studies IV–VI sought to increase the amount of training data available by exploiting large-scale social and editorial tag data. The results of Study IV showed that a model using a semantic mood space inferred from social tags as an intermediate layer, to which audio features are mapped, can outperform models that are trained to directly map audio to listener ratings. Overall, the proposed SLP technique yielded a relatively high performance. However, the performance for valence remained at a low level, in line with the findings of previous studies, indicating that audio-based prediction of valence is more challenging than that of arousal. These results, i.e., the efficiency of SLP and the low prediction accuracy for valence, were further generalized by subjecting the editorial tag data to SLP in Study V. These results indicate that the trade-off between (1) the amount of data available when exploiting tags and (2) the reliability of training data when exploiting listener ratings is not crucial if the tag data are processed using appropriate semantic computing techniques.

Study VI aimed to improve the performance of audio-based mood prediction by exploiting genre information. The novel technique ACT+SLPwg produced the most accurate predictions by adapting both the semantic mood space and audio-based models to different genres. In particular, as compared to models not exploiting genre information, ACT+SLPwg largely improved the prediction of valence, suggesting that semantic relationships between mood tags are not static across different genres of music and that audio-based modeling benefits from taking into account these genre-specific aspects. The results for valence indicate different ways in which positive and negative aspects of moods are conceptualized in different musical genres. The study also showed that inferring a semantic mood space in a bottom-up manner from music-specific tag data was more efficient in producing music mood representations than the space built on the basis of normative data collected in studies in the field of affective sciences. This result indicates that although the emotion models grounded on everyday emotions can be applied as the general framework to represent music mood, music-specific data are needed to produce robust music mood representations.

6.1 Methodological Considerations

Computational modeling of music mood is such a wide topic that many important methodological aspects were not addressed in the present work. The first consideration is related to the listener ratings on the basis of which a major proportion of the evaluation was carried out. Although the ratings were collected on clearly defined mood scales, the choice of mood term-based scales might have caused a bias in the evaluation. In Study II, the ratings were collected only for seven terms, all related to the most typical moods expressed by music. If the proposed techniques were to be evaluated using more uncommon terms, such as "dramatic", "quirky", or "bittersweet", perhaps, the low-dimensional mood model would not have been as efficient as a representative as the studies indicated. Although the low-dimensional models have been widely supported in music research, it could have been more adequate to use higher dimensional representations such as the GEMS (Zentner et al., 2008) as the basis for mood modeling. Moreover, the choice of collecting responses on Likert scales might have exaggerated the fit of the dimensional model to the rating data. Therefore, it would be useful to cross-validate the listener ratings using other self-report methods such as adjective checklists, open-ended questionnaires, or non-verbal measures (e.g., assessing the similarity in music clips in terms of mood).

The second methodological consideration is related to the music corpora. The corpora employed in Studies II–VI consisted of sampled tracks associated with both mood and genre tags, which led to the omission of a large proportion of the original data. Since the number of tags associated with music items was arguably correlated with popularity, this undoubtedly caused a bias toward more popular tracks not representative of the whole corpora. Further analysis would be needed to determine how the proposed techniques would perform with music that was methodologically disregarded from the analysis. As shown in Study II, performance of tag-based mood annotation is influenced by the scarcity of tag data. One way to increase the model performance with sparsely tagged music items would be to complement the tag data with textual data from other sources, such as web-mined text or lyrics. Further research is thus needed to assess such options. Another possible bias caused by the choice of music corpora is related to the included musical genres. As the studies focused on Western popular music, and, in particular, on music that is popular on online music services, the analysis largely ignored classical and non-Western music. For example, instead of considering different popular music genres in the genre-adaptive analysis, a plausible genre configuration could have been obtained by distinguishing classical music, popular music, and non-western, non-popular music. It remains to be seen how the proposed genre-adaptive technique would perform with data showing such a wide variety of musical and cultural characteristics.

The third consideration is related to the semantic modeling techniques. Smoothing out noise in the tag data is considered, in general, a favorable feature of semantic modeling, but this also causes the loss of inherent nonlinearities or ir-

regularities that might be conceptually important characteristics. Disregarding these characteristics as noise reduces the granularity at which moods are modeled. Also, the use of low-dimensional semantic mood space may have exaggerated the semantic similarity of certain mood terms, such as "intense" and "dark" or "sleepy" and "peaceful." Higher-dimensional semantic models or nonlinear techniques might be more suitable for dealing with these issues. Moreover, the use of existing ontologies such as the Wordnet (Fellbaum, 1998) to infer semantic knowledge about tags was left unexplored. This avenue could provide a potential alternative to deal with semantic data.

The fourth consideration is related to content-based analysis. The performance of audio-based techniques is highly dependent on the quality of the extracted audio features. Prediction errors in audio-based mood annotation can be partly attributed to the inaccuracy of audio features at capturing perceptually meaningful features of music. For example, audio features modeling particular musical features such as mode or rhythmic clarity have been inconsistent in terms of listener evaluations of the target features (Friberg, Schoonderwaldt, Hedblad, Fabiani, & Elowsson, 2014). Development and fine-tuning of the audio feature extraction is therefore an effective way to enhance the audio-based mood prediction accuracy. Moreover, an avenue that was left unexplored in the content-based analysis was the use of lyrics as a source of information for mood prediction. Lyrics can either detract or enhance the emotions perceived in music (Ali & Peynircioğlu, 2006), and studies predicting mood based on a combination of lyrics and audio have yielded positive results (Kim et al., 2010) although disputing evidence has also been presented (Schuller, Weninger, & Dorfner, 2011). Acknowledging the multi-modal nature of music-related emotional expression could be another way to improve the performance of the proposed techniques.

6.2 Future Directions

A major benefit of the explored mood modeling approach is that the produced models can be adapted to real-world data produced by everyday music listening activities. This capability also entails flexible modification of the models as music culture evolves, new genres emerge, and online user communities change their tagging behavior. Application of the proposed techniques is by no means restricted to the music domain. The techniques could be adapted to any multimedia content such as films and books as well as to online shopping and social networking. Semantic computing methods will be increasingly useful as online activity grows and the smallest details of our daily lives are tracked. For example, linking various modalities and sources of information could produce a system that facilitates safe traffic by tracking drivers' mood and health based on music listening, recently read news articles and purchased items, heart-rate changes and affective qualities of recent social interactions.

All tag and audio data employed in the present work were obtained from

proprietary sources, so the practical use of the obtained data is restricted by copyright licensing. Restrictions are also applicable to the evaluation data: while the collected listener ratings can be freely disseminated to researchers, the related audio is copyright protected. This is a common problem faced during the process of conducting reproducible research in the MIR field. A potential future direction is to circumvent copyright issues by examining how efficient mood annotation models could be obtained by relying on open-access music catalogues and web-mined textual data.

Another future direction is to develop models that take into account the subjective aspects and situational factors of music mood perception. As noted earlier, listener's mood, personality, musical preferences, and listening context play a significant role in how moods are perceived (Vuoskoski & Eerola, 2011; Scherer & Zentner, 2001). Personalized music mood annotation systems are therefore highly needed, and some advances have already been made toward this aim (Yang & Chen, 2012). In the simplest form, rather than dealing with the averaged listener ratings as done in the present work, ratings of each listener could be modeled separately, and data from a listener background questionnaire could be employed to assess personal differences. For example, mood prediction models trained with one genre could be particularly applicable to predicting the ratings of the listeners that are most familiar with that genre. A more ambitious direction would be to retain the user information of the tag data and infer personal and situational factors for each tag from the users' listening profiles and linked data related to social networking, activities, and events attended. Modeling all these aspects together would produce personalized mood models, and these models could be validated by behavioral experiments. Methodologically, this type of research would be positioned in the wider context of opinion mining and sentiment analysis (Pang & Lee, 2008). Also, assessing the moods induced by music would be beneficial for this approach in order to gain a more complete picture of the affective phenomena related to music listening.

TIIVISTELMÄ

Musiikin ilmaisemien tunnetilojen mallinnus käyttäen semanttisen laskennan ja koneoppimisen menetelmiä

Musiikin vetovoima ja merkitys perustuvat vahvasti musiikin kykyyn ilmaista tunnetiloja. Siksi on tarpeellista kehittää menetelmiä, jotka mahdollistavat musiikin löytämisen musiikin ilmaisemiin tunnetiloihin perustuen. Nykypäivän verkkopohjaiset musiikkipalvelut tarjoavat käyttäjilleen kuunneltavaksi miljoonia kappaleita käsittäviä kokoelmia, ja useat musiikkipalvelut hyödyntävät sosiaalisten verkkoyhteisöjen tai musiikkitoimittajien tuottamia merkintöjä, niin sanottuja semanttisia tageja. Tunnetiloja kuvaavat tagit eivät kuitenkaan ole yhtä luotettavia kuin kuuntelukokeissa perinteisin menetelmin kerätyt tunnearviot, koska tagit sisältävät niiden käyttötavoista johtuen monenlaisia epätarkkuuksia. Semanttisen laskennan menetelmillä on mahdollista tehostaa tagien perusteella tehtäviä päätelmiä. Myös yleisesti käytetyillä sisältöpohjaisilla tunnistusmenetelmillä kyetään arvioimaan tunnetiloja digitaalisesta musiikkitiedostosta laskennallisesti irrotettuihin audio-piirteisiin sekä koneoppimiseen perustuen.

Tämän väitöskirjan päätavoitteena oli tarkastella ja kehittää laskennallisia menetelmiä, joiden avulla voidaan hyödyntää musiikin tunnetageja sekä annotoida musiikkikappaleita automaattisesti musiikin ilmaisemilla tunnetiloilla aiempaa tehokkaammin. Työssä kehitettiin uusi semanttisen laskennan menetelmä nimeltään Affective Circumplex Transformation (ACT), jolla pyrittiin parantamaan tunnetagien tarkkuutta aiempiin menetelmiin verrattuna peilaamalla tagiaineisto emootiopsykologiasta tuttuun dimensionaaliseen emootiomalliin. Työssä pyrittiin myös parantamaan audio-piirteisiin pohjautuvien tunnetila-arvioiden yleistettävyyttä käyttäen piirteiden valintamenetelmiä sekä semanttista laskentaa. Tähän tarkoitukseen kehitettiin työssä uusi menetelmä nimeltään Semantic layer Projection (SLP), joka hyödyntää laajaa tagiaineistoa ACT-mallien avulla. Lisäksi työssä tutkittiin voidaanko audio-pohjaisen tunnetilojen tunnistuksen tarkkuutta parantaa mukauttamalla malleja eri musiikkityylien ominaispiirteisiin.

Väitöskirjan pohja-aineistona käytettiin mittavia populaari- ja tuotantomusiikkia sisältäviä musiikkikokoelmia, jotka yhdistettiin sosiaalisiin ja toimituksellisiin tageihin sekä audio-tiedostoihin. Menetelmien arviointi suoritettiin empiirisesti kuuntelukokeissa. Tulokset paljastivat, että ACT-menetelmän tuottamat tunnearviot ovat ylivertaisia verrattuna alkuperäiseen tagitietoon sekä perinteisiin semanttisen laskennan menetelmien tuottamiin arvioihin. Sosiaalisten ja toimituksellisten tagien pohjalta muodostettujen ACT-mallien havaittiin lisäksi olevan hyvin yhtenevät, joten niitä on mahdollista käyttää ristikkäin ilman merkittävää annotointitarkkuuden putoamista. Audio-pohjaisten menetelmien arvoinnissa havaittiin, että SLP kykenee tuottamaan samantasoisia tai jopa tarkempia arvioita kuin ACT. Tämä osoitti yllättäen, että käytettäessä tehokkaita menetelmiä audio-tiedostosta lasketut piirteet voivat olla malleille hyödyllisempiä tunnetilojen vihjeitä kuin semanttinen tagitieto. Sekä tagi- että audio-pohjaisten mal-

lien mukauttaminen eri musiikkityyleihin paransi tunnistustarkkuutta entisestään, erityisesti arvioitaessa musiikin positiivisia ja negatiivisia tunnetiloja.

Väitöskirjassa osoitettiin ensi kertaa kattavasti laajan tagiaineiston hyödyt musiikin tunnetilojen mallinnuksessa. Lisäksi väitöskirjassa kehitetyt uudet menetelmät ovat käyttökelpoisia kehitettäessä entistä tehokkaampa musiikkisovelluksia.

REFERENCES

Ali, S. O., & Peynircioğlu, Z. F. (2006). Songs and emotions: are lyrics and melodies equal partners? *Psychology of Music*, *34*(4), 511–534.

Ames, M., & Naaman, M. (2007). Why we tag: Motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 971–980).

Aucouturier, J., Defreville, B., & Pachet, F. (2007). The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *Journal of the Acoustic Society of America*, *122*, 881-891.

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

Barthet, M., Fazekas, G., & Sandler, M. (2012). Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models. In *Proceedings of the 9th International Symposium on Computer Music Modeling and Retrieval (CMMR)* (pp. 492–507).

Beedie, C., Terry, P., & Lane, A. (2005). Distinctions between emotion and mood. *Cognition & Emotion*, *19*(6), 847–878.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., & Sandler, M. B. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, *13*(5), 1035–1047.

Bertin-Mahieux, T., Eck, D., Maillet., F., & Lamere, P. (2008). Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, *37*(2), 115-135.

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR).*

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 & Emotion*, *19*, 1113–1139.

Bischoff, K., Firan, C., Paiu, R., Nejdl, W., Laurier, C., & Sordo, M. (2009). Music mood and theme classification a hybrid approach. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR).*

Bischoff, K., Firan, C. S., Nejdl, W., & Paiu, R. (2008). Can all tags be used for search? In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (pp. 193–202).

Blumensath, T., & Davies, M. (2006). Sparse and shift-invariant representations of music. *IEEE Transactions on Audio, Speech, and Language Processing*, *1*(1), 50–57.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavioral Therapy and Ex-*

perimental Psychiatry, *25*, 49–59.

Bradley, M. M., & Lang, P. J. (1999). *Affective norms for english words (ANEW): Instruction manual and affective ratings* (Tech. Rep.). Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.

Bu, J., Tan, S., Chen, C., Wang, C., Wu, H., Zhang, L., & He, X. (2010). Music recommendation by unified hypergraph: combining social media information and music content. In *Proceedings of the 18th ACM International Conference on Multimedia* (pp. 391–400).

Calvo, R. A., & Mac Kim, S. (2012). Emotions in text: Dimensional and categorical models. *Computational Intelligence*.

Casey, M. A., Veltkamp, R., Goto, M., Leman, M., Rhodes, C., & Slaney, M. (2008). Content-based music information retrieval: Current directions and future challenges. *Proceedings of the IEEE*, *96*(4), 668–696.

Coviello, E., Chan, A. B., & Lanckriet, G. (2011). Time series models for semantic music annotation. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(5), 1343–1359.

Cunningham, P. (2008). Dimension reduction. In M. Cord & P. Cunningham (Eds.), *Machine learning techniques for multimedia* (pp. 91–112). Springer Berlin Heidelberg.

Deerwester, S., Dumais, S. T., Furnas, G. W., & Landauer, T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407.

Dieleman, S., & Schrauwen, B. (2013). Multiscale approaches to music audio feature learning. In *Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR)* (pp. 3–8).

Eck, D., Lamere, P., Bertin-Mahieux, T., & Green, S. (2007). Automatic generation of social tags for music recommendation. *Advances in Neural Information Processing Systems*, *20*(20), 1–8.

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., Lartillot, O., & Toiviainen, P. (2009). Prediction of multidimensional emotional ratings in music from audio using multivariate regression models. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)* (pp. 621–626).

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

Eerola, T., & Vuoskoski, J. K. (2013). A review of music and emotion studies: Approaches, emotion models and stimuli. *Music Perception*, *30*(3), 307–340.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*, 169-200.

Ekman, P. (1999). Handbook of cognition and emotion. In T. Dalgleish & M. Power (Eds.), (pp. 45–60). New York: John Wiley & Sons.

Fellbaum, C. (Ed.). (1998). *Wordnet: An electronic lexical database*. Cambridge, MA: MIT Press.

Feng, Y., Zhuang, Y., & Pan, Y. (2003). Popular music retrieval by detecting mood. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval.* Toronto.

Friberg, A., Schoonderwaldt, E., Hedblad, A., Fabiani, M., & Elowsson, A. (2014). Using listener-based perceptual features as intermediate representations in music information retrieval. *The Journal of the Acoustical Society of America*, *136*(4), 1951–1963.

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

Gabrielsson, A. (2002). Emotion perceived and emotion felt: Same or different. *Musicae Scientiae*, *2001-2002*, 123-147.

Gabrielsson, A. (2010). Strong experiences with music. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 547–574). Oxford University Press.

Gabrielsson, A., & Lindström, E. (2001). The influence of musical structure on emotional expression. In P. N. Juslin & J. A. Sloboda (Eds.), *Music and Emotion: Theory and Research* (pp. 223–248). Oxford University Press.

Garcia-Silva, A., Corcho, O., Alani, H., & Gomez-Perez, A. (2012). Review of the state of the art: Discovering and associating semantics to tags in folksonomies. *The Knowledge Engineering Review*, *27*(01), 57–85.

Golder, S. A., & Huberman, B. A. (2006, April). Usage patterns of collaborative tagging systems. *Journal of Information Science*, *32*(2), 198-208.

Gomez, P., & Danuser, B. (2004). Affective and physiological responses to environmental noises and music. *International Journal of Psychophysiology*, *53*(2), 91–103.

Goto, M. (2001). An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, *30*(2), 159–171.

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes problems* (Vol. 3). Oxford University Press.

Hamel, P., & Eck, D. (2010). Learning features from music audio with deep belief networks. In *Proceedings of the 11th International Conference of Music Information Retrieval (ISMIR)* (pp. 339–344).

Han, B., Rho, S., Jun, S., & Hwang, E. (2009). Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 1-28.

Han, J., & Kamber, M. (2001). *Data mining : concepts and techniques* (D. D. Serra, Ed.). San Francisco (CA) : Morgan Kaufmann.

Hawkins, D. M. (2004). The problem of overfitting. *Journal of Chemical Information and Computer Sciences*, *44*(1), 1–12.

Hevner, K. (1936). Experimental studies of the elements of expression in music. *The American Journal of Psychology*, *48*(2), 246-268.

Hevner, K. (1937). The affective value of pitch and tempo in music. *The American Journal of Psychology*, 621–630.

Heymann, P., Ramage, D., & Garcia-Molina, H. (2008). Social tag prediction. In

Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 531–538).

Hoffman, M. D., Blei, D. M., & Cook, P. R. (2009). Easy as CBA: A simple probabilistic model for tagging music. In *Proceedings of the 11th International Conference on Music Information Retrieval (ISMIR)* (pp. 369–374).

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*(1-2), 177–196.

Hopkins, B., & Skellam, J. G. (1954). A new method for determining the type of distribution of plant individuals. *Annals of Botany*, *18*(2), 231-227.

Hu, X., & Downie, J. S. (2007). Exploring mood metadata: relationships with genre, artist and usage metadata. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR).*

Hu, X., Downie, J. S., & Ehmann, A. F. (2009). Lyric text mining in music mood classification. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)* (pp. 411–416).

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 the 9th International Conference on Music Information Retrieval (ISMIR)* (pp. 462–467).

Jensen, D., & Cohen, P. R. (2000). Multiple comparisons in induction algorithms. *Machine Learning*, *38*, 309-338.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Proceedings of the International Conference on Machine Learning* (pp. 121–129).

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

Juslin, P. N. (2013b). What does music express? basic emotions and beyond. *Frontiers in Psychology*, *4*.

Juslin, P. N., Harmat, L., & Eerola, T. (2013). What makes music emotionally significant? exploring the underlying mechanisms. *Psychology of Music*, *42*(4), 599–623.

Juslin, P. N., & Sloboda, J. A. (2009). *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*(5), 559–575.

Kaminskas, M., & Ricci, F. (2012). Contextual music information retrieval and recommendation: State of the art and challenges. *Computer Science Review*, *6*(2), 89–119.

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

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., … Turnbull, D. (2010). Music emotion recognition: A state of the art review. In *Proceedings of the 11th International Conference of Music Information Retrieval*

(ISMIR) (pp. 255–266).

Knees, P., Pampalk, E., & Widmer, G. (2004). Artist classification with web-based data. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR).*

Knees, P., Pohle, T., Schedl, M., & Widmer, G. (2007). A music search engine built upon audio-based and web-based similarity measures. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 447–454).

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, *97*(1-2), 273-324.

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

Lam, X. N., Vu, T., Le, T. D., & Duong, A. D. (2008). Addressing cold-start problem in recommendation systems. In *Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication* (pp. 208–211).

Lamere, P. (2008). Social tagging and music information retrieval. *Journal of New Music Research*, *37*(2), 101-114.

Lartillot, O., & Toiviainen, P. (2007, September). A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects (DAFx).*

Laurier, C., Sordo, M., Serra, J., & Herrera, P. (2009). Music mood representations from social tags. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)* (pp. 381–86).

Law, E., Settles, B., & Mitchell, T. (2010). Learning to tag from open vocabulary labels. In J. Balcázar, F. Bonchi, A. Gionis, & M. Sebag (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 6322, pp. 211–226). Springer Berlin Heidelberg.

Law, E., & Von Ahn, L. (2009). Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1197–1206).

Lee, H., Pham, P., Largman, Y., & Ng, A. Y. (2009). Unsupervised feature learning for audio classification using convolutional deep belief networks. In *Advances in Neural Information Processing Systems* (pp. 1096–1104).

Levy, M. (2012). *Retrieval and annotation of music using latent semantic models* (Unpublished doctoral dissertation). School of Electronic Engineering and Computer Science, Queen Mary, University of London.

Levy, M., & Sandler, M. (2007). A semantic space for music derived from social tags. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR).*

Levy, M., & Sandler, M. (2008). Learning latent semantic models for music from social tags. *Journal of New Music Research*, *37*(2), 137-150.

Levy, M., & Sandler, M. (2009). Music information retrieval using social tags and audio. *IEEE Transactions on Multimedia*, *11*(3), 383–395.

Li, T., & Ogihara, M. (2003). Detecting emotion in music. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)* (pp. 239–240).

Lin, Y.-C., Yang, Y.-H., & Chen, H.-H. (2009). Exploiting genre for music emotion classification. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)* (pp. 618–621).

Lin, Y.-C., Yang, Y.-H., & Chen, H. H. (2011). Exploiting online music tags for music emotion classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *7*(1), 26.

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the 1st International Conference on Music Information Retrieval (ISMIR).*

Lu, L., Liu, D., & Zhang, H.-J. (2006, Jan.). Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Speech and Audio Processing*, *14*(1), 5-18.

MacDorman, S., & Ho, S. (2007). Automatic emotion prediction of song excerpts: Index construction, algorithm design, and empirical comparison. *Journal of New Music Research*, *36*(4), 281-299.

Mandel, M. I., & Ellis, D. P. (2005). Song-level features and support vector machines for music classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)* (pp. 594–599).

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

Marlin, B., Zemel, R., Roweis, S., & Slaney, M. (2007). Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence.*

Miotto, R., & Lanckriet, G. (2012). A generative context model for semantic music annotation and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(4), 1096–1108.

Miotto, R., & Orio, N. (2012, May). A probabilistic model to combine tags and acoustic similarity for music retrieval. *ACM Transactions on Information Systems*, *30*(2), 8:1–8:29.

Mirkin, B. (1996). *Mathematical classification and clustering*. Kluwer Academic Press, Dordrecht.

Nanopoulos, A., Rafailidis, D., Symeonidis, P., & Manolopoulos, Y. (2010). Musicbox: Personalized music recommendation based on cubic analysis of social tags. *IEEE Transactions on Audio, Speech, and Language Processing*, *18*(2), 407–412.

Ness, S. R., Theocharis, A., Tzanetakis, G., & Martins, L. G. (2009). Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs. In *Proceedings of the 17th ACM International Conference on Multimedia* (pp. 705–708).

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *2*(1-2), 1–135.

Pauws, S. (2004). Musical key extraction from audio. In *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR).*

Peng, J., Zeng, D. D., & Huang, Z. (2008, October). Latent subject-centered modeling of collaborative tagging: An application in social search. *ACM Transactions on Management Information Systems*, *2*(3), 15:1–15:23.

Peter, R., Shivapratap, G., Divya, G., & Soman, K. (2009). Evaluation of SVD and NMF methods for latent semantic analysis. *International Journal of Recent Trends in Engineering*, *1*(3), 308–310.

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Emotion: Theory, Research, and Experience*, *1*(3), 3–33.

Reunanen, J. (2007). Model selection and assessment using cross-indexing. In *Proceedings of the 20th International Joint Conference on Neural Networks (IJCNN'07)* (pp. 2581–2585).

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

Russell, J. A., & Barrett, L. F. (1999). Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant. *Journal of Personality and Social Psychology*, *76*(5), 805–819.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613–620.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web* (pp. 285–295).

Scaringella, N., Zoia, G., & Mlynek, D. (2006). Automatic genre classification of music content. *IEEE Signal Processing Magazine : Special Issue on Semantic Retrieval of Multimedia*, *23*, 133-141.

Schedl, M., Hauger, D., & Urbano, J. (2013). Harvesting microblogs for contextual music similarity estimation: a co-occurrence-based framework. *Multimedia Systems*, 1–13.

Schedl, M., & Knees, P. (2009). Context-based music similarity estimation. In *Proceedings of the 3rd International Workshop on Learning Semantics of Audio Signals* (pp. 59–74).

Scherer, K. R. (1984). Emotion as a multicomponent process: A model and some cross-cultural data. In *Review of Personality and Social Psychology* (Vol. 5, pp. 37–63). Beverly Hills: CA: Sage.

Scherer, K. R. (2000). Emotion. In M. Hewstone & W. Stroebe (Eds.), *Introduction to social psychology: A European perspective* (3rd ed., pp. 151–191). Oxford: Blackwell.

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

Scherer, K. R. (2005). What are emotions? and how can they be measured? *Social Science Information*, *44*(4), 695–729.

Scherer, K. R., & Zentner, M. R. (2001). Music and emotion: Theory and research. In P. N. Juslin & J. A. Sloboda (Eds.), (pp. 361–392). Oxford University Press.

Schimmack, U., & Grob, A. (2000). Dimensional models of core affect: A quan-

titative comparison by means of structural equation modeling. *European Journal of Personality*, *14*(4), 325–345.

Schmidt, E. M., & Kim, Y. E. (2011). Learning emotion-based acoustic features with deep belief networks. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)* (pp. 65–68).

Schmidt, E. M., Scott, J. J., & Kim, Y. (2012). Feature learning in dynamic environments: Modeling the acoustic structure of musical emotion. In *Proceedings of the 13th International Conference on Music Information Retrieval (ISMIR)* (pp. 325–330).

Schmidt, E. M., Turnbull, D., & Kim, Y. E. (2010). Feature selection for content-based, time-varying musical emotion regression. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval* (pp. 267–274).

Schubert, E. (2003). Update of the hevner adjective checklist. *Perceptual and Motor Skills*, *96*, 117–1122.

Schuller, B., Hage, H., Schuller, D., & Rigoll, G. (2010). 'Mister D.J., cheer me up!': Musical and textual features for automatic mood classification. *Journal of New Music Research*, *39*(1), 13–34.

Schuller, B., Weninger, F., & Dorfner, J. (2011). Multi-modal non-prototypical music mood analysis in continuous space: Reliability and performances. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)* (pp. 759–764).

Seung, D., & Lee, L. (2001). Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems*, *13*, 556–562.

Sinclair, J., & Cardew-Hall, M. (2008). The folksonomy tag cloud: when is it useful? *Journal of Information Science*, *34*(1), 15–29.

Song, Y., Dixon, S., Pearce, M., & Halpern, A. R. (2013). Do online social tags predict perceived or induced emotional responses to music? In *Proceedings of the 14th International Conference on Music Information Retrieval (ISMIR)* (pp. 89–94).

Song, Y., Zhang, L., & Giles, C. L. (2011). Automatic tag recommendation algorithms for social recommender systems. *ACM Transactions on the Web (TWEB)*, *5*(1), 4.

Sordo, M., Laurier, C., & Celma, O. (2007). Annotating music collections: How content-based similarity helps to propagate labels. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR).*

Symeonidis, P., Ruxanda, M. M., Nanopoulos, A., & Manolopoulos, Y. (2008). Ternary semantic analysis of social tags for personalized music recommendation. In *Proceedings of the 9th International Conference of Music Information Retrieval (ISMIR)* (pp. 219–224).

Thayer, R. E. (1989). *The biopsychology of mood and arousal.* Oxford University Press, New York, USA.

Tingle, D., Kim, Y. E., & Turnbull, D. (2010). Exploring automatic music annotation with acoustically-objective tags. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval* (pp. 55–62).

Tolonen, T., & Karjalainen, M. (2000). A computationally efficient multipitch

analysis model. *IEEE Transactions on Speech and Audio Processing*, *8*(6), 708-716.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. (2008). Multilabel classification of music into emotions. In *Proceedings of the 9th International Conference on Music Information Retrieval (ISMIR).*

Turnbull, D., Barrington, L., Torres, D., & Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(2), 467-476.

Turnbull, D. R., Barrington, L., Lanckriet, G., & Yazdani, M. (2009). Combining audio content and social context for semantic music discovery. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 387–394).

Tzanetakis, G., & Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*(5), 293-302.

Vieillard, S., Peretz, I., Gosselin, N., Khalfa, S., Gagnon, L., & Bouchard, B. (2008). Happy, sad, scary and peaceful musical excerpts for research on emotions. *Cognition & Emotion*, *4*(720–752).

Vuoskoski, J. K., & Eerola, T. (2011). The role of mood and personality in the perception of emotions represented by music. *Cortex*, *47*(9), 1099-1106.

Wang, J. C., Yang, Y. H., Chang, K., Wang, H. M., & Jeng, S. K. (2012). Exploring the relationship between categorical and dimensional emotion semantics of music. In *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies* (pp. 63–68).

Warriner, V., Amy Bethand Kuperman, & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 1–17.

Wetzker, R., Umbrath, W., & Said, A. (2009). A hybrid approach to item recommendation in folksonomies. In *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval* (pp. 25–29).

Wieczorkowska, A., Synak, P., Lewis, R., & Ras, Z. W. (2005). Extracting emotions from music data. In M.-S. Hacid, N. Murray, Z. Ras, & S. Tsumoto (Eds.), *Proceedings of the ISMIS'05: Foundations of Intelligent Systems* (pp. 456–465). New York: Springer.

Wieczorkowska, A., Synak, P., & Zbigniew, R. (2006). Multi-label classification of emotions in music. In *Proceedings of the Intelligent Information Processing and Web Mining* (pp. 307–315).

Wold, E., Blum, T., Keislar, D., & Wheaton, J. (1996). Content-based classification, search, and retrieval of audio. *IEEE Multimedia*, *3*(2), 27-36.

Wundt, W. (1897). *Outlines of psychology*. Lepzig: Englemann. (trans. C. H. Judd)

Yang, Y.-H., & Chen, H. H. (2012). Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology*, *3*(3).

Yang, Y.-H., Lin, Y.-C., Lee, A., & Chen, H. H. (2009). Improving musical concept detection by ordinal regression and context fusion. In *Proceedings of the 10th International Conference of Music Information Retrieval (ISMIR)* (pp. 147–152).

Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008, Feb.). A regression ap-

proach 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 ACM International Conference on Multimedia* (pp. 81–84).

Yang, Y.-H., Su, Y.-F., Lin, Y.-C., & Chen, H. H. (2007). Music emotion recognition: the role of individuality. In *HCM '07: Proc. International Workshop on Human-centered Multimedia* (pp. 13–22). ACM.

Yaslan, Y., & Cataltepe, Z. (2006). Audio music genre classification using different classifiers and feature selection methods. In *Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06)* (Vol. 2).

Yik, M. S., Russell, J. A., & Barrett, L. F. (1999). Structure of self-reported current affect: Integration and beyond. *Journal of Personality and Social Psychology*, *77*(3), 600–619.

Zentner, M. R., & Eerola, T. (2010). Self-report measures and models. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of Music and Emotion: Theory, Research, Applications* (pp. 187–221). Boston, MA: Oxford University Press.

Zentner, M. R., Grandjean, D., & Scherer, K. (2008). Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, *8*(4), 494–521.

Zhou, N., Cheung, W., Qiu, G., & Xue, X. (2011). A hybrid probabilistic model for unified collaborative and content-based image tagging. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(7), 1281–1294.

ORIGINAL PAPERS

I

GENERALIZABILITY AND SIMPLICITY AS CRITERIA IN FEATURE SELECTION: APPLICATION TO MOOD CLASSIFICATION IN MUSIC

by

Pasi Saari, Tuomas Eerola & Olivier Lartillot 2011

IEEE Transactions on Audio, Speech, and Language Processing, 19 (6), 1802-1812

Generalizability and Simplicity as Criteria in Feature Selection: Application to Mood Classification in Music

Pasi Saari, Tuomas Eerola, and Olivier Lartillot, *Member, IEEE*

***Abstract*—Classification of musical audio signals according to expressed mood or emotion has evident applications to content-based music retrieval in large databases. Wrapper selection is a dimension reduction method that has been proposed for improving classification performance. However, the technique is prone to lead to overfitting of the training data, which decreases the generalizability of the obtained results. We claim that previous attempts to apply wrapper selection in the field of Music Information Retrieval (MIR) have led to disputable conclusions about the used methods due to inadequate analysis frameworks, indicative of overfitting and biased results. This paper presents a framework based on cross-indexing for obtaining realistic performance estimate of wrapper selection by taking into account the simplicity and generalizability of the classification models. The framework is applied on sets of film soundtrack excerpts that are consensually associated with particular basic emotions, comparing Naive Bayes, k-NN and SVM classifiers using both forward selection (FS) and backward elimination (BE). K-NN with BE yields the most promising results – 56.5% accuracy with only four features. The most useful feature subset for k-NN contains mode majorness and key clarity, combined with dynamical, rhythmical, and structural features.**

***Index Terms*—Music and emotion, musical features, feature selection, wrapper selection, overfitting, cross-indexing.**

I. Introduction

AUTOMATIC recognition of emotions in musical audio has gained increasing attention in the field of Music Information Retrieval (MIR) during the past few years. The development in the field has coincided with the need for managing large collections of digital audio for the public via web services such as Spotify[1] and Last.fm[2]. This is reflected, for example, in the increasing number of submitted systems in the annual Audio Music Mood Classification (AMC) contest part of the Music Information Retrieval Evaluation eXchange[3] (MIREX). The substantial variance in the submitted systems in the contest and in the approaches in the MIR field in general indicates a lack of consensus concerning the choice of an underlying psychological model for emotions or moods in music and the establishment of precise machine learning conventions.

Despite research in musicology studying the influence of specific cues in the musical structure on emotional expressions [1], there is no complete analytical consensus about the required 'ingredients' – i.e. acoustical features extracted from music – to build the optimal models for emotion recognition, in the limits imposed by the subjectivity of emotions. Machine learning comprises methods developed for detecting these types of relations automatically by taking into account the interrelations between features, but the potential of these methods is limited by the ability of the chosen set of features to describe music in the same way as how listeners perceive it [2] or by the complexity or technical deficiencies in the feature sets [3].

A part of the research field has adopted the view that regression models are more useful for understanding emotions in music than classifiers [4]. The most adequate linear models transform the feature space used in learning into few dimensions constituting of sets of input features while retaining the predictive information about the target concept. On the other hand, classifiers traditionally exploit the features independently when building the model. The prevalent downside to this approach in MIR has been the large dimensionality of the feature space, which leads to models that are quite difficult to interpret, contributing therefore rather modestly to the understanding of the phenomenon under study.

Another 'curse' relating to the high dimensionality of the input data given to the learning method is that it leads to overfitting, which is reflected in the low degree of generalizability of the models in classifying unknown data. Different dimension reduction methods applied to the input data have been developed to deal with problems related to high dimensionality in machine learning. Wrapper selection [5] is a method that can be used to find a subset of input features optimal for a given classifier. Perhaps surprisingly, wrapper selection too is highly prone to overfitting when the found subsets are used to build classification models [6]. The analysis of the previous research in MIR (detailed in the section V) shows that the use of wrapper approach has almost constantly led to disputable results indicating overfitting of the data. This pitfall was addressed in music

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the authors. This includes the analyzed feature sets and MATLAB scripts that enable experimenting with the used methods and reproducing some of the reported results and visualizations. This material is 156 KB in size.

P. Saari, T. Eerola and O. Lartillot are with the Finnish Centre of Excellence in Interdisciplinary Music Research, Music Department of the University of Jyväskylä, Finland.

[1]http://www.spotify.com/

[2]http://www.last.fm/

[3]http://www.music-ir.org/mirex/20xx/index.php/Main_Page

classification in [7] by applying guidelines given in machine learning studies [6] and [8] for building a wrapper selection and classification framework. Since then, the guidelines have been developed further in a proposed cross-indexing algorithm that has been shown to yield unbiased estimates of the performance of classification models [9], [10]. Applying the algorithm leads to a framework that is essentially a realization of a model parameter selection problem where the optimal dimensionality of the feature space for a given classifier is searched. Cross-indexing has not been used in music classification previously.

This study aims at developing a framework for studying wrapper selection in music classification based on the cross-indexing algorithm. It will be shown that classification in MIR, specifically in the recognition of expressed emotions in music, can lead to interpretable and efficient models when the number of input features is reduced dramatically, taking into account the simplicity and generalizability of the models. This required a novel modification proposed to the cross-indexing algorithm. The advantage of the chosen approach is that rather than combining a large number of features to represent few dimensions open to interpretations as in linear modeling, the gained dimensions will be exactly those single features itself whose relations to the phenomenon under study are understood, at least abstractly.

II. Feature Selection

Given the data, comprised of instances described by features and the targets, the problem of *feature selection* can be formulated as the task of finding a subset of the original features that maximizes the performance of a given learning algorithm run on the data. By reducing the amount of data used in learning, feature selection can reduce the problems related to feature redundancy, feature irrelevance and the curse of dimensionality [3]. Consequently, feature selection is crucial in reducing the required computational effort in learning and classification, reducing the complexity of the obtained models and increasing the generalization capabilities of the models.

Feature ranking, perhaps the most straightforward feature selection technique, evaluates each feature independently from the other features according to a pre-defined measure. N top-ranked features can then be used in the final subset. Because of the independent evaluation, feature relevance and redundancy cannot be taken into account in feature ranking, thus potentially harming the predictive capabilities of the results. To avoid such drawback, feature selection commonly implements subset selection, which evaluates features in the context of the whole feature subset. In such approach, called *feature subset selection*, an initial feature set – which can be the empty set – is iteratively transformed based on pre-specified evaluation criterion and search method until a given termination condition has been met.

Two main approaches can be distinguished based on the type of evaluation criterion used in the search: *filter* and *wrapper*. In the filter approach, a chosen information-theoretic measure is used as evaluation criterion. In the wrapper approach [5], [11], each feature subset is evaluated using the learning algorithm that will ultimately perform the final data classification. The approach is based on a claim [11] that subset selection must take into account the biases of the induction algorithm in order to select a subset with highest possible prediction accuracy of that algorithm on unseen data. In an extensive evaluation on various real-world and artificial data, wrapper performed favorably in comparison to a Relieved-F filter algorithm and induction with no selection [5].

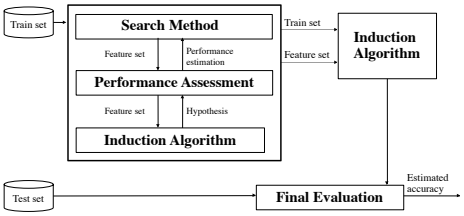


Fig. 1. Wrapper selection (chart adopted and slightly modified from [5]).

Fig. 1 illustrates the wrapper selection process. First the data is split into train and test sets. The train set is used for feature selection while keeping the test set only for the final evaluation of the performance of the induction algorithm. The search is conducted by iteratively evaluating each candidate subset with respect to the performance of the induction algorithm. The performance is assessed usually either through cross-validation or using a validation set that is separate from the train and test sets. After the terminating condition is met the learning phase is conducted on the train set represented by the selected feature subset. Last, the output model is evaluated based on the test set.

Finding the optimal subset of features requires searching the whole feature space, involving as many evaluations as there are possible feature subsets. When the number of features increases, an *exhaustive search* encompassing all possible subsets becomes prohibitively expensive. Therefore, *heuristic* search methods have been developed in order to reduce the number of evaluated subsets. Many popular heuristic algorithms used in feature selection exploit *greedy selection heuristics*, where choices are based on the local context defined at each successive step of the search [12, p. 370]. By making locally optimal choices, greedy heuristic does not guarantee that the search will lead to the globally optimal subset. Two of the most famously used greedy selection algorithms are *forward selection* (FS) and *backward elimination* (BE). They are called *stepwise algorithms* since they involve addition or removal of only one feature at each modification of the feature subset.

III. Minimizing Overfitting by Cross-Indexing

Overfitting relates to a problem that occurs when a learning algorithm fits the training data too well, considering peculiarities in the data such as noise or possible outliers as important characteristics of the phenomenon under analysis. While the problem is well-known in classification and regression models, the effect grows in feature selection, especially in the wrapper approach. Overfitting in wrapper selection is caused by the large number of evaluated feature subsets, which makes it likely that one subset leads to high predictive accuracy on the hold-out data used in selection [5]. Therefore evaluation and reporting of wrapper selection must be based on the performance on a test set unseen to the search process – or on test sets generated by an *outer loop of cross-validation* – not on the (cross-validation) estimates used in the selection process [5]. In fact, the cross-validation estimates used in the selection process were found to be seriously misleading estimates of the performance of wrappers in subsequent studies [6], [8].

The problem of overfitting in wrapper selection was further analyzed in [9] and [10] for the purpose of estimating the optimal subset size for a given learning algorithm and assessing its performance. The results obtained in [9] showed that the outer loop of cross-validation was able to decrease bias in performance assessment only to a certain degree, but the proposed *cross-indexing algorithms A* and *B* decreased the bias virtually to zero. The algorithms were proposed to circumvent bias emerging from using the same estimates in picking a certain model from a large set of candidates as those used in assessing the performance obtainable with the model.

In [10] the two algorithms were merged into a *generalized* $(N, K-N)$*-fold cross-indexing algorithm*. The difference between the outer loop of cross-validation and cross-indexing is that in cross-indexing, rather than by the averaged performances in the K iterations, the optimal subset size is estimated separately at each iteration $k \in [1, 2, \ldots, K]$ by the maximum averaged performance of N $(1 < N < K)$ iterations. Then, the performance at the kth iteration, with the subset size, is obtained by the averaged performance of $K - N$ other iterations. The final estimates for the optimal subset size and its performance is then obtained by averaging the K estimates. The parameter N can be considered as a trade-off between leaning towards accurate estimation of the optimal subset and towards accurate estimation of the performance attainable with the obtained subset size. Based on the preliminary results with the algorithm [10], the choice of $1 < N < K-1$ is a good trade-off between these two.

IV. Emotion and Mood Recognition in MIR

To enable direct comparison of the systems dedicated to audio mood recognition, classification in the annual Audio Music Mood Classification (AMC) task organized by MIREX is conducted on a collectively agreed large ground-truth set, measuring the performance with 3-fold cross-validation [13]. During the three years that the contest has been held, the average classification accuracies of all submitted systems have increased from 53% to 58% and the performance of the winning systems have increased from 62% to 66%[4]. The 66% accuracy in music classification can be considered rather high considering the estimated glass-ceiling performance of around 70% in music classification based on timbre similarity [14].

Some previous studies on classification of music according to emotions such as [15], [16] and [17] have reported drastically higher performances than AMC suggests. However, comparing the results obtained in the AMC evaluations with previous research into mood recognition is problematic due to fundamental differences in ground-truth, performance evaluation and reporting.

V. Wrapper Selection in MIR

The bias of the cross-validation estimates used in wrapper selection process has not been given enough consideration within MIR community. In fact, reviewing the few studies in MIR using wrapper selection reveals likely inadequate frameworks for obtaining performance estimates for the used methods. For example, wrapper selection with genetic search algorithm was used in [18], comparing the performance of the Decision Tree, 3-NN, Naive Bayes, Multi-layer Perceptron and Support Vector Machine in genre classification. The results were reported to indicate that wrapper selection procedure is effective for Decision Tree, 3-NN and Naive Bayes but the authors did not base their arguments on classification performance on independent test sets. This may have led to positively biased results given that overfitting in wrapper selection with genetic algorithms is an acute problem [19]. Similar conclusions can be drawn from [20] that exploited wrapper selection with fuzzy classifiers and BE to emotion recognition. The results showed increasing performance with both classifiers with the estimated optimal subset size. However, the found effects might have been optimistic since the final performance assessment and reporting of the results was based only on cross-validation accuracies used in the selection process.

In [21] wrapper selection was applied to genre classification using FS and BE with ten different classifiers. The dataset was first split randomly into train and test sets constituting of 90% and 10% of the excerpts, respectively. Then, wrapper selection was run with the train set and finally the test set accuracies with different subset sizes were reported. However, since the whole process was run only once, the results might have been somewhat random. Using the outer loop of cross-validation would have improved the validity to some extent.

In the single exception [7] within MIR community that specifically addressed overfitting in feature selection, the authors re-evaluated the claims made in their previous study [22] where feature weighting with genetic algorithm

[4]The results of the corresponding years are shown in http://www.music-ir.org/mirex/20xx/index.php/Audio_Music_Mood_Classification_Results.

TABLE I
SUMMARY OF THE DATASETS.

Set	Features	Classes	Excerpts
Primary set	66	*anger/fear, happy, sad, tender*	64
Validation set	66	*anger/fear, happy, sad, tender*	160

increased the performance of k-NN in timbre recognition from snare-drum attack and beat-box sounds. Performance assessment and evaluation of the results in [22] was based solely on cross-validation accuracies obtained in the selection process while the re-evaluation was based on the guidelines given in [6] and [8]. Re-evaluation incorporated the outer loop of cross-validation and concentrated on the wrapper with k-NN and FS with the same datasets as in [22]. The results, obtained in the same manner as in the previous study, indicated significant degree of overfitting as the performance of k-NN decreased in the independent test sets. Moreover, the performance improvement with the selected feature sets, when compared to full sets, provided little or no benefit. To gain further evidence on the usability of wrapper selection, the framework was evaluated in a genre classification task. Also Principal Component Analysis (PCA) was used for dimensionality reduction for comparison purposes. In genre classification, wrapper selection provided significant improvement in terms of test set accuracies when compared to classification without feature selection but PCA yielded similar increase in accuracy in a fraction of computation time. The authors concluded by stressing the importance of a well-founded framework for feature selection in MIR areas.

VI. Proposed Cross-Indexing Framework

The aim of the study is reached by testing the behaviors and benefits of feature selection, namely wrapper selection, on a task of classification of musical audio according to the expressed emotion. The analysis was conducted in a comparative manner to obtain information about different selection methods. For the sake of convenience, the analysis was split into two experiments. In Experiment 1 all selection methods were used. The most promising ones in Experiment 1 were chosen for Experiment 2 whose aim was to give more reliable and detailed information about the performance of these methods. The framework, illustrated in Fig. 2, is explained in the forthcoming sections.

A. Data Pre-Processing

This section describes how the primary and validation sets were obtained and how they were preprocessed for the purposes of the analysis. The sets are summarized in Table I.

1) Audio Material and Its Annotation: The audio material[5] constitutes of two sets of film soundtrack excerpts containing 360 and 110 excerpts with length from 10 to 30 seconds detailed in [23]. Both of the sets were annotated by the listeners in terms of basic emotions *anger, fear, happy, sad* and *tender*, each emotion on a scale from 1 to 7. First, the set of 360 excerpts was rated by 12 musicology experts. Then, 110 of the excerpts with unambiguous emotion content were rated by 116 non-musician participants.

[5]The material is downloadable at https://www.jyu.fi/hum/laitokset/musiikki/en/research/coe/materials/emotion/soundtracks.

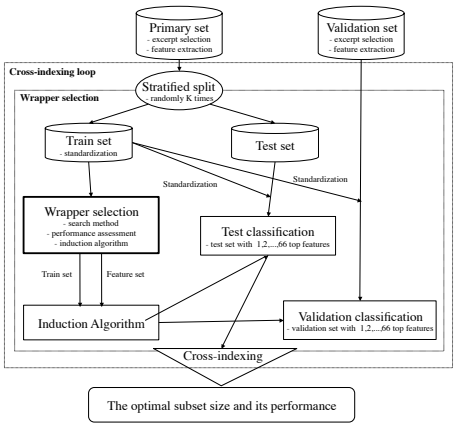


Fig. 2. A chart describing the cross-indexing framework.

High correlation between ratings of *anger* and *fear* indicates that they might not be easily distinguished in the analyzed data [23]. Therefore these emotions were grouped into one category in the present study. The annotated material was adapted for classification purposes by assigning the highest rated emotion of each musical excerpt as the single representative emotion label *anger/fear*, *happy*, *sad*, or *tender*.

2) Excerpt Selection: The excerpts from Experiments 1 and 2 in [23] were evaluated according to their applicability for classification purposes. The aim was to omit excerpts whose expressed emotions had been ambiguous according to the participants' ratings. A selection measure for each excerpt, similar to that in [23], was computed by taking a ratio between the mean rating of the highest rated emotion category and the mean rating of the emotion that was rated second highest. Thus the same amount of excerpts was preserved for each emotion, corresponding to those scoring highest in terms of the selection measure. In this way, the primary set with 64 excerpts was selected from the set of 110 samples by preserving 16 excerpts in each of the four emotion categories. The validation set was formed by first removing the excerpts that overlapped with the primary set. Then, 40 excerpts in each emotion category were selected in the same manner as the examples in the primary set. The fact that the primary and validation sets were rated by different groups of participants can induce slight inconsistency between the sets, but gives also an opportunity to assess model performances with data that are truly separate from the data used in the learning phase.

TABLE II
EXTRACTED FEATURE SET. m = MEAN, d = STANDARD DEVIATION, l = SLOPE, h = ENTROPY.

Category	No.	Feature	Acronyms
Dynamics	1-3	RMS energy	Em, Ed, El
	4	Low-energy ratio	LEm
	5	Attack time	ATm
	6-7	Attack slope	ASm, ASd
Rhythm	8	Event density	EDm
	9-10	Fluctuation peak (pos., mag.)	FPm, FMm
	11	Fluctuation centroid	FCm
	12-13	Tempo	Tm, Td
	14-15	Pulse clarity	PCm, PCd
Pitch	16-17	Pitch	Pm, Pd
	18-21	Chromagram (unwrapped) centr.	Cm, Cd, Cl, Ch
Harmony	22-23	Key clarity	KCm, KCd
	24-25	Key mode (majorness)	Mm, Md
	26	HCDF	Hm
	27	Entropy (oct. collapsed spectr.)	ESm
	28	Roughness	Rm
	29-30	Inharmonicity	Im, Id
Timbre	31-32	Brightness (cut-off 110 Hz)	Bm, Bd
	33-34	Spectral centroid	SCm, SCd
	35-36	Zerocross	Zm, Zd
	37	Spread	Sm
	38	Skewness	Km
	39-40	Spectral entropy	SEm, SEd
	41	Spectral flux	SFm
	42	Flatness	Fm
	43-44	Regularity	REm, REd
	45-46	1st MFCC + delta	M1m, D1m
	⋮ ⋮		⋮ ⋮
	57-58	7th MFCC + delta	M7m, D7m
Structure	59-60	Repetition (spectrum)	RSm, RSd
	61-62	Repetition (rhythm)	RRm, RRd
	63-64	Repetition (tonality)	RTm, RTd
	65-66	Repetition (register)	RGm, RGd

3) Feature Extraction: A total of 66 audio features, presented in table II were extracted with *MIRtoolbox*[6] [24]. The set contained 52 features suggested in [25] and 14 MFCC features.

The extraction was done with the frame-based approach [26], with 46 ms, 50% overlapping frames for most of the features. For low-energy ratio and high-level features that require longer frame length (fluctuation, harmony-related features), as well as tempo and pulse clarity, the analysis frame was 2 seconds with 50% overlap whereas the structure-related features were computed with frame length of 100 ms and 50% overlap. The values of frame length and overlap were based on the aforementioned analysis. In the case of MFCC features, the window was the same as with most other low-level features, i.e. 46 ms with 50% overlap.

B. Wrapper Selection

At each wrapper selection run the primary set was split in a random stratified manner into train and test sets. Stratified splitting was used to create random equal-sized subsets of samples for analysis while maintaining the relative numbers of excerpts expressing each emotion. In each of the K folds, a different random number seed was used for shuffling the dataset. The train set was standardized at each run to give each feature initially the same importance in the feature selection process and test and validation sets were z-score-transformed based on means and standard deviations of the train sets. Wrapper selection was done in Weka[7] software [27].

1) Performance Assessment: The performance of each candidate subset was assessed by cross-validation[8] with 4 folds. This relatively low number of folds in cross-validation was chosen to limit the computation time as the number corresponds to the number of times the learning algorithm must be used to estimate the performance of a single candidate subset.

a) Classifiers: The choice of the classifiers used in the wrapper selection is based on their popularity and efficiency in previous studies in musical emotion or mood classification. Three classifiers were chosen: Naive Bayes with Flexible Bayes modification [28], k-Nearest Neighbor (k-NN) with $k = 10$, and Support Vector Machine (SVM) with Pairwise coupling [29], Sequential Minimal Optimization [30] and fitting into linear regression models. Although k-NN and Naive Bayes are simple classifiers, they are competitive and sometimes can outperform more sophisticated classifiers, as pointed out in [31, p. 60]. Moreover, it was hypothesized that the performance of k-NN and Naive Bayes in particular could be significantly improved by wrapper selection since they are highly influenced by redundancy and irrelevance of features. SVM, on the other hand, is an example of a more sophisticated classifier with successful implementations in MIR.

b) Search Methods: Two search methods, FS and BE, were used to avoid reliance on one particular technique. These methods were chosen because of their popularity and simplicity, found as an advantage for example in [6].

2) Classification: Once the features were selected, the classifier taught on the whole train set, represented by the subsets of 1 to 66 top-ranked features according to the wrapper selection, was applied to the test and validation sets. In addition to the classification of the test data, the cross-validation accuracies, used in the selection process, were also recorded.

C. Cross-Indexing Loop

The wrapper selection methods were evaluated and compared primarily in terms of estimates obtained by the *generalized* $(N, K-N)$*-fold cross-indexing algorithm* [10] outlined in Fig. 3. Step 1 refers to the cross-indexing loop visualized in Fig. 2. Stratified splitting, feature selection and classification are conducted as described in the previous sections. Only the classification results on the test set are taken into account in the algorithm.

The loop in Step 2 produces the K optimal subset size and performance estimates. Each optimal subset size estimate is based on the values computed in Step 3 by

[6]Version 1.1.17. *MIRtoolbox* is available from www.jyu.fi/music/coe/material/mirtoolbox.

[7]Weka is available from http://www.cs.waikato.ac.nz/ml/weka.

[8]This refers to the inner cross-validation used in the selection process as opposed to the outer loop of cross-validation.

averaging N classification accuracies on the test set and the estimate of the performance attainable with the obtained optimal subset size in the kth fold is computed in Step 12 by averaging the other $K - N$ classification accuracies not used in Step 3. The final estimate of the optimal subset size and performance are obtained in Step 15 and Step 16. Based on the variances of the K size and performance estimates obtained with the algorithm, the value of $N = \frac{K}{2}$ was chosen as it produced low amount of variance in both estimates.

When analyzing the preliminary results, the initial algorithm was found to produce relatively large optimal subset size estimates while almost as high prediction accuracy estimates would have been obtained with notably smaller subsets. This means that the algorithm found the estimate of the optimal subset size although there existed a subset size that would potentially produce less complex models with nearly equally high prediction potential. Therefore a modification, outlined in Steps 4–11, was added in the algorithm. The rationale behind the modification was avoiding the curse of dimensionality by determining a cost for increasing the feature subset size at each cross-indexing fold.

The modification proceeds as follows: In each of the K folds, all D values obtained in Step 3 are considered as in the original algorithm. That is, the subset size estimate is determined by examining all the local maximum values rather than the global maximum of the D values. First, the local maximum with the smallest subset size is considered as the optimal. Then, comparison with the other local maxima is conducted by a selection criterion, which states that, when increasing the size by one feature, an increase of at least s percents in the maximum values is required for a bigger subset size to win the comparison. In such case, the new subset size is selected, replacing the smaller subset size estimate previously selected. These comparisons are repeated iteratively until all maxima have been evaluated.

Parameter s of the modification controls the emphasis given to smaller subset sizes. The value $s = 0$ yields the original cross-indexing algorithm since the cost of the subset size is omitted. Therefore the proposed algorithm can be thought of as a generalized version of the original cross-indexing algorithm. Increasing the value of s presumably yields smaller subset sizes. Although the algorithm aims at retaining the level of accuracy estimates, an excessively high value of s strongly reduces the accuracies. Therefore it is rather crucial to choose the right parameter value. Preliminary results indicated that the value $s = 1$ both reduced the estimated optimal subset size significantly while maintained the prediction accuracy at similar level. Increasing the value of s above 1 generally reduced the accuracy estimates by presumably over-emphasizing the small subset sizes. The modification also reduced the variance of the optimal subset size estimates across folds.

TABLE III
CROSS-INDEXING ESTIMATES OBTAINED IN EXPERIMENT 1 WITH THE MODIFIED ALGORITHM ($s = 1$). THE LAST COLUMN SHOWS THE RELATIVE EFFECT IN THE SUBSET SIZE AND ACCURACY ESTIMATES WHEN THE MODIFICATION WAS USED, COMPARED TO THE INITIAL ALGORITHM (S=0).

Method	Subset Size	Accuracy (%)	Size / Acc. (%)
NB FS	16.3 $\pm$ 2.9	59.4 $\pm$ 3.4	-22.6 / 1.9
NB BE	12.3 $\pm$ 7.2	52.3 $\pm$ 3.7	-53.8 / 0.8
k-NN FS	11.0 $\pm$ 3.9	52.7 $\pm$ 5.0	-25.4 / -3.6
k-NN BE	3.5 $\pm$ 0.6	57.4 $\pm$ 1.5	0 / 0
SVM FS	8.8 $\pm$ 5.6	55.5 $\pm$ 3.7	-56.8 / 0.7
SVM BE	3.0 $\pm$ 1.4	57.4 $\pm$ 5.5	-73.3 / -4.5

VII. RESULTS AND DISCUSSION

A. Experiment 1

In Experiment 1 the cross-indexing loop was run with four iterations. Table III summarizes the cross-indexing estimates and shows the effect of the modification ($s = 1$) compared to the initial algorithm ($s = 0$). It can be seen that the modified cross-indexing algorithm generally retained the level of the accuracy estimates, with maximum 4.5% decrease, while reducing the needed amount of features.

Fig. 4 displays the averaged accuracy estimates as well as the cross-indexing estimates for the optimal subset size and its performance. The cross-validation (cv) and test set accuracies correspond to the measures used to guide the selection and to the ones produced by the traditional external cross-validation loop, respectively. It is evident that the traditional methods – the cross-validation measures used to guide the selection as well as the outer loop of cross-validation – failed at predicting the performance of the feature selection methods when the obtained models were used for validation, as discussed in Section III. It would therefore be misleading to use the averaged values for estimating the optimal subset sizes for a given learning method and to use these values for estimating the performance of the method with that size.

When comparing the cross-indexing estimates (denoted by circles in the figure and summarized in Table III) to the validation set performances, it is notable that these estimates generally correspond fairly well to the optimal subset sizes in terms of validation. Naive Bayes with FS as well as k-NN and SVM with BE yielded the highest performances. However, the accuracies varied notably between the iterations (with standard deviations from 1.5% to 5.5%) causing uncertainty into the comparison of the mean accuracies, which differ by 6.7% at most. Therefore a high importance must be given to the subset sizes, which should be as small as possible in order to keep the models simple and interpretable. In this comparison SVM and k-NN with BE consistently yielded the smallest subset sizes with under 5 features whereas the subset sizes found with Naive Bayes and FS were the largest. Therefore SVM and k-NN with BE were chosen for Experiment 2.

1: Conduct feature selection and classification K times, each time with a given train-test split specific to the kth run ($k = 1, \ldots, K$). Test set accuracies with every subset size, denoted by $\left(x_1^{(k)}, x_2^{(k)}, \ldots, x_D^{(k)}\right)$ are obtained.
2: **for** $k = 1$, **to** K **do**
3: For each subset size ($i = 1, \ldots, D$), compute the mean of N estimates: $\dot{x}_i^{(k)} = \frac{1}{N}\left(\sum_{j=\alpha}^{K} x_i^{(j)} + \sum_{j=\beta}^{k} x_i^{(j)}\right)$, where $\alpha = K + k - N + 1$ and $\beta = \max(1, k - N + 1)$.
4: Use them to estimate the optimal model complexity:
5: Find local maxima $p_l^{(k)}$ from $\dot{x}^{(k)}$. The maxima are ordered by the subset size, starting from the maximum with the smallest size (starting points of local maximum plateaus are considered as local maxima). Initialize the complexity $d^{(k)} = p_1^{(k)}$.
6: **for** each remaining local maximum $l = 2$, **to** $numberofmaxima$ **do**
7: **if** $\dot{x}^{(k)}_{p_l^{(k)}} > \dot{x}^{(k)}_{d^{(k)}} + \dot{x}^{(k)}_{d^{(k)}} \cdot \frac{s}{100} \cdot (p_l^{(k)} - d^{(k)})$ **then**
8: Assign the current maximum as the complexity $d^{(k)} = p_l^{(k)}$.
9: **end if**
10: **end for**
11: The remaining complexity $d^{(k)}$ is considered the best subset size in the kth fold.
12: Calculate the average performance at this level of complexity for the other $K - N$ folds: $\ddot{x}^{(k)}_{d^{(k)}} = \frac{1}{K-N}\left(\sum_{l=1}^{k-N} x^{(l)}_{d^k} + \sum_{l=k+1}^{\gamma} x^{l}_{d^{(k)}}\right)$, where $\gamma = \min(K, K + k - N)$.
13: For comparison purposes, the performance for the full feature set in the kth fold ($\ddot{x}_D^{(k)}$) can also be recorded.
14: **end for**
15: Average all the K subset sizes $d^{(k)}$ obtained during the different executions of Step 11. This average is the estimate for the optimal subset size.
16: Average all the K performance estimates obtained in Step 12. This average is the estimate for the performance of the best subset having the size discovered during Step 15.

Fig. 3. Generalized $(N, K - N)$-fold cross-indexing algorithm, modified from [10]. In this study, the value $N = \frac{K}{2}$, the maximum subset size $D = 66$ and $s = 1$ were used. (When $a > b, \sum_a^b(\cdot) = 0$.)

TABLE IV
CROSS-INDEXING ESTIMATES OBTAINED IN EXPERIMENT 2.

Method	Subset Size	Accuracy (%)	Size / Acc. (%)
k-NN BE	3.5 ± 0.9	56.5 ± 2.8	-86.6 / 3.5
SVM BE	6.3 ± 1.8	54.3 ± 1.9	-84.1 / -12.2

B. Experiment 2

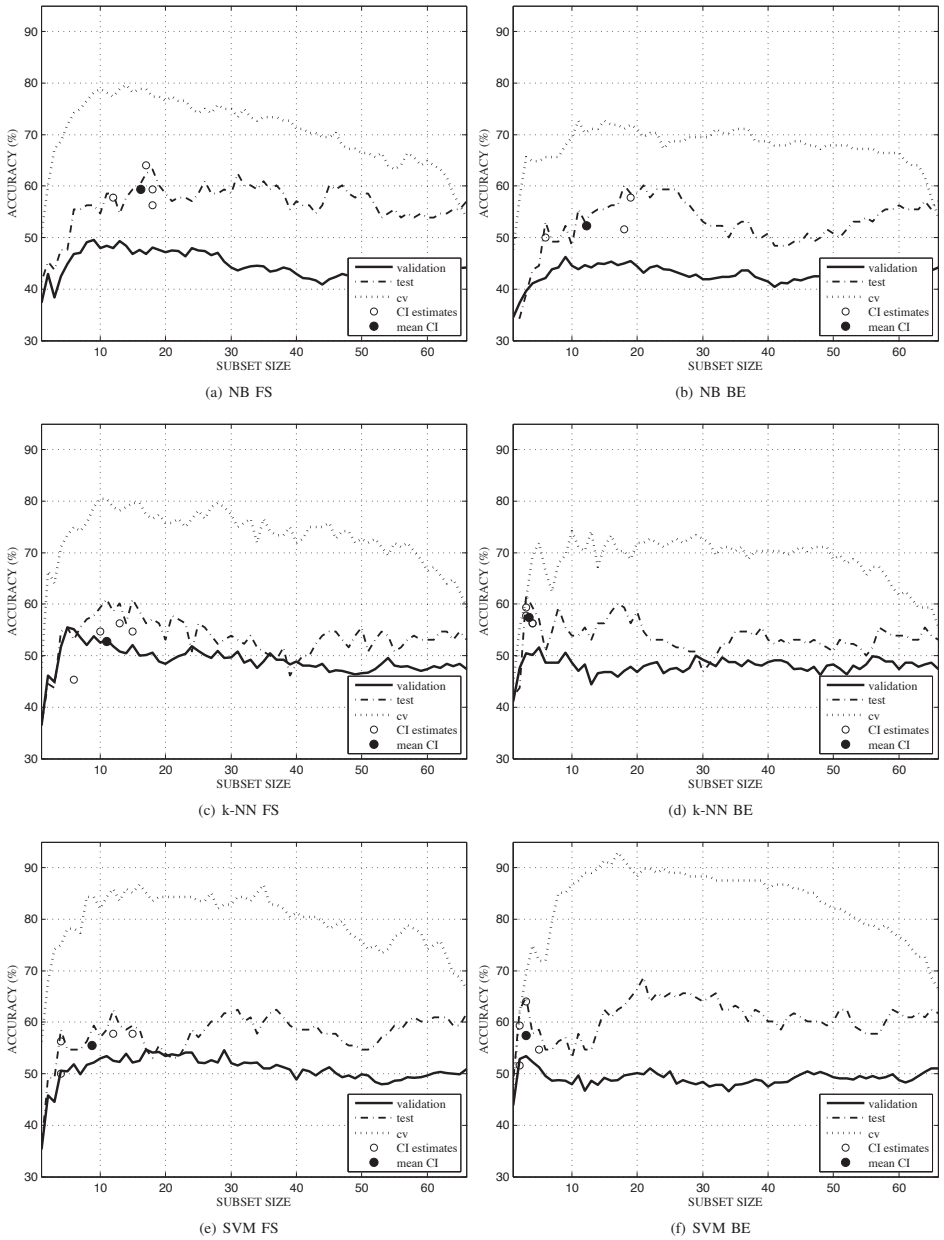
Cross-indexing was run with 30 iterations in Experiment 2 to obtain reliable estimates. Table IV shows that k-NN yielded better results than SVM in all measures. Although difference in the estimated accuracies between the methods was minor due to considerable variation, the difference in the obtained subset sizes was larger. The estimated optimal subset size was 3.5 for k-NN and 6.3 for SVM. The modification reduced the obtained subset sizes substantially with both k-NN and SVM and also increased the accuracy estimates with k-NN. However, the accuracy estimates decreased rather significantly with SVM.

Cross-indexing estimates were again compared to validation, visualized in Fig. 5(a) and Fig. 5(b). When using k-NN, it is notable that the average validation set accuracies generally decreased after adding more than six features in the set. This indicates that k-NN truly benefits from BE. Fig. 5(a) shows that the mean cross-indexing estimate of the optimal subset size corresponds to validation: the maximal validation performance level was attained with three features, which exactly matches most of the subset size estimates.

While the optimal subset size in terms of validation was estimated reliably, there exists large variation between the accuracy estimates, which are also slightly optimistic in general when compared to validation. However, it must be noted that the groups of participants who rated the primary and validation sets were different, which could have induced slight inconsistency between the sets.

Fig. 5(b) shows the performance of SVM with BE. Validation implies that SVM gives relatively good performances with subsets of 4 to 7 features, but models built on larger sets with more than 25 features generalize also well. This indicates that SVM is not overly prone to overfitting. However, the reason for adding the modification to the cross-indexing algorithm was to weight the algorithm towards finding small but effective feature sets. The figure shows that cross-indexing yielded accurate estimates of the validation set performance peaks with small subset sizes.

To test the efficiency of the modification to the cross-indexing algorithm in more detail, the mean estimates obtained with the initial algorithm ($s = 0$), represented by triangles in Fig. 5(a) and Fig. 5(b), were compared to the validation accuracies. It can be seen that the subset size estimate obtained with the initial algorithm and k-NN does not correspond to the peak in the validation accuracies, and with SVM the accuracy with the obtained subset size is over-optimistic compared to validation. This gives strong indication that the modified version of the algorithm gives more reliable estimates than the original version.

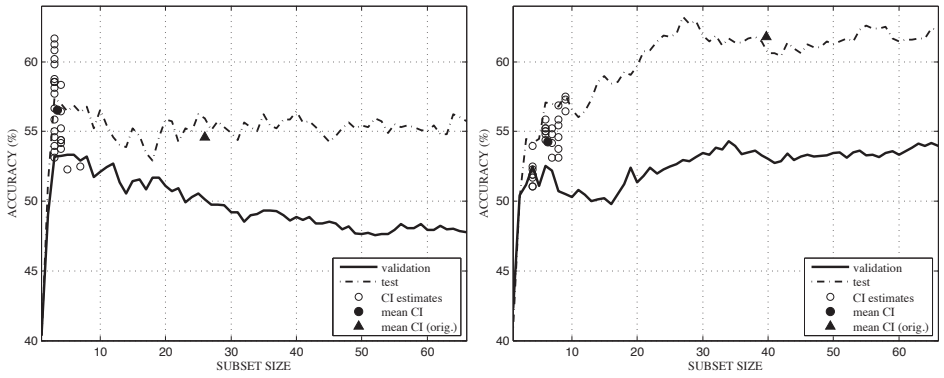


(a) NB FS (b) NB BE

(c) k-NN FS (d) k-NN BE

(e) SVM FS (f) SVM BE

Fig. 4. Cross-indexing (CI) estimates and the averaged accuracies obtained in Experiment 1.



(a) k-NN BE (b) SVM BE

Fig. 5. Cross-indexing (CI) estimates and the averaged accuracies obtained in Experiment 2.

TABLE V
CONFUSION MATRICES OF THE 4-FEATURE MODELS WITH K-NN AND 6-FEATURE MODELS WITH SVM, AVERAGED OVER 30 RUNS. VALUES ARE REPRESENTED IN PERCENTS. ALSO AVERAGED PRECISION (THE PROPORTION OF TRUE POSITIVES TO THE SUM OF TRUE POSITIVES AND FALSE POSITIVES), RECALL (THE PROPORTION OF TRUE POSITIVES TO THE SUM OF TRUE POSITIVES AND FALSE NEGATIVES), AND F-MEASURE (THE HARMONIC MEAN OF PRECISION AND RECALL) FOR THE CLASSES ARE PRESENTED.

(a) k-NN BE.

	Predicted						
Actual	Anger/Fear	Happy	Sad	Tender	Precision	Recall	F
Anger/Fear	14.5	3.9	5.0	1.7	0.60	0.58	0.58
Happy	3.4	11.8	2.4	7.4	0.59	0.47	0.50
Sad	4.9	1.8	12.8	5.6	0.52	0.51	0.50
Tender	2.2	4.2	4.5	14.2	0.50	0.57	0.52

(b) SVM BE.

	Predicted						
Actual	Anger/Fear	Happy	Sad	Tender	Precision	Recall	F
Anger/Fear	12.3	4.1	6.1	2.4	0.54	0.49	0.50
Happy	3.1	13.5	2.1	6.2	0.62	0.54	0.56
Sad	5.6	1.6	13.0	4.8	0.51	0.52	0.50
Tender	3.0	3.8	4.6	13.7	0.50	0.55	0.52

1) Confusions Between Emotions: As the overall accuracies of the models left room for improvement, the confusion patterns in validation were explored in detail. Table V(a) and Table V(b) show that most confusion between classes concentrated within class pairs *anger/fear - sad* and *happy - tender*, which represent similar affective dimension in valence (negative or positive emotions). Especially *happy* and *sad* as well as *anger/fear* and *tender* were well-separated. Most confusions between the class pairs were evident in the excerpts belonging to the class *sad*, which had tendency to be classified as *tender*. Again, these confusions, whilst infrequent, are understandable due to the fact that these two emotions are situated in the same dimension in the affective space (low arousal).

2) Best Models: Finally, the k-NN models with the obtained subset size four were further analyzed in order to find explanation for the variances in the validation accuracies and to contribute to the understanding of the feature combinations explaining emotion perception in music. The 4-feature sets were found to vary across the iterations of the cross-indexing loop – total of 33 features appeared in the 30 sets and none of the sets appeared more than once –, which explains the variance in the obtained validation accuracies. Analysis focused therefore on similarities between the sets. To that end, dissimilarity matrix D of all 30 feature sets was computed.

The dissimilarity between feature sets X and Y containing four features X_i and Y_i ($i \in \{1, 2, 3, 4\}$) was computed by

$$D_{XY} = 1 - max(\frac{1}{4}\sum_{i=1}^{4} corr(X_i, Y_{P^{(j)}_i})), \qquad (1)$$

where $corr(X_i, Y_i)$ is correlation between the features in the validation set and $P^{(j)}$ is a permutation of $\{1, 2, 3, 4\}$. The maximal averaged correlation was expected for particular permutation of Y with overlapping and highly correlating features aligned according to X.

Non-metric Multi-Dimensional Scaling (MDS) according to Kruskals Stress-1 criterion [32] into two dimensions was applied to visualize the obtained dissimilarity matrix. MDS is a set of mathematical techniques for exploring dissimilarity data by representing objects geometrically in a space of the desired dimensionality, where the distances between objects in the obtained space approximate a monotonic transformation of the corresponding dissimilarities in the original data. Interpretation of the MDS results should be done mainly in terms of inter-object distances. The stress value of the MDS result denotes the variance in the original data not accounted for in the scaled distances. The results of MDS are displayed in Fig. 6.

MDS results indicated that harmony-related features Mm

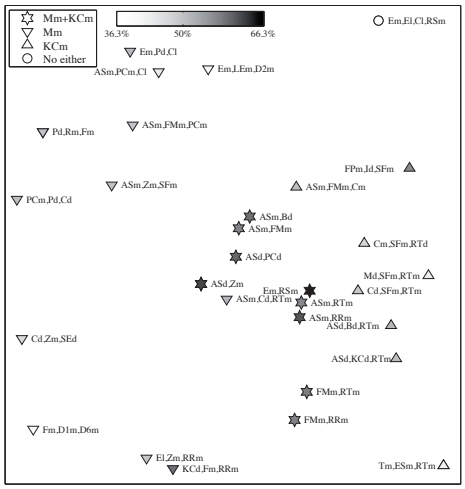


Fig. 6. Multi-dimensional scaling (stress = 0.18) of dissimilarities between four-feature sets obtained with k-NN. As a way to assess the efficiency of the feature sets, the color bar in the figure indicates the corresponding validation accuracies. For explanation of the acronyms, see Table II.

and KCm were found to be most useful for k-NN since all except one set contained at least one of them and 9 out of 10 of the most efficient sets in terms of validation accuracy contained both of the features, yielding 61.6% mean accuracy. It was also notable that Mm and KCm were almost exclusively coupled with one of the highly co-varying dynamics-related features ASm, ASd, or Em (the average inter-correlation $r_{mean} = 0.80$), one of the structural features RSm, RRm, or RTm ($r_{mean} = 0.83$) or/and rhythm-related FMm or PCd. Therefore it can be concluded that harmony and dynamics, structure and/or rhythm together constitute the most efficient models explaining emotions in the analyzed data sets.

VIII. Conclusion

The results in this study showed that classification in emotion recognition, or generally in the field of MIR, can be improved by wrapper selection when the methods are evaluated by taking into account the generalizability and simplicity of the produced models. Moreover, the proposed framework gave reliable estimates for the optimal subset sizes for the utilized classifiers, shown by the performances of the validation set. The modified cross-indexing algorithm was found to considerably improve the performance of the framework compared to the original algorithm and the outer loop of cross-validation.

The results also indicated that simple classifiers perform favorably in comparison to the more complex SVM classifier. This exhibits somewhat contrasting view to the common assumption in the field that have stated the superiority of SVM in emotion and mood recognition. Wrapper selection with k-NN and BE was found to yield the most promising performance – 56.5% classification rate with only 4 features. It was shown that relatively high level features – mode majorness and key clarity were most useful whereas most of the low-level timbral features were not found to increase the performance of the classifier. The effect of mode majorness and key clarity was congruent with their perceptual meanings.

The suggested framework is not intended to produce a single 'best' feature subset for a given classifier but rather a set of subsets for deriving general conclusions about the features useful for a chosen classifier. It must be noted that suggesting a single optimal subset would be trivial both research-wise and application-wise. Considering the sensitivity of the problem, applicability of such subset would be highly hypothetical in any utilization out of this study.

Although the results of the present framework were promising, the classification rate of 56.5% was not entirely satisfactory as previous studies have reported accuracies up to 66% [13]. However, several reasons for the lower overall accuracy can be identified. First, serious effort was made to minimize overfitting and avoid the reporting of optimistic recognition accuracies with the selected classifiers. Finding more efficient classifiers or tuning the model parameters in order to obtain higher classification accuracies was left out of the scope of this study. Secondly, the sets of only 32 excerpts used in training, standardization, and testing were rather small in order to account for all important aspects of emotions in the analyzed music style, possibly reducing the classification rates and causing high variance in the results obtained at each run. Third, the classification categories were established by a simple selection principle, which may hide the underlying patterns of closely related emotions. Hierarchical classifiers [33] could potentially uncover such structures more easily than single-level classifiers. It is clear that the performance of the framework is still essentially limited by the degree at which the features capture perceptually relevant information. Therefore the performance of the framework could be improved by studying the representativeness and reliability of the features by modifying certain parameters in the feature extraction such as the frame length or filtering. Moreover, one must acknowledge that the pursuable accuracies of classification according to perceptual categories are always bounded by the certainty at which humans recognize the concepts – previous research has reported 80% accuracies in the human recognition of clearly expressed emotions in artificial musical examples [34], [35].

Acknowledgment

The study was supported by Finnish Centre of Excellence in Interdisciplinary Music Research.

REFERENCES

[1] A. Gabrielsson and E. Lindström, "The influence of musical structure on emotional expression," in *Music and Emotion*, P. N. Juslin and J. A. Sloboda, Eds. New York: Oxford University Press, 2001, ch. 10, pp. 223–248.

[2] J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *J. Acoust. Soc. Amer.*, vol. 122, pp. 881–891, 2007.

[3] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learning Research*, vol. 3, pp. 1157–1182, 2003.

[4] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[5] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1-2, pp. 273–324, 1997.

[6] J. Reunanen, "Overfitting in making comparisons between variable selection methods," *J. Mach. Learn. Res.*, vol. 3, pp. 1371–1382, 2003.

[7] R. Fiebrink and I. Fujinaga, "Feature selection pitfalls and music classification," in *Proc. 7th Int. Conf. Music Inf. Retrieval (ISMIR'06)*, October 2006, pp. 340–341.

[8] J. Reunanen, "A pitfall in determining the optimal feature subset size," in *Proc. 4th Int. Workshop Pattern Recognition Inf. Syst. (PRIS'04)*, April 2004, pp. 176–185.

[9] ——, "Less biased measurement of feature selection benefits," in *Subspace, Latent Structure and Feature Selection*, C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, Eds. Berlin - Heidelberg, Germany: Springer, 2006, pp. 198–208.

[10] ——, "Model selection and assessment using cross-indexing," in *Proc. 20th Int. Joint Conf. Neural Netw. (IJCNN'07)*, Aug. 2007, pp. 2581–2585.

[11] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant features and the subset selection problem," in *Proc. Int. Conf. Mach. Learn.*, 1994, pp. 121–129.

[12] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*, 2nd ed. Cambridge, MA: MIT Press, 2001.

[13] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *Proc. 9th Int. Conf. Music Inf. Retrieval (ISMIR'08)*, 2008, pp. 462–467.

[14] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?" *J. Negative Results Speech Audio Sci.*, vol. 1, no. 1, pp. 1–13, 2004.

[15] T. Li and M. Ogihara, "Detecting emotion in music," in *Proc. Int. Symp. Music Inf. Retrieval (ISMIR'03)*, 2003, pp. 239–240.

[16] A. Wieczorkowska, P. Synak, R. Lewis, and Z. W.Ras, "Extracting emotions from music data," in *Proc. ISMIS'05: Foundations of Intelligent Systems*, M.-S. Hacid, N. Murray, Z. Ras, and S. Tsumoto, Eds. New York: Springer, 2005, pp. 456–465.

[17] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[18] C. Silla Jr, A. Koerich, and C. Kaestner, "Feature selection in automatic music genre classification," in *Proc. 10th IEEE Int. Symp. Multimedia (ISM'08)*, Washington, DC, Dec. 2008, pp. 39–44.

[19] J. Loughrey and P. Cunningham, "Using early-stopping to avoid overfitting in wrapper-based feature selection employing stochastic search," Dept. Comput. Sci., Trinity College, Dublin, Tech. Rep., 2005.

[20] Y.-H. Yang, C.-C. Liu, and H. H. Chen, "Music emotion classification: a fuzzy approach," in *Proc. 14th Annu. ACM Int. Conf. Multimedia (ACM MM'06)*, 2006, pp. 81–84.

[21] Y. Yaslan and Z. Cataltepe, "Audio music genre classification using different classifiers and feature selection methods," in *Proc. 18th Int. Conf. Pattern Recognition (ICPR'06)*, vol. 2, 2006.

[22] R. Fiebrink, C. McKay, and I. Fujinaga, "Combining d2k and jgap for efficient feature weighting for classification tasks in music information retrieval," in *Proc. 6th Int. Conf. Music Inf. Retrieval (ISMIR'05)*, 2005, pp. 510–513.

[23] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music." *Psychol. Music*, vol. 39, no. 1, pp. 18–49, 2011.

[24] O. Lartillot and P. Toiviainen, "Mir in matlab (ii): A toolbox for musical feature extraction from audio," in *Proc. 8th Int. Conf. Music Inf. Retrieval (ISMIR'07)*, 2007, pp. 127–130.

[25] T. Eerola and R. Ferrer, "Setting the standards: Normative data on audio-based musical features for musical genres," Poster Session presented at: 7th Triennial Conf. Eur. Soc. for the Cognitive Sci. of Music (ESCOM'09), Jyväskylä, Aug. 2009.

[26] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[27] I. H. Witten and E. Frank, *Data Mining: Practical machine learning tools with Java implementations*, 2nd ed., J. Gray, Ed. San Francisco, CA: Morgan Kaufmann, 2005.

[28] G. H. John and P. Langley, "Estimating continuous distributions in bayesian classifiers," in *Proc. 11th Conf. Uncertainty Artif. Intell.*, San Mateo, CA, 1995, pp. 338–345.

[29] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," *Ann. Stat.*, vol. 26, no. 2, pp. 451–471, 1998.

[30] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in kernel methods: support vector learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1998, ch. 12, pp. 185–208.

[31] J. H. Friedman, "On bias, variance, 0/1-loss, and the curse of dimensionality," *Data Mining Knowl. Disc.*, vol. 1, pp. 55–77, 1997.

[32] J. P. Kruskal and M. Wish, *Multidimensional Scaling*. Newbury Park, CA: Sage, 1978.

[33] S. Dumais and H. Chen, "Hierarchical classification of web content," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval (SIGIR'06)*, 2000, pp. 256–263.

[34] N. Gosselin, I. Peretz, M. Noulhiane, D. Hasboun, C. Beckett, M. Baulac, and S. Samson, "Impaired recognition of scary music following unilateral temporal lobe excision," *Brain*, vol. 128, no. 3, pp. 628–640, 2005.

[35] D. Dellacherie, N. Ehrlé, and S. Samson, "Is the neutral condition relevant to study musical emotion in patients? is the neutral condition relevant to study musical emotion in patients?" *Music Percept.*, vol. 25, no. 4, pp. 285–294, 2008.

Pasi Saari received MSc degree in Computer Science in 2008 and MA degree in Music, Mind & Technology master's programme in 2010 from the University of Jyväskylä, Finland.

He is currently working as a Doctoral Student at the Finnish Centre of Excellence in Interdisciplinary Music Research within the University of Jyväskylä, Finland. His research interests are in content-based analysis of audio and musical feature extraction and analysis.

Tuomas Eerola graduated from the University of Jyväskylä, Finland with a PhD dissertation (2003) in musicology that concerned the dynamical aspects and cross-cultural correlates of melodic expectancies.

He currently acts as a Professor of Musicology at the Department of Music at the University of Jyväskylä, Finland. He is affiliated with the Finnish Centre of Excellence in Interdisciplinary Music Research. His research interest are in music cognition, including induction of emotions and perception of melody, rhythm and musical structure in general.

Olivier Lartillot (M'10) received in parallel an Engineer degree at Supélec Grande École and an Arts Degree in musicology at Sorbonne University, Paris. He also obtained a Ph.D. degree in electrical and computer engineering from UPMC University in Paris.

Since March 2004, he has been with the Department of Music, University of Jyväskylä, where he is now an Academy of Finland Research Fellow and a member of the Finnish Centre of Excellence in Interdisciplinary Music Research. His professional interests lie in the areas of computational and cognitive modeling dedicated to sound and music analysis.

Dr. Lartillot is a member of the European Society for the Cognitive Sciences of Music (ESCOM).

II

SEMANTIC COMPUTING OF MOODS BASED ON TAGS IN SOCIAL MEDIA OF MUSIC

by

Pasi Saari & Tuomas Eerola 2014

IEEE Transactions on Knowledge and Data Engineering, 26 (10), 2548-2560

Semantic Computing of Moods Based on Tags in Social Media of Music

Pasi Saari and Tuomas Eerola

Abstract—Social tags inherent in online music services such as Last.fm provide a rich source of information on musical moods. The abundance of social tags makes this data highly beneficial for developing techniques to manage and retrieve mood information, and enables study of the relationships between music content and mood representations with data substantially larger than that available for conventional emotion research. However, no systematic assessment has been done on the accuracy of social tags and derived semantic models at capturing mood information in music. We propose a novel technique called Affective Circumplex Transformation (ACT) for representing the moods of music tracks in an interpretable and robust fashion based on semantic computing of social tags and research in emotion modeling. We validate the technique by predicting listener ratings of moods in music tracks, and compare the results to prediction with the Vector Space Model (VSM), Singular Value Decomposition (SVD), Nonnegative Matrix Factorization (NMF), and Probabilistic Latent Semantic Analysis (PLSA). The results show that ACT consistently outperforms the baseline techniques, and its performance is robust against a low number of track-level mood tags. The results give validity and analytical insights for harnessing millions of music tracks and associated mood data available through social tags in application development.

Index Terms—Semantic analysis, social tags, music, Music Information Retrieval, moods, genres, prediction.

1 Introduction

Mining moods inherent in online content, such as web forums and blogs [1], [2], images [3], and news stories [4], brings benefits to document categorization and retrieval due to the availability of large data. The need for automatic mood-based music management is increasingly important as music listening, consumption and music-related social behaviors are shifting to online sources, and a large proportion of all recorded music is found online. An extensive body of research in music psychology has shown that moods[1] are, in many aspects, fundamental to music [5]: music expresses and evokes moods, appeals to people through moods, and is conceptualized and organized according to moods. Online music services based on social tagging, such as Last.fm,[2] exhibit rich information about moods related to music listening experience. Last.fm has attracted wide interest from music researchers, since crowd-sourced social tags enable study of the links between moods and music-listening in large music collections; these links have been unattainable in the past research, which has typically utilized laborious survey-based annotations.

Social tags can be defined as free-form labels or keywords collaboratively applied to documents by users in online services, such as Del.icio.us (web bookmarks), Flickr (photos), and Pinterest (images, videos, etc.)[3]. Obtaining semantic information from social tags is, in general, a challenge not yet met. Due to the free-form nature of social tags, they contain a large amount of user error, subjectivity, polysemy and synonymy [6]. In particular, the sparsity of social tags, referring to the fact that a typical document is associated to only a subset of all relevant tags, is a challenge to the indexing and retrieval of tagged documents. Various techniques in semantic computing, such as Latent Semantic Analysis (LSA) [7], that infer semantic relationships between tags from within-document tag co-occurences, provide solutions to tackle these problems, and techniques have been proposed to automatically predict or recommend new tags to documents bearing incomplete tag information [8], [9]. However, no agreement exists on how to map tags to the space of semantic concepts, as indicated by the large number of approaches dedicated to the task [10].

In the music domain, the majority of social tags are descriptors of the type of music content, referring typically to genres [11], but also to moods, locales and instrumentations, which are well represented in the data as well. In particular, moods are estimated to account for 5% of the most prevalent tags [12]. Several studies in the field of Music Information Retrieval (MIR) have applied bottom-up semantic computing techniques, such as LSA to uncover mood representations emerging from the semantic relationships between social tags [13], [14]. These representations have resembled

• *P. Saari and T. Eerola are with the Department of Music, University of Jyväskylä, Jyväskylä, Finland.*
E-mail: pasi.saari@jyu.fi.

This work was funded by the Academy of Finland (Finnish Centre of Excellence in Interdisciplinary Music Research).

1. In this paper, we use terms mood, emotion, and affect interchangeably.

2. Last.fm: http://www.last.fm/.

3. Del.icio.us: http://www.delicious.com; Flickr: http://www.flickr.com; Pinterest: http://pinterest.com.

mood term organizations in the dimensional [15] or categorical [16], [17] emotion models, which have regularly been used to model moods in music [18]. However, we claim that the previous studies in tag-based music mood analysis have not given comprehensive evaluation of the models proposed, utilized knowledge emerging from emotion modeling to the full potential, or presented systematic evaluation of the accuracy of the models at the track level.

In this paper we propose a novel technique called Affective Circumplex Transformation (ACT), optimized to uncover the mood space of music by bottom-up semantic analysis of social tags. The key aspect of ACT is that it is a predictive model that can be used to predict the expressed moods in novel tracks based on associated tags. We train ACT with a large collection of approximately 250,000 tracks and associated mood tags from Last.fm and evaluate its predictive performance with a separate test set of 600 tracks according to the perceived moods rated by a group of participants. We compare ACT to predictive models devised based on various semantic analysis techniques, as well as to the predictions based on raw tag data. We also estimate the applicability of ACT to large collections of weakly-labeled tracks by assessing ACT performance as a factor of the number of tags associated to tracks. Furthermore, we gain insights into the general views on mood modeling in music by examining the structure of the mood semantic space inherent in social tags.

The rest of the paper is organized as follows: Section 2 goes through related work in semantic computing and emotion modeling. Section 3 describes the process of obtaining tracks and associated social tags from Last.fm and details the method for semantic analysis of the data. The semantic structures of the data are examined in Section 3.6. Section 4 presents the ACT technique and Section 5 introduces the baseline techniques for comparatively evaluating its prediction performance on listener ratings of the perceived mood in music. The test set used in the evaluation is described in Section 6. The results are presented and discussed in Section 7 and conclusions are drawn in Section 8.

2 Related Work

2.1 Semantic Analysis of Social Tags

Latent Semantic Analysis (LSA) [7], has been widely used to infer semantic information from tag data. To enable computational analysis, tag data is first transformed into the Vector Space Model (VSM) [19], representing associations between documents and tags in a sparse term-document matrix. Semantically meaningful information is then inferred from a low-rank approximation of the VSM, alleviating the problems with synonymy, polysemy and data sparsity. Low-rank approximation is typically computed by Singular Value Decomposition (SVD), but other techniques such as Nonnegative Matrix Factorization (NMF) [20] and Probabilistic Latent Semantic Analysis (PLSA) [21] have been proposed for the task as well.

SVD has been used in past research for music auto-tagging [22] and music mood modeling [13], [14]. Variants of NMF have been exploited for collaborative tagging of images [23] and user-centered collaborative tagging of web sites, research papers and movies [24]. PLSA has been used for collaborative tagging of web sites [25] and topic modeling of social tags in music [26]. In the latter paper, SVD and PLSA were compared in a task of genre and artist retrieval based on social tags for music, showing the advantage of PLSA in these tasks. Performance of SVD and NMF were compared in [27], in a bibliographic metadata retrieval task, but no significant difference was found. On the other hand, NMF outperformed SVD and PLSA in classification of text documents into mood categories [28].

2.2 Structure of Moods

Emotion modeling in psychology and music psychology research typically relies on explicit – textual or scale-based – participant assessments of emotion term relationships [15], [29], [30] and their applicability to music [31], [16], [32]. Based on these assessments, dimensional [15] and categorical [17] models of emotions have been proposed. Categorical emotion models either stress the existence of a limited set of universal and innate basic emotions [17], or explain the variance between moods by means of a few underlying affect dimensions [30] or a larger number of emotion dimensions based on factor analyses [16]. With regards to music, an ongoing related theoretical debate considers whether moods in music can most realistically be described as categories or dimensions [33]. Two variants of the dimensional models of emotions [15], [30] are particularly interesting here since these have received support in music-related research [18]. Russell's [15] affective circumplex postulates two orthogonal dimensions, called Valence and Arousal, and these dimensions are thought to have distinct physiological substrates. Thayer's popular variant [30] of this dimensional model assumes the two dimensions to be rotated by $45°$, labeling them as Tension and Energy. However, divergent views exist as to whether two dimensions is enough to represent affect. In particular, a three-dimensional model of Sublimity, Vitality and Unease has been proposed as underlying dimensions of affect in music [16], whereas a model of Arousal, Valence and Dominance has been proposed as a normative reference for English words [34].

Importantly, these models lend themselves to a coherent spatial representation of the individual affect terms, which is valuable property with respect to semantic analysis of mood-related social tags.

Past accounts of mood detection in MIR have applied the various emotion models coupled with advanced techniques of machine learning and signal processing to identify acoustic substrates for moods. Both categorical [35] and dimensional [36] models of emotions have been used to represent the mood in music tracks. These studies prove that insights and findings from emotion modeling research are useful to new computational approaches to automatic mood modeling. Moreover, and as noted above, past studies have recovered mood spaces based on semantic analysis of social tags that resemble the emotion models [13], [14]. Here, we go further by quantifying the predictive value of applying insights from the psychology of emotion to the analysis of large-scale and diffuse meta-data, such as information provided by social tags.

3 Semantic Analysis

This section describes the process of collecting tracks and associated social tags from Last.fm, and details the semantic analysis method to infer spatial representation of tags.

3.1 Gathering Vocabularies

To obtain the corpus of tracks and associated tags from Last.fm, we systematically crawled the Last.fm online database through a dedicated API[4]. We used extensive vocabularies of mood- and music genre-related terms as search words for populating the corpus. This approach suits our purposes since it controls the relevance of the track content and to some degree balances the data according to mood and genre.

The mood vocabulary was aggregated from several research papers and from an expert-generated word list. A number of research fields provided relevant sources: affective sciences [37], music psychology studying the use of emotion words in music [16], [38], [15] and MIR [39], [14], studying the mood prevalence in social tags. As an expert-generated source we used an extensive mood word list at Allmusic[5] web service. The vocabulary was then edited manually to identify inflected terms, such as "depressed", "depressing", "depression" and "depressive".

Genre vocabulary was aggregated from several expert-generated sources. A reference for music genres and styles available through Allmusic[6] was used as the main source. This included over 1,000 popular and classical music styles. Moreover, we included several Finnish music styles out of curiosity. Manual editing was then carried out for the genre vocabulary to aggregate regular alternate spellings, such as "rhythm and blues", "R'n'B", and "R&B" as well as "indie pop" and "indiepop".

The number of terms in the resulting vocabularies was 568 for moods and 864 for genres (1,083 and 1,603 including the inflected forms, respectively).

Moreover, the following reference vocabularies were collected for evaluating the mood structures in Section 3.6: **Locations – 464 terms:** Country names including nationality-related nouns and adjectives (e.g., "Finland", "Finn", "Finnish"), as well as continents and certain geographical terms (e.g., "arctic"). **Instruments – 147 terms:** Comprehensive list of instrument names. **Opinions – 188 terms:** Manually identified from the tags associated to more than 1,000 tracks, and not included in the other vocabularies (e.g., "favorite", "one star", "wicked", "check out").

3.2 Fetching Tags and Tracks from Last.fm

The mood and genre vocabularies, including the inflected terms, were used as search words via the Last.fm API[7] to populate the track corpus. The process is visualized in Fig. 1.

The tracks were collected using two tag-specific API functions: *tag.getTopTracks* returning up to 50 top tracks and *tag.getSimilar* returning up to 50 most similar tags. First, the top tracks for each term were included in the corpus, amounting to up to $2,686 \times 100 = 134,300$ tracks. In parallel, for each term we fetched the similar tags and included the associated top tracks. This process potentially visited up to $2,686 \times 50 \times 50 = 6,715,000$ tracks, and using both fetching processes combined we were able to fetch up to approximately 7M tracks. In practice, the number was reduced by many overlapping tracks and similar tags.

Finally, track-level tags in the final corpus were fetched using the function *track.getTopTags*, returning up to 100 tags. The returned track-level tags are represented by normalized "counts" indicating the relative number of times each tag has been applied to a track. Although the exact definition of these counts is not publicly available, they are often used in semantic analysis [12], [14]. All tags were cleaned by lemmatizing [40] and by removing non-alphanumeric characters. The final set consisted of 1,338,463 tracks and 924,230 unique tags.

3.3 Vector Space Modeling

A standard Vector Space Model (VSM) [19] was built separately for each of the vocabularies. Tags related to the vocabulary terms were identified from the corpus following the bag-of-words approach also taken in [26]. All tags that included a term as a separate word (or separate consecutive words in the case of multi-word terms) were associated with the corresponding

4. http://www.last.fm/api, accessed during November - December 2011.
5. http://www.allmusic.com/moods
6. http://www.allmusic.com/genres
7. Find detailed information on the used functions from the API documentation referenced above.

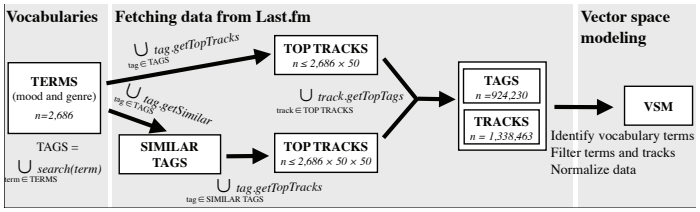


Fig. 1. Data collection process.

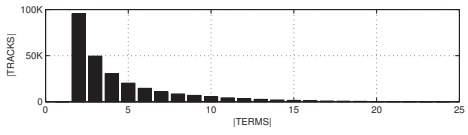


Fig. 2. Distribution of the number of mood terms associated to tracks in the test set.

TABLE 1
Statistical measures related to each vocabulary-related corpus.

	|Tracks|	|Terms|	# Terms per track (Avg.)
Moods	259,593	357	4.44
Genres	746,774	557	4.83
Instruments	46,181	53	4.44
Locales	72,229	126	2.26
Opinions	305,803	192	5.67

terms. We also filtered out those track-specific term associations where a term was included in either track title or artist name. This was due to the fact that many social tags describe these actual track metadata.

To avoid obtaining overly sparse and uncertain information, we excluded all terms that were associated to less than 100 tracks. At this point 493,539 tracks were associated to at least one mood term. However, we excluded tracks associated to only one mood term, as it was assumed that these tracks would provide little additional information for the further semantic analysis of mood term relationships. This resulted in a corpus of 259,593 tracks and 357 mood terms. As shown in Fig. 2, distribution of the number of terms associated to each track was exponential, indicating the sparsity of the data. Similar procedures were applied to all other vocabularies as well. Statistical measures related to the resulting corpora are shown in Table 1. The five most frequently applied within each corpora are as follows: **Moods**: "chill", "mellow", "relaxing", "dark" and "melancholy"; **Genres**: "rock", "pop", "alternative", "electronic" and "metal"; **Instruments**: "guitar", "bass", "drum", "piano" and "acoustic guitar"; **Locales**: "British", "UK", "American", "USA" and "German"; and **Opinions**: "favorite", "love", "beautiful", "awesome" and "favourite".

Finally, the normalised counts $n_{i,j}$ provided by Last.fm for term (w_i) – track (t_j) associations were used to form the VSM N defined by Term Frequency-Inverse Document Frequency (TF-IDF) weights $\hat{n}$ in a similar manner as in [26]:

$$\hat{n}_{i,j} = (n_{i,j} + 1) \log(\frac{R}{f_i}), \quad (1)$$

where R is the total number of tracks, f_i is the number of tracks term w_i has been applied to. Separate models were formed for each vocabulary-related corpora.

3.4 Singular Value Decomposition

SVD is the typical low-rank matrix approximation technique utilized in LSA to reduce the rank of the TF-IDF matrix, alleviating problems related to term synonymy, polysemy and data sparsity. SVD decomposes a sparse matrix N so that $N = USV^T$, where matrices U and V are orthonormal and S is the diagonal matrix containing the singular values of N. Rank k approximation of N is computed by $N^k = U^k S^k (V^k)^T$, where the i:th row vector U_i^k represents a term w_i as a linear combination of k dimensions. Similarly, V_j^k represents track t_j in k dimensions. Based on a rank k approximation, dissimilarity between terms w_i and $w_{\hat{i}}$ is computed by the cosine distance between $U_i^k S^k$ and $U_{\hat{i}}^k S^k$.

In the present study, all data sets summarized in Table 1 are subjected to LSA. While the main content of this paper deals with the Mood corpus, we use Genres to balance our data sampling in Section 6, and the other sets for comparison of different concepts in Section 3.6.

3.5 Multidimensional Scaling

Past research in emotion modeling, reviewed above, suggests two to three underlying dimensions of emotions, which indicates that very concise representation of the mood data at hand would successfully explain most of its variance. Therefore, we develop further processing steps to produce a semantic space of moods congruent with the dimensional emotion model. Genres, Locales, Instruments and Opinions were subjected to the same procedures to allow comparative analysis described in Section 3.6.

We applied non-metric Multidimensional Scaling (MDS) [41] according to Kruskal's Stress-1 criterion into three dimensions on the term dissimilarities produced by SVD with different rank k-values. MDS is a set of mathematical techniques for exploring dissimilarity data by representing objects geometrically in a space of a desired dimensionality, where the distances between objects in the obtained space approximate a monotonic transformation of the corresponding dissimilarities in the original data. When used with a low number of dimensions, MDS allows for concise representation of data, which is why it is a typical tool for data visualization. In particular, [42] showed with several high-dimensional biochemical data sets that the combination of SVD followed by MDS is more efficient at dimension reduction than either technique alone.

The resulting mood and genre term configurations with $k = 16$ are shown in Fig. 3. The stress ϕ_k, indicating the goodness-of-fit varied between ($\phi_4 = 0.02, \phi_{256} = 0.29$) depending on the rank k. Similar values were obtained for both moods and genres.

To represent a track in the MDS term space, we applied projection based on the positions of the associated terms. Given an MDS term configuration $y_i = (y_{i1}, y_{i2}, y_{i3})$, $i \in (1, ..., |w|)$, position of a track represented by a sparse TF-IDF-weighted term vector q is computed by the center-of-mass:

$$\hat{t} = \frac{\Sigma_i q_i y_i}{\Sigma_i q_i}. \quad (2)$$

For example, the position of a track associated to "happy", with no other terms assigned, coincides with the position of the term. On the other hand, a track with "happy" and "atmospheric" is positioned along the segment *happy–atmospheric*. In general, tracks are located in the MDS space within a *convex polyhedron* with vertices defined by positions of the associated terms.

3.6 Mood Structures Emerging from the Semantic Data

Because of the different views on how to treat mood-related data, whether as categories or dimensions, we used semantic information of music tracks obtained by the MDS analysis to gain evidence on this issue. If tracks in the MDS space would have clear cluster structure, we should choose the categorical representation; whereas, if tracks would scatter somewhat evenly across the space, continuous description of moods would be appropriate.

Hopkins' index [43] can be used to estimate the degree of clusterability of multidimensional data. It is based on the hypothesis that the clustering tendency of a set of objects is directly reflected in a degree of non-uniformity in their distribution. Non-uniformity is estimated by comparing the sum of

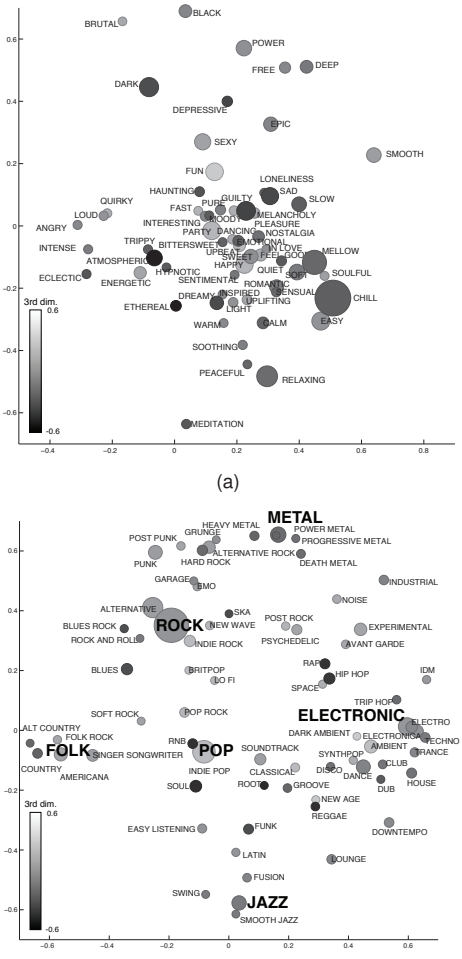


Fig. 3. MDS configurations ($k = 16$) of (a) mood and (b) genre terms in three dimensions (bubble size = prevalence, where prevalence $\geq$ 4,000 and 10,000 for (a) and (b)). Six highlighted genres refer to listening experiment (see Section 6).

nearest-neighbor distances R_j within a set of real objects to the sum of distances A_j between artificial objects and their nearest real neighbors:

$$H = \frac{\sum A_j}{\sum A_j + \sum R_j}. \quad (3)$$

Following an extension by [44], artificial objects are sampled from univariate distributions that match

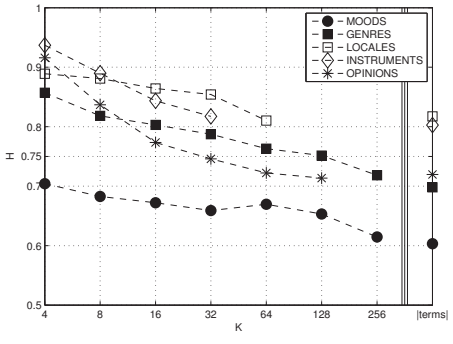


Fig. 4. The mean Hopkins' indices for each vocabulary-related corpus across various k ($sd \leq .0067 \; \forall \; k$).

those of the real objects. Value $H \approx 0.50$ indicates uniform structure ($\sum R_j \approx \sum A_j$), whereas $H \approx 1.0$ indicates perfect clusterability. In particular, the value $H = 0.75$ indicates that null hypothesis of uniform structure can be rejected at 90% confidence level.

For each corpus, we computed Hopkins' index for the track positions in the MDS spaces (see Eq. 2) obtained with ranks $k = (4, 8, 16, ..., 256)$ and $k = |terms|$. The latter corresponds to computing MDS without LSA, i.e. based on term distances in the original TF-IDF matrices. Preliminary analyses indicated that Hopkins' index is affected by the number of terms associated to each track. Since the characteristics of the vocabulary-related corpora differed in this respect, we randomly sampled for each corpus a subset of 4088 tracks with exponential terms-per-track distribution ($2048 + 1024 + 512 + ... + 8$ tracks associated to $2, 3, 4, ..., 10$ terms, respectively) and computed H for the subset. The results shown in Fig. 4 are computed as an average of ten separate runs of this process.

The results showed that Hopkins' indices for Moods remained at the range of $0.6 < H < 0.7$, which means that track positions are uniformly distributed across the mood space. This suggests that the optimal representation of Moods is continuous rather than categorical. Comparison to other corpora supports this view, as mood values remain at a clearly lower level than those of any other set. Genre-related values indicated that genre-data is fairly clusterable ($H > .75$, when $k \leq 128$), supporting the common practice of assigning songs categorically into genres. Furthermore, semantic spaces of Instruments and Locales had the clearest cluster structure. This is in line with the intuition that music tracks can, in general, be characterized with distinct instruments or instrumentations and geographical locations. Clusterability of data related to Opinions was in general at the same level as that of Genres. However, Opinions yielded particularly high values of H with low k. We consider this high dependence on k as an artefact caused by ill-conditioned distances between Opinion terms: almost all of the most prevalent terms were highly positive ("favorite", "killer", "amazing", "awesome", etc.), and the computed distances between these terms may not reflect any true semantic relationships.

In summary, the results support the use of the dimensional representation of mood information of music tracks. In the next section we develop further processing steps to comply with this finding.

4 Affective Circumplex Transformation

Typical MDS operations, described above, may not be adequate to characterize moods, since the dimensions obtained do not explicitly represent the dimensional models of emotion. We therefore propose a novel technique called *Affective Circumplex Transformation* (ACT) influenced by Russell's affective circumplex model of emotions [15] to conceptualize the dimensions of the MDS mood spaces. First, reference positions for mood terms on the Valence-Arousal (VA) space are obtained from past research on emotion modeling. Then, the MDS space is linearly transformed to conform to the reference. Finally, explicit mood information of music tracks is computed by projecting those onto the transformed space.

4.1 ACT of Mood Term Space

Reference locations for a total of 101 unique mood terms on the VA space were extracted from Russell's [15, p. 1167] and Scherer's [29, p. 54] studies. In the case of seven overlapping mood terms between the two studies, Scherer's term positions were chosen since they are scattered on a larger part of the plane and thus may provide more information. Furthermore, the model by [30] was projected on the space diagonally against the negative valence and positive arousal to obtain explicit representation of the tension dimension.

Three-dimensional MDS spaces were conformed to the extracted VA space by first identifying the corresponding mood terms in the semantic data. Identification of mood terms resulted in a set of 47 mood terms out of the 101 candidates. The fact that less than half of the mood terms used in the past studies exist in the semantic mood data may indicate the difference between affect terms used to describe everyday experiences in general versus terms used in the context of the aesthetic experience.

Transformation of the MDS space to optimally conform to the VA reference was determined by classical Procrustes analysis [45], using sum of squared errors as goodness-of-fit. Given the MDS configuration $y_{\hat{i}} = (y_{\hat{i}1}, y_{\hat{i}2}, y_{\hat{i}2})$ and VA reference $x_{\hat{i}} = (x_{\hat{i}1}, x_{\hat{i}2})$ for

mood terms $\hat{i}$ matched between the two, Procrustes transformation gives $\hat{x}_{\hat{i}} = By_{\hat{i}}T + C$, where B is an isotropic scaling component, T is an orthogonal rotation and reflection component, and C is a translation component. B, T, and C minimize the goodness-of-fit measure $X^2 = \Sigma_{\hat{i}}(x_{\hat{i}} - \hat{x}_{\hat{i}})^2$. Based on the components, configuration $\hat{x}_i$ including all mood terms can be obtained by

$$\hat{x}_i = By_iT + C. \tag{4}$$

Procrustes retains the relative distances between objects since it allows only translation, reflection, orthogonal rotation and isotropic scaling. Therefore, the relative configuration of the terms in the original MDS space is not affected. Changing the rank parameter in SVD had no significant effect on the goodness-of-fit of the Procrustes transformation. The criterion varied between $0.75 < X^2 < 0.79$.

A peculiarity of ACT is in conforming the three-dimensional MDS space to two-dimensional reference. The transformation is thus provided with an additional degree of freedom, producing two explicitly labeled dimensions and a third residual dimension. Using three dimensions in the MDS space is based on the unresolved debate of whether the underlying emotion space is actually two- or three-dimensional (see Section 2.2).

Fig. 5 shows the transformed mood term configuration based on SVD with rank 16, also indicating Russell's dimensions of Arousal and Valence, and Thayer's dimensions of Energy and Tension. VA-reference and the transformed term positions correspond well, in general, as they are located roughly at the same area of the space. For example, positions of terms "happy", "joy", "sad", "tense" and "peaceful" have only minor discrepancy between the reference. Moreover, dimension labels and the dimensions implied by the mood term organization correspond as well and the positions of popular mood terms not used as reference for the transformation make sense in general. For example, "fun", "party" and "upbeat" all have positive valence and arousal, "dark" has negative valence and negative arousal, whereas "brutal" has negative valence and positive arousal.

However, certain terms such as "solemn", "delight", "distress" and "anxious" show larger discrepancy, and the terms "atmospheric" and "ethereal", which could intuitively be considered as neutral or even positive, both have negative valence. The cause of these inconsistencies could again be traced back to the difference between aesthetic and everyday affective experience, but could also be due to the subjectivity of mood-related associations in music listening. For example, a solemn or atmospheric track that one enjoys may be regarded as depressing by another. This multi-faceted aspect of music listening is discussed in [32].

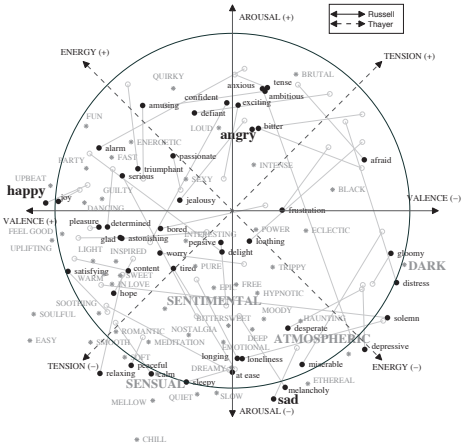


Fig. 5. Mood space obtained by ACT based on *rank* 16 *semantic space*. Mood term positions are shown in black dots and the reference positions in grey circles. Mood terms having no reference, but that are associated to at least 4,000 tracks are shown in grey stars. Highlighted terms relate to the seven scales rated in listening experiment (see Section 6).

4.2 Mood Prediction of Tracks with ACT

In this section we describe how ACT can be used to predict the prevalence of moods in novel tracks that have been associated to one or more mood terms. The prediction performance of ACT related to mood dimensions and individual mood terms is evaluated later in Section 7 with listener ratings of the perceived mood in a separate test set.

Projection in Eq. 2 can be used in itself to estimate the valence and arousal of a track – the estimations are represented explicitly by the dimensions in the projection. However, in order to estimate the prevalence of a certain mood term in a track, another projection is needed.

We assigned continuous mood term-specific weight for a track by projecting the track position given by Eq. 2 in the MDS space along the direction determined by the term. A track with position $\hat{t} = (\hat{t}_1, \hat{t}_2, \hat{t}_3)$ in the transformed mood space was projected according to the direction of a mood term with position $\hat{x}_i = (\hat{x}_{i1}, \hat{x}_{i2}, \hat{x}_{i3})$ by

$$P_i = \frac{\hat{x}_i}{|\hat{x}_i|} \cdot \hat{t}, \tag{5}$$

where $|\cdot|$ denotes the l^2-norm. To obtain an estimate for Tension we projected the track along the direction $(-1, 1, 0)$ (note the inverted valence axis according to a convention used in emotion modeling).

5 Baseline Techniques for Mood Prediction

We compared the performance of ACT in predicting moods of tracks to several techniques based on low-rank matrix approximation. For a track t represented by a sparse TF-IDF-weighted term vector q, we computed rank k mood weight related to term w_i with SVD, NMF and PLSA. All of the techniques involve computing the low-rank approximation of the TF-IDF matrix, transforming an unknown track t to the VSM by Eq. 1 and folding it into the low-rank semantic space, and approximating the weight of a mood w_i related to the track. In addition to the low-rank approximation of the VSM, we used the original sparse VSM representation q as a baseline as well.

5.1 Singular Value Decomposition

Track represented with a sparse TF-IDF-weighted term vector $q = (q_1, q_2, ..., q_{|w|},)$ is first folded in to the rank k space obtained with SVD by:

$$\hat{q}^k = (S^k)^{-1}(U^k)^T q. \tag{6}$$

The weight N_i^k related to the track and a mood term w_i is then computed by

$$N_i^k = U_i^k S^k (\hat{q}^k)^T. \tag{7}$$

5.2 Nonnegative Matrix Factorization

NMF [20] is a method proposed for low-rank approximation of a term-document matrix. The method distinguishes from SVD by its use of nonnegative constraints to learn parts-based representation of object semantics. Given a nonnegative TF-IDF matrix $N \subset \mathbb{R}^{C \times D}$ and a desired rank parameter k, NMF constructs nonnegative matrices $W^k \subset \mathbb{R}^{C \times k}$ containing k basis components and $H^k \subset \mathbb{R}^{k \times D}$ such that $N \approx W^k H^k$. This is done by optimizing

$$\min_{W^k, H^k} f(W^k, H^k) = \frac{1}{2}||N - W^k H^k||_F^2, \; s.t. \; W^k, H^k > 0, \tag{8}$$

where F denotes the Frobenius norm. We solve the optimization problem using multiplicative updating rules in an iterative manner [20]. The ith row of W can be interpreted as containing k "importance" weights a mood term w_i has in each basis component. Similarly, the jth column of H can be regarded as containing k corresponding weighting coefficients for track t_j.

Folding in a new track represented by a TF-IDF-weighted term vector q to obtain $\hat{q}^k$ is achieved by solving an optimization problem by keeping H^k fixed:

$$\min_{\hat{q}^k} f(\hat{q}^k, H^k) = \frac{1}{2}||q - \hat{q}^k H^k||_F^2, \; s.t. \; \hat{q}^k > 0. \tag{9}$$

Finally, to estimate the weight N_i^k related to track t and mood term w_i, we compute

$$N_i^k = W_i^k \hat{q}^k. \tag{10}$$

5.3 Probabilistic Latent Semantic Analysis

In the core of PLSA [21], is the statistical *aspect model*, a latent variable model for general co-occurrence data. Aspect model associates an unobserved class variable $z \in Z = (z_1, ..., z_k)$ with each occurrence of a term w_i in a track t_j.

PLSA states that the probability $P(t_j, w_i)$ that term w_i is associated with a track t_j can be expressed as a joint probability model using latent class variable z:

$$P(t_j, w_i) = P(t_j)P(w_i|t_j) = P(t_j) \sum_{z \in Z} P(w_i|z)P(z|t_j), \tag{11}$$

where $P(t)$ is the probability of a track t_j, $P(z|t_j)$ is the probability of a latent class z in track t_j, and $P(w_i|z)$ is the probability of a term w_i in the latent class. The model is fitted to the collection of tracks by maximizing log-likelihood function

$$L = \sum_t \sum_w N_{i,j} log P(t_j, w_i), \tag{12}$$

where $N_{i,j}$ is the nonnegative TF-IDF matrix. The procedure for fitting the model to training data is the Expectation Maximization (EM) algorithm [21]. To estimate the probability $P(q, w_i)$ of a mood term w_i for a new track represented by a TF-IDF-weighted term vector q, we first fold in the track using EM, keeping the parameters of $P(w_i|z)$ fixed and then calculate weights $P(z|q)$. The mood weight for the track is finally computed by

$$P(q, w_i) = P(w_i|z)P(z|q). \tag{13}$$

5.4 Predicting the Mood Dimensions

Since all baseline techniques predict mood primarily according to explicit mood terms, the techniques must be optimised to achieve mood dimension predictions comparable to ACT. We considered that a mood term representative of a mood dimension would yield the highest predictive performance for the corresponding dimension. We assessed the representativeness of the mood terms by computing the angle between each mood dimension and mood term location in the ACT configurations with $k \in [4, 8, 16, ..., 256]$, and limited the choice to terms associated to at least 10% of all tracks in the corpus. This yielded the following terms, indicating the number of track associations and the maximum angle across k between the term position in the ACT configurations and the corresponding dimension: "happy" for Valence ($n = 28,982$, $\alpha_k \leq 9.29°$), "melancholy" for Arousal ($n = 31,957$, $\alpha_k \leq 5.11°$) and "mellow" for Tension ($n = 46,815$, $\alpha_k \leq 4.48°$)

6 Ground-Truth Data of Moods in Music

We evaluated the performance of ACT and the baseline techniques by comparing the estimates produced

by these methods to listener ratings of the perceived moods in music tracks. Participants listened to short music clips (15s) and rated their perception of moods expressed by music in terms of ten scales. The test set of tracks was retrieved from the Last.fm in a random fashion, balancing the sampling to cover semantic genre and mood spaces. This section describes the ground-truth collection process in detail[8].

6.1 Choosing Moods and Genres as Focus

To systematically cover the concurrent characterizations of moods in music, ratings were done for both the dimensional mood model and individual mood terms. All ratings were given in nine-step Likert-scales to capture the continuous nature of mood uncovered in Section 3.6. We used bipolar and unipolar scales for the mood dimensions and terms, respectively.

For dimensional model we used three scales: Valence, Arousal and Tension, later denoted as VAT; whereas for the mood term representation we used seven scales: Atmospheric, Happy, Dark, Sad, Angry, Sensual and Sentimental. The choice was based on several criteria: *i)* to cover the semantic space as well as the basic emotion model; *ii)* to use representative terms as implied by high prevalence in the data ("sentimental" used 4,957 times – "dark" 33,079 times); and *iii)* to comply with research in the affect prevalence and applicability in music [31], [16], [32].

Six popular and distinct genres according to the Last.fm track collection (see Fig. 3 (b)) – Rock, Pop, Electronic, Metal, Jazz and Folk – were chosen as the focus of the study to retain a wide variance in the stylistic characteristics of popular music.

6.2 Sampling of Tracks

We fetched a set of 600 tracks from Last.fm, separate to the mood track corpus used in the semantic modeling, to be rated in the listening experiment. To obtain a track collection that allows multifaceted comparison between tag information and the ratings, we utilized balanced random sampling of tracks based on: *i)* mood coverage – reciprocal of the track density in the rank 16-based MDS mood space; and *ii)* genre coverage – closeness of track positions in the MDS genre space to one of the six chosen genre terms. Moreover, quality and variability of semantic information in the data was ensured by: *i)* favoring tracks associated to many mood tags; *ii)* favoring tracks with many listeners according to statistics provided by Last.fm; and *iii)* choosing no more than one track from each artist.

Tracks in the resulting test set are associated with 8.7 mood terms on average, which is a higher number than that of the larger mood set due to sampling according to the number of associated mood terms. The details of the term occurrences are shown in Fig. 6. The number of times each mood term related to the chosen scales appear in the set are: 90 (Atmospheric), 137 (Happy), 109 (Dark), 166 (Sad), 28 (Angry), 43 (Sensual) and 52 (Sentimental). For genres, the corresponding figures are: 422 (*rock*), 353 (*pop*), 149 (*electronic*), 139 (*metal*), 147 (*jazz*) and 144 (*folk*). Considering the high frequency of genres such as *rock* and *pop*, these genres have naturally wider representation in the set – a track in the *electronic* genre has likely been tagged with *pop*, for instance.

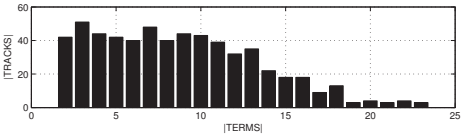


Fig. 6. Distribution of the number of mood terms associated to tracks in the test set.

6.3 Listening Experiment

An online interface was used to allow participants to login on their own computers and save their ratings on a server in real time. At each session, tracks were presented in a randomized order. Participants were allowed to rate as many or as few songs as they liked. However, to encourage the rating of many tracks, the task was rewarded by Spotify[9] and Amazon[10] gift cards proportional to the amount of tracks rated.

The task was to rate 15 second clips of the tracks in terms of the perceived moods *expressed* by music, rather than moods *induced* by music. VAT scales were presented with bipolar mood term labels: "negative"/"positive", "calm"/"energetic" and "relaxed"/"tense", respectively. In addition to mood, participants rated their personal liking of the tracks, and in half of the cases, genre representativeness. In this paper, however, we utilize only the mood ratings.

We based the sampling of each song on the audio previews on Last.fm service, arguing that, since the previews are track summarizations sampled for marketing purposes, consisting of the most prolific section, they are fairly representative of the full tracks. The previews typically consist of a build-up and part of chorus, starting either at 30 or 60 seconds into the beginning. While some studies have highlighted the difference between clip- and track-level content [46], it has been argued that using short clips lessens the burden of human evaluation and reduces problems in annotation caused by time variation of moods [47].

A total of 59 participants, mostly Finnish university students (mean age 25.8 years, SD = 5.1 years, 37 fe-

8. Ground-truth and semantic mood data are publicly available at http://hdl.handle.net/1902.1/21618.

9. http://www.spotify.com/

10. http://www.amazon.co.uk/

TABLE 2
Correlations (r_s) between mood ratings.

	Valence	Arousal	Tension
Valence		$-.073$	$-.639^{***}$
Arousal			$.697^{***}$
Atmospheric	$.180^{***}$	$-.901^{***}$	$-.687^{***}$
Happy	$.940^{***}$	$.114^{**}$	$-.478^{***}$
Dark	$-.940^{***}$	$.059$	$.640^{***}$
Sad	$-.413^{***}$	$-.662^{***}$	$-.253^{***}$
Angry	$-.687^{***}$	$.633^{***}$	$.876^{***}$
Sensual	$.320^{***}$	$-.733^{***}$	$-.688^{***}$
Sentimental	$.114^{**}$	$-.722^{***}$	$-.621^{***}$

Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$, $df = 599$.

males), took part in the experiment. Musical expertise of the participants spanned from listeners ($N = 23$), to musicians ($N = 28$) and trained professionals ($N = 8$). Each participant rated 297 clips on average, and 22 participants rated all 600 clips. Cronbach's alpha for mood scales vary between 0.84 (sentimental) and 0.92 (arousal), which indicates high internal consistency [48]. Such high agreement among the participants gives support for (a) using all participants in further analysis, and (b) representing each song by single value on each mood scale, computed as the average across participants.

Spearman's rho correlations (r_s) between mood ratings in different scales, presented in Table 2, showed no correlation between valence and arousal, which supports treating these moods as separate dimensions. On the other hand, tension is highly correlated with arousal and negative valence, which in turn supports projecting tension diagonally against these dimensions. Ratings of all 7 mood terms are highly related to valence (happiness, darkness), arousal (atmospheric, sentimental), or a combination of these (sad, angry, sensual). This extends previous findings about high congruence between term-based and dimensional emotion models in emotion ratings of film soundtracks [49] to a large variety of tracks in popular music genres.

7 Results and Discussion

We compared the prediction rates of ACT with various rank values $k \in (4, 8, 16, ..., 256)$ to those of the baseline techniques SVD, NMF, PLSA and VSM. All prediction rates were computed by correlating the estimates with the listener ratings of moods, using Spearman's rank correlation (r_s). Fig. 7 shows the results in detail with different rank k values, while Table 3 summarizes the results into the average performance across k, assessing also the significance of the performance differences between ACT and the baseline techniques. Section 7.3 (Table 3: ACT alt.) provides results obtained with alternative configurations of ACT. Finally, Section 7.4 assesses the performance of ACT as a factor of the number of terms applied to tracks in the test set.

7.1 Performance for VAT Dimensions

Fig. 7 shows that ACT yielded the highest performance for all VAT scales, outperforming the baseline techniques consistently across k. For Valence the median performance of ACT was $r_s = .576$, varying between $.519 < r_s < .606$. The performance was slightly higher for Arousal (Mdn $r_s = .643$, $.620 < r_s < .683$) and Tension (Mdn $r_s = .622$, $.585 < r_s < .642$). Performance difference to the baseline techniques was significant for all scales – NMF gave the highest median performances ($r_s = .348, .514, .579$), while SVD performed the worst ($r_s = .302, .414, .443$) at predicting Valence, Arousal and Tension, respectively. VSM yielded performance levels comparable to the baseline methods, outperforming SVD for all three scales, and PLSA for Valence and Arousal. However, devising baseline techniques to infer predictions for VAT scales from highly prevalent mood terms possibly benefits VSM more than the other techniques. While SVD, NMF and PLSA utilize the semantic relationships with other terms in making predictions, VSM predictions rely solely on the individual terms. The chosen mood terms are popular also within the test set ($n = 137, 189, 227$ for "happy", "melancholy" and "mellow", respectively).

The results also show that ACT is less sensitive to the value of k than SVD, NMF and PLSA. While ACT performance varied by $\Delta r_s \leq .087$, SVD ($\Delta r_s \leq .222$) and PLSA ($\Delta r_s \leq .412$) were clearly more inconsistent. For Valence and Arousal, PLSA yielded particularly low performance with $k < 16$. NMF was more robust than other baseline techniques against k as shown by the performance differences of $\Delta r_s \leq .112$.

The high prediction rate of Arousal compared to that of Valence bears similarity to the results from prediction of affect dimensions from the musical features across different genres of music [50]. This was also highlighted by an analysis of ACT prediction rates at the genre-level. The median r_s across k for subsets of the test tracks associated to different main genres was consistently high for Arousal regardless of genre ($.585 < r_s < .701$), whereas for Valence the rates spanned $r_s = .390$ (Jazz) and $r_s = .614$ (Metal).

In summary, the results suggest that conventional techniques of semantic analysis are inadequate at reliably inferring mood predictions congruent with the dimensional model of emotions, whereas ACT yields consistently high performance at this task.

7.2 Performance for Individual Mood Terms

Since the rated mood term scales relate to the mood term associations explicitly represented in the test set, comparison between ACT and the baseline techniques is more direct than with VAT dimensions. Still, the same patterns in the performances were prevalent. ACT, again, clearly gave the highest overall performance, while NMF was the most successful baseline

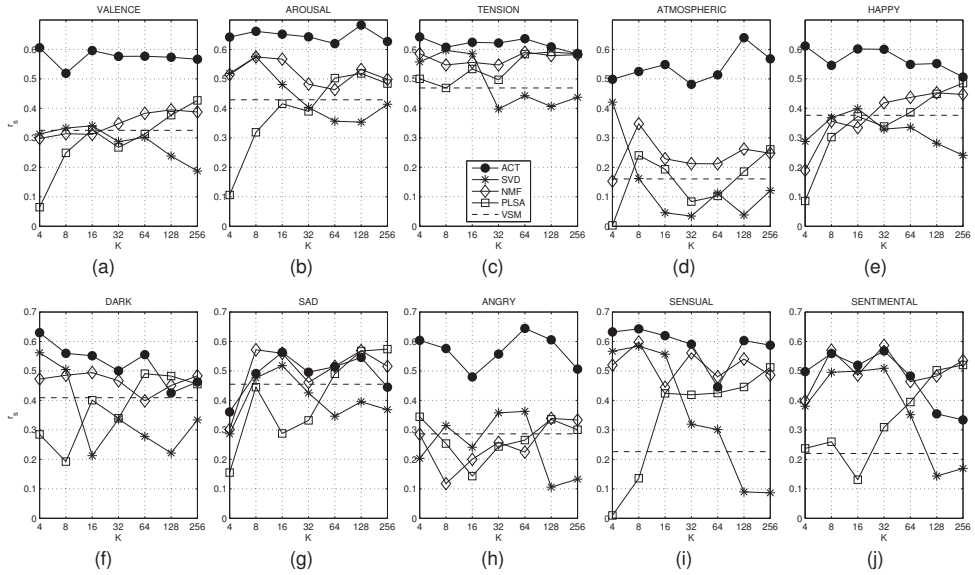


Fig. 7. Prediction rates (r_s) of listener ratings in (a – c) VAT scales and (d – j) mood term scales.

TABLE 3
Comparison of the performances of ACT, baseline techniques, and alternative ACT implementations (*Mdn* = median across k). Significances of the performance differences were computed by Wilcoxon rank sum test for equal medians between ACT and SVD, NMF and PLSA, and Wilcoxon signed rank test for median r_s between ACT and VSM, SVD only, and MDS only.

	ACT			BASELINE				ACT alt.	
				SVD	NMF	PLSA	VSM	SVD-only	MDS-only
	r_s *(Mdn)*	r_s *(min)*	r_s *(max)*	r_s *(Mdn)*	r_s *(Mdn)*	r_s *(Mdn)*	r_s	r_s	r_s
Valence	**.576**	.519	.606	$.302^{***}$	$.348^{***}$	$.313^{***}$	$.326^{*}$	$.475^{*}$	.558
Arousal	**.643**	.620	.683	$.414^{***}$	$.514^{***}$	$.416^{***}$	$.429^{*}$	$.373^{*}$	**.643**
Tension	**.622**	.585	.642	$.443^{**}$	$.579^{**}$	$.534^{**}$	$.469^{*}$	$.591^{*}$	$.596^{*}$
Atmospheric	**.525**	.482	.640	$.112^{***}$	$.229^{***}$	$.186^{***}$	$.161^{*}$	$.247^{*}$	**.581**
Happy	**.552**	.506	.612	$.330^{***}$	$.419^{***}$	$.373^{***}$	$.376^{*}$	$.279^{*}$	$.455^{*}$
Dark	**.552**	.425	.630	.334	.472	$.401^{*}$	$.409^{*}$	**.595***	$.239^{*}$
Sad	.496	.361	.563	.396	**.516**	.445	.455	$.328^{*}$	.469
Angry	**.576**	.480	.644	$.241^{***}$	$.258^{***}$	$.265^{***}$	$.286^{*}$	$-.131^{*}$	$.432^{*}$
Sensual	**.603**	.446	.643	$.319^{**}$	$.520^{*}$	$.424^{**}$	$.226^{*}$	.589	.542
Sentimental	**.498**	.334	.568	.380	.486	.309	$.220^{*}$	.420	.356

Note: $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$, $df = 6$.

method. NMF outperformed ACT only at predicting Sad, but this difference was not, however, statistically significant.

In general, median performances of ACT were lower for the individual mood scales than for VAT dimensions, ranging from $r_s = .496$ (Sad) to $r_s = .603$ (Sensual). Performance difference between ACT and baseline techniques was the most notable for Atmospheric and Angry. While ACT yielded median performances $r_s = .525$ for the former scale and $r_s = .576$ for the latter, the most successful baseline techniques (NMF and VSM, respectively) produced only $r_s = .229$ and $r_s = .286$.

ACT performance was generally more sensitive to the value of k for the individual mood terms than for the VAT dimensions. The performance range was smallest for Happy ($\Delta r_s = .105$, $.506 \leq r_s \leq .612$) and largest for Sentimental ($\Delta r_s = .234$, $.334 \leq r_s \leq .568$). However, the corresponding numbers were higher for all baseline techniques.

All in all, these results show that ACT is efficient at predicting the individual mood terms and gives consistent performance for mood terms (Atmospheric, Angry), which the baseline techniques fail at predict-

ing. Together with the findings for VAT dimensions, this suggests that domain knowledge on moods can be utilized to great benefit in semantic computing.

7.3 ACT with Alternative Implementations

While ACT clearly outperformed the baseline techniques at predicting the perceived mood, we carried out further comparative performance evaluation with ACT to assess the optimality of the technique. In particular, we were interested to find whether it is beneficial to implement ACT with dimension reduction in two stages, involving low-rank approximation with SVD and mood term configuration with MDS. For this evaluation we analyzed the performance of two models: *a)* SVD-only applying Procrustes directly on the SVD mood term configuration $u_i = U_i^k S^k$ ($k = 3$) without the MDS stage; and *b)* MDS-only applying MDS on the cosine distances between mood terms computed from the raw TF-IDF matrix instead of the low-rank representation. In must be noted, however, that the latter model effectively corresponds to the original ACT with $k = |terms|$ but is computationally heavier than the original ACT when the TF-IDF matrix is large.

The results presented in Table 3 show that both ACT implementations yielded performance mostly comparable to that of the original ACT. The original ACT generally outperformed both alternative implementations. This difference was statistically significant in seven moods for SVD-only and in four moods for MDS-only. SVD-only outperformed the original ACT for Dark, whereas MDS-only yielded the highest performance for Arousal and Atmospheric. However, the performance differences for MDS-only were not statistically significant. The clearest difference was between ACT and SVD-only for Angry, where SVD-only failed to produce positive correlation.

The results suggest that mood prediction performance of ACT is significantly boosted by utilizing both SVD and MDS.

7.4 The Effect of Tag Sparsity on ACT Performance

As noted in the introduction, social tag data is sparse, meaning that a typical document is associated to only a subset of all relevant tags. In the mood data fetched from Last.fm *493,539* tracks are associated to at least one mood term, whereas only 38,450 tracks are associated to at least 8 terms, which is approximately the average within the test set. If we consider only the level of data sparsity, we can assume that the performance presented above extends to approximately *38,000* tracks. The question is, how high could prediction performance be expected for the larger set of almost *500,000* tracks?

To study this question, we carried out systematic performance assessment with ACT as a factor of the number of mood terms associated to the test tracks. Starting with the full test set, we iteratively removed one term-track association at a time until all tracks were associated to only one term. The association to be removed was sampled in a weighted random manner at each iteration, weighting tracks directly proportional to the number of associated terms, and terms with lower raw within-track counts. We recorded ACT performance at each iteration, and calculated the mean performance across ten separate runs. The process can be seen as imitating user tagging, where a novel track in a system is most likely first applied with clearly descriptive tags. The results of the analysis are summarized in Table 4, showing the median performance across k obtained with $1, 2, 4, 6$ and 8 terms associated to each track in average.

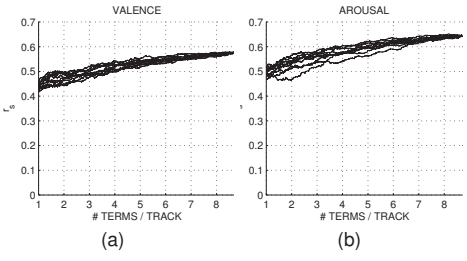


Fig. 8. The relationship between the number of tags for each track and the ACT performance ($k = 32$) for (a) Valence and (b) Arousal.

The results suggest that tag sparsity and prediction performance are in a strong linear positive relationship, supporting the assumption that tag sparsity primes the ACT prediction performance. This relationship held also at each of the ten separate runs (see Fig. 8). Depending on the mood, performance achieved with only one tag in each track was approximately $r_s = .433$ and varied between $r_s = .352$ (Sentimental) and $r_s = .496$ (Tension). Difference between the performances obtained with the full test set to that with only one term for each track was on average $\Delta r_s = .132$, $.086 \leq \Delta r_s \leq .151$, which is not a drastic drop considering that the prediction based on one term alone deals with a lot less track-level information.

These results suggest that ACT prediction is robust against the low number of track-level mood tags. Based on the results, we estimate that the correlations of $r_s = .433$ between the perceived mood and ACT mood predictions extend to the large set of almost *500,000* tracks extracted from Last.fm. This gives positive implications for utilizing sparse but abundant social tags to manage and retrieve music.

TABLE 4
Median performances (r_s) across k obtained with ACT when the specified numbers of mood terms in average were associated to each track in the test set. *# Tracks* refers to the number of the fetched Last.fm tracks with at least *# Terms*.

# Terms / Track	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**8.71 (Full)**
# Tracks	*493,539*	*259,593*	*164,095*	*114,582*	*84,206*	*64,018*	*49,393*	*38,450*	
Valence	.445	.474	.498	.521	.535	.548	.558	.568	.576
Arousal	.492	.530	.560	.578	.600	.615	.627	.639	.643
Tension	.496	.535	.559	.576	.590	.598	.607	.617	.622
Atmospheric	.375	.419	.445	.462	.477	.493	.509	.519	.525
Happy	.418	.454	.479	.497	.513	.525	.535	.543	.552
Dark	.413	.447	.471	.495	.512	.527	.539	.545	.552
Sad	.368	.387	.410	.429	.451	.471	.482	.491	.496
Angry	.490	.511	.525	.540	.546	.554	.562	.570	.576
Sensual	.475	.510	.535	.550	.567	.578	.586	.595	.603
Sentimental	.352	.382	.410	.428	.450	.463	.477	.489	.498

8 Conclusions

This paper marks the first systematic assessment of the potential of social tags at capturing mood information in music. We used large-scale analysis of social tags coupled with existing emotion models to construct robust music mood prediction.

We proposed a novel technique called Affective Circumplex Transformation to represent mood terms and tracks in a space of Valence, Arousal and Tension. Use of the dimensional emotion model to represent moods was supported by our analysis of the structure of the tag data. ACT outperformed the baseline techniques at predicting listener ratings of moods in a separate test set of tracks spanning multiple genres. Furthermore, the results showed that mood prediction with ACT is robust against the low number of track-level mood tags, and suggested that moderate to good fit with the dimensional emotion model can be achieved in extremely large data sets.

The present study facilitates information retrieval according to mood, assists in building large-scale mood data sets for music research, gives new ways to assess the validity of emotion models on large data relevant to current music listening habits, and makes large data available for training models that automatically annotate music based on audio.

A limitation of the present study is the possible discrepancy between track-level and clip-level moods, which may have reduced the prediction rates presented. This is because tags associated to full tracks may not adequately describe the representative clip rated in the listening experiment. Moreover, further research is needed to assess the mood-related associations of music genres and genre-related implications to mood modeling.

The implications of the present study also extend to mood mining in other online content, as it was shown that domain knowledge of moods is highly beneficial to semantic computing. Moreover, the techniques developed here can be applied to social tags as well as to other types of textual data.

Acknowledgments

The authors wish to thank the reviewers for their invaluable feedback and suggestions.

References

[1] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 9, pp. 1168 –1180, sept. 2008.

[2] T. Nguyen, D. Phung, B. Adams, and S. Venkatesh, "Mood sensing from social media texts and its applications," *Knowledge and Information Systems*, pp. 1–36, 2013.

[3] S. Schmidt and W. G. Stock, "Collective indexing of emotions in images. a study in emotional information retrieval," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 863–876, 2009.

[4] S. Bao, S. Xu, L. Zhang, R. Yan, Z. Su, D. Han, and Y. Yu, "Mining social emotions from affective text," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 9, pp. 1658 –1670, sept. 2012.

[5] P. N. Juslin and J. A. Sloboda, *Handbook of music and emotion: Theory, research, applications*. Oxford University Press, 2009.

[6] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, April 2006.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[8] P. Heymann, D. Ramage, and H. Garcia-Molina, "Social tag prediction," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 531–538.

[9] Y. Song, L. Zhang, and C. L. Giles, "Automatic tag recommendation algorithms for social recommender systems," *ACM Transactions on the Web (TWEB)*, vol. 5, no. 1, p. 4, 2011.

[10] A. Garcia-Silva, O. Corcho, H. Alani, and A. Gomez-Perez, "Review of the state of the art: Discovering and associating semantics to tags in folksonomies," *The Knowledge Engineering Review*, vol. 27, no. 01, pp. 57–85, 2012.

[11] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?" in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 193–202.

[12] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[13] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[14] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags," in *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 381–86.

[15] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[16] M. Zentner, D. Grandjean, and K. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.
[17] P. Ekman, "An argument for basic emotions," *Cognition & Emotion*, vol. 6, pp. 169–200, 1992.
[18] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models and stimuli," *Music Perception*, vol. 30, no. 3, pp. 307–340, 2012.
[19] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, Nov. 1975.
[20] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.
[21] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1-2, pp. 177–196, 2001.
[22] E. Law, B. Settles, and T. Mitchell, "Learning to tag from open vocabulary labels," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, J. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds. Springer Berlin Heidelberg, 2010, vol. 6322, pp. 211–226.
[23] N. Zhou, W. Cheung, G. Qiu, and X. Xue, "A hybrid probabilistic model for unified collaborative and content-based image tagging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 7, pp. 1281–1294, 2011.
[24] J. Peng, D. D. Zeng, and Z. Huang, "Latent subject-centered modeling of collaborative tagging: An application in social search," *ACM Transactions on Management Information Systems*, vol. 2, no. 3, pp. 15:1–15:23, Oct. 2008.
[25] R. Wetzker, W. Umbrath, and A. Said, "A hybrid approach to item recommendation in folksonomies," in *Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval*. ACM, 2009, pp. 25–29.
[26] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.
[27] R. Peter, G. Shivapratap, G. Divya, and K. Soman, "Evaluation of svd and nmf methods for latent semantic analysis," *International Journal of Recent Trends in Engineering*, vol. 1, no. 3, pp. 308–310, 2009.
[28] R. A. Calvo and S. Mac Kim, "Emotions in text: Dimensional and categorical models," *Computational Intelligence*, 2012.
[29] K. R. Scherer, *Emotion as a multicomponent process: A model and some cross-cultural data*. Beverly Hills: CA: Sage, 1984, pp. 37–63.
[30] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, USA, 1989.
[31] P. Juslin and P. Laukka, "Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening," *Journal of New Music Research*, vol. 33, pp. 217–238, 2004.
[32] P. Juslin, S. Liljeström, P. Laukka, D. Västfjäll, and L. Lundqvist, "Emotional reactions to music in a nationally representative sample of swedish adults prevalence and causal influences," *Musicae scientiae*, vol. 15, no. 2, pp. 174–207, 2011.
[33] M. R. Zentner and T. Eerola, *Handbook of Music and Emotion*. Boston, MA: Oxford University Press, 2010, ch. Self-report measures and models, pp. 187–221.
[34] M. M. Bradley and P. J. Lang, "Affective norms for english words (anew): Instruction manual and affective ratings," Technical Report C-1, The Center for Research in Psychophysiology, University of Florida, Tech. Rep., 1999.
[35] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 6, pp. 1802 –1812, aug. 2011.
[36] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.
[37] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion knowledge: further exploration of a prototype approach." *Journal of Personality and Social Psychology*, vol. 52, no. 6, pp. 1061–86, 1987.
[38] K. Hevner, "Experimental studies of the elements of expression in music," *The American Journal of Psychology*, vol. 48, no. 2, pp. 246–268, 1936.
[39] X. Hu and J. S. Downie, "Exploring mood metadata: relationships with genre, artist and usage metadata," in *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.
[40] C. Fellbaum, Ed., *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.
[41] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.
[42] C. Bécavin, N. Tchitchek, C. Mintsa-Eya, A. Lesne, and A. Benecke, "Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition," *Bioinformatics*, vol. 27, no. 10, pp. 1413–1421, May 2011.
[43] B. Hopkins and J. G. Skellam, "A new method for determining the type of distribution of plant individuals," *Annals of Botany*, vol. 18, no. 2, pp. 231–227, 1954.
[44] R. G. Lawson and P. C. Jurs, "New index for clustering tendency and its application to chemical problems," *Journal of Chemical Information and Computer Sciences*, vol. 30, no. 1, pp. 36–41, 1990.
[45] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*. Oxford University Press Oxford, 2004, vol. 3.
[46] M. I. Mandel, R. Pascanu, D. Eck, Y. Bengio, L. M. Aiello, R. Schifanella, and F. Menczer, "Contextual tag inference," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 7S, no. 1, pp. 32:1–32:18, October 2011.
[47] X. Hu, J. S. Downie, C. Laurier, M. Bay, and A. F. Ehmann, "The 2007 mirex audio mood classification task: Lessons learned," in *Proceedings of 9th International Conference on Music Information Retrieval (ISMIR)*, 2008, pp. 462–467.
[48] J. Nunnally, *Psychometric theory*. New York: McGraw-Hill, 1978.
[49] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music." *Psychol. Music*, vol. 39, no. 1, pp. 18–49, 2011.
[50] T. Eerola, "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*, vol. 40, no. 4, pp. 349–366, 2011.

Pasi Saari received MSc degree in Computer Science in 2008 and MA degree in Musicology in 2010 from the University of Jyväskylä, Finland.

He is currently working as a Doctoral Student at the Finnish Centre of Excellence in Interdisciplinary Music Research within the University of Jyväskylä, Finland. His research interests are in semantic computing of moods in music and content-based analysis of musical audio.

Tuomas Eerola received a PhD in musicology from the University of Jyväskylä, Finland and is currently a full Professor of Music at this institution. He is also affiliated with the Finnish Centre of Excellence in Interdisciplinary Music Research.

His research interest are in music cognition, particularly the perception of melody, rhythm, timbre, and induction of emotions by music.

III

SEMANTIC MODELS OF MOOD EXPRESSED BY MUSIC: COMPARISON BETWEEN CROWD-SOURCED AND CURATED EDITORIAL ANNOTATIONS

by

Pasi Saari, Mathieu Barthet, György Fazekas, Tuomas Eerola & Mark Sandler
2013

In IEEE International Conference on Multimedia and Expo Workshops (ICMEW)

SEMANTIC MODELS OF MUSICAL MOOD: COMPARISON BETWEEN CROWD-SOURCED AND CURATED EDITORIAL TAGS

Pasi Saari[*], *Mathieu Barthet*[†], *György Fazekas*[†], *Tuomas Eerola*[*], *and Mark Sandler*[†]

[*]Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland
{pasi.saari, tuomas.eerola}@jyu.fi
[†]Centre for Digital Music, Queen Mary University of London, United Kingdom
{mathieu.barthet, gyorgy.fazekas, mark.sandler}@eecs.qmul.ac.uk

ABSTRACT

Social media services such as Last.fm provide crowd-sourced mood tags which are a rich but often noisy source of information. In contrast, editorial annotations from production music libraries are meant to be incisive in nature. We compare the efficiency of these two data sources in capturing semantic information on mood expressed by music. First, a semantic computing technique devised for mood-related tags in large datasets is applied to Last.fm and I Like Music (ILM) corpora separately (250,000 tracks each). The resulting semantic estimates are then correlated with listener ratings of arousal, valence and tension. High correlations (Spearman's rho) are found between the track positions in the dimensional mood spaces and listener ratings using both data sources ($0.60 < r_s < 0.70$). In addition, the use of curated editorial data provides a statistically significant improvement compared to crowd-sourced data for predicting moods perceived in music.

Index Terms— Semantic computing, dimensional emotion model, affective circumplex transformation, music moods.

1. INTRODUCTION

Empirical evidence have shown that music has the ability to express emotion or mood (perceived emotions) and to evoke emotion in listeners (felt emotions) [1]. This is reflected in the prevalence of mood-related tags, i.e. free-form labels applied to tracks, albums, artists, etc. in popular online music tagging services such as Last.fm[1], and in the importance of editorial mood-related metadata in production music catalogues.

The collection and management of multimedia document tags is a widely-used practice in online services and content providers with large population of users. Typically, very large corpora of data describing semantic information on multimedia documents can be obtained straightforwardly from the end-users. Music related tags may contain information of any kind including genre, locale, mood, opinion and instrumentation. The importance of mood tags was highlighted in several studies including [2], claiming that mood tags account for 5% of the most commonly used tags, and [3], which reported that 15% of the song queries on Last.fm are made using mood tags. Mood-related metadata are also considered important for searching and finding suitable tracks from production music catalogues, especially for specific purposes in creative media production involving music, such as movie making. In order to build models and applications[2] to classify tracks according to moods, there is a need to develop robust semantic representations of mood tags, in line with judgements from human listeners.

We wish to compare the reliability of semantic tags and mood representations based on tags obtained from two different sources of data: (i) crowd-sourced tags available from Last.fm, and (ii) curated editorial annotations used in production music catalogues. Moreover, this study seeks to assess how wide is the gap between the semantic representations of mood from these two data sources by applying semantic models across the sources. We assess the reliability of semantic representations using listener ratings collected for each source. For production music we use a unique source of curated editorial tags extracted from I Like Music's[3] (ILM) collection, aggregated from 29 individual production music catalogues.

In order to represent semantically meaningful information in a low-rank space, tag data can be analysed using Latent Semantic Analysis (LSA) [4]. The technique reduces noise resulting from spelling variations, the frequent use of synonyms, the polysemy of words, and largely subjective annotations occurring in crowd-sourced tag data. This is achieved by

This work was supported by the Academy of Finland (project numbers 7118616 and 125710) and partly funded by the TSB project 12033-76187 "Making Musical Mood Metadata" (TS/J002283/1).

[1] http://www.last.fm

[2] See e.g. Gracenote Habu, https://www.gracenote.com/case_studies/habu, or Spotify Moodagent apps http://www.moodagent.com/spotify.

[3] http://www.ilikemusic.com/

learning the latent structure of the semantic space in an unsupervised manner, in other words it learns context-specific relationships between tags from domain-specific data. The process of LSA involves Singular Value Decomposition (SVD) to find a low-rank approximation of a term-document matrix, leading to the above-mentioned semantic space with reduced dimensionality and data sparsity.

Past research in Music Information Retrieval (MIR) has successfully established relationships between the semantic spaces of crowd-sourced tags based on LSA and expert-based taxonomies for moods [5] and genres [6]. Moreover, [7] recently examined the reliability of mood-related tag data obtained from Last.fm by comparing semantics emerging from track-level tags to listener ratings on the corresponding mood scales. The authors proposed the Affective Circumplex Transformation (ACT) method based on LSA and mood models in the field of affective sciences to explicitly estimate the mood expressed by music tracks. The results showed medium high correlations ($r_s \approx 0.60$) between the estimates and ratings for *valence*, *arousal*, and *tension*. No similar study has been conducted yet using music mood tags curated by music librarians and professional music experts.

The remainder of this article is organised as follows: In Section 2, we present the system devised to uncover semantic music mood models from metadata. In Section 3, we describe the cross-evaluation framework used to assess the semantic music mood models obtained using the Last.fm and ILM datasets. The results of the evaluation process are presented and discussed in Section 4. Section 5 summarises the findings and proposes future developments of this work.

2. SEMANTIC ANALYSIS

The following procedures are applied separately for Last.fm tags and ILM annotations. A detailed description of Last.fm data and analysis is given in [7].

2.1. Vector Space Modelling

First, mood and genre vocabularies were collected by aggregating and lemmatising words listed in several research papers in affective sciences, music psychology and MIR, as well as in the Allmusic.com web service. The genre vocabulary was used to select tracks for the listening test detailed in Section 3.2 to ensure a well-balanced representation of genres. Synonyms and inflected forms of the vocabulary terms were identified and aggregated, or added manually, such as (happy, happiness) and (rhythm and blues, r'n'b, R&B). The resulting vocabularies consist of 560 unique mood words and 865 distinct genre names.

Last.fm [7] and ILM mood and genre vocabulary terms were identified from tags using a bag-of-words approach similar to that used in [8]. Vocabulary terms were then applied to associated tracks accordingly. To avoid obtaining overly sparse and uncertain information, we excluded tracks with less than two mood (and genre) terms, and terms applied to less than 100 tracks. Finally, both datasets were normalised by computing term frequency-inverse document frequency (TF-IDF) weights: $\hat{n}_{w,t} = (n_{w,t} + 1)\log(\frac{R}{f_w})$, where $n_{w,t}$ is the original frequency weight related to term w and track t, R is the total number of tracks, and f is the number of tracks the term w was applied to. Statistics describing the mood data associated with the Last.fm and ILM datasets are summarised in Table 1.

Table 1. Statistics of the mood term sets.

	# Tracks	# Terms	# Applied terms / track
LAST.FM	259,593	357	4.44
ILM	226,344	288	4.81

2.2. Singular Value Decomposition

Low-rank approximations of the resulting mood (and genre) TF-IDF matrices was then computed by Singular Value Decomposition (SVD). SVD decomposes a sparse matrix N into orthogonal matrices U and V, and diagonal matrix S, such that $N = USV^T$. S contains singular values in decreasing order. A rank k approximation of N is then computed by $\tilde{N}_k = U^k S^k (V^k)^T$, where each row vector U_i^k represents the term w_i with k relative weights for each dimension. Similarly, V_j^k represents track t_j as k relative weights. Based on the rank k approximation, dissimilarity between terms w_i and $w_{\hat{i}}$ can be computed using the cosine distance between the $U_i^k S^k$ and $U_{\hat{i}}^k S^k$ vectors.

2.3. Affective Circumplex Transformation

We use the Affective Circumplex Transformation (ACT) proposed in [7] to infer explicit representation of *valence*, *arousal*, and *tension* for the annotated tracks. The rationale behind ACT is based on research in psychology [9, 10, 11] which showed that the variance between various mood states could be modelled using only a few underlying affective dimensions.

To represent mood terms in a low-dimensional space, non-metric Multidimensional Scaling (MDS) [12] is applied on the term dissimilarities obtained by rank k approximation of mood TF-IDF obtained by SVD. Three-dimensional mood spaces were obtained, yielding similar stress values (Kruskal's stress 1, denoted ϕ_k) for the Last.fm ($\phi_4 = 0.02, \phi_{256} = 0.29$) and ILM ($\phi_4 = 0.02, \phi_{256} = 0.25$) datasets.

Next, the resulting MDS mood term space is made to fit the space of *arousal* and *valence* (AV), using AV values of 101 mood terms given in [9, p. 1167] and [10, p. 54]. To ob-

tain explicit representation of tension, the model by [13] can be projected on the space diagonally against negative valence and positive arousal. First, mood term co-occurrences are found between the MDS and AV-spaces, yielding 47 and 37 matches for Last.fm and ILM, respectively. Then, the MDS term space is transformed to match the AV space using classical Procrustes analysis [14] with sum of squared errors used as goodness-of-fit. The method retains the relative distances between objects in the original MDS configuration, since it allows only translation, reflection, orthogonal rotation, and isotropic scaling. Given AV values $x_{\hat{i}} = (x_{\hat{i}1}, x_{\hat{i}2})$, and MDS configuration $y_{\hat{i}} = (y_{\hat{i}1}, y_{\hat{i}2}, y_{\hat{i}3})$, where $\hat{i}$ denotes the mood terms matched between MDS and AV, the Procrustes transformation gives $\hat{x}_{\hat{i}} = By_{\hat{i}}T + C$, where B is an isotropic scaling component, T is an orthogonal rotation and reflection component, and C is a translation component. B, T, and C minimise the goodness-of-fit measure X^2: $X^2 = \Sigma_{\hat{i}}(x_{\hat{i}} - \hat{x}_{\hat{i}})^2$. To this end, AV values $\hat{x}_{\hat{i}}$ are zero-padded into three dimensions. Configuration $\hat{x}_i$ composed of all mood terms is then obtained by using the transformation $\hat{x}_i = Bx_iT + C$.

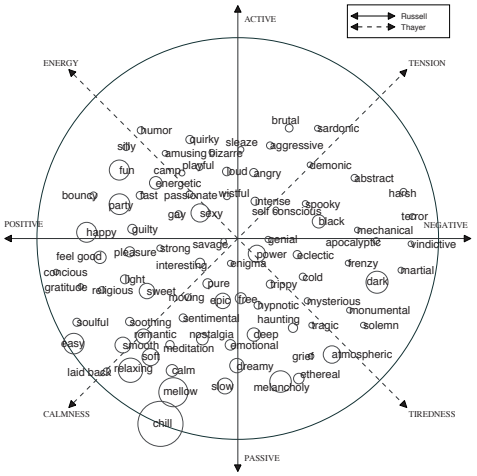
Fig. 1 shows AV values of the resulting mood term configurations ($k = 16$) for both Last.fm tags and ILM terms. The frequencies of the terms shown span from 110 tracks ("vindictive") to 79,524 tracks ("chill") for Last.fm, and 346 tracks ("narrative") to 39,892 tracks ("uplifting") for ILM. It can be seen that ILM terms have more positive valence in general, which may reflect the different music genres covered by these corpora. Moreover, the Last-fm configuration shows certain unexpected term positions. For example, positive valence of "guilty" may be explained by a frequent term combination "guilty pleasure", which yields low distance between these terms.

Tracks are projected onto the resulting mood space based on the term positions and sparse TF-IDF term vectors of tracks. Given a configuration of terms $\hat{x}_i$, $i \in (1, ..., n)$, where n is the number of terms, track positions are computed by taking the euclidean mean of the term positions, weighted by the sparse TF-IDF vector q of the track: $\hat{t} = (\Sigma_i q_i \hat{x}_i)/(\Sigma_i q_i)$. This way, any track associated with one or more mood terms can be projected.

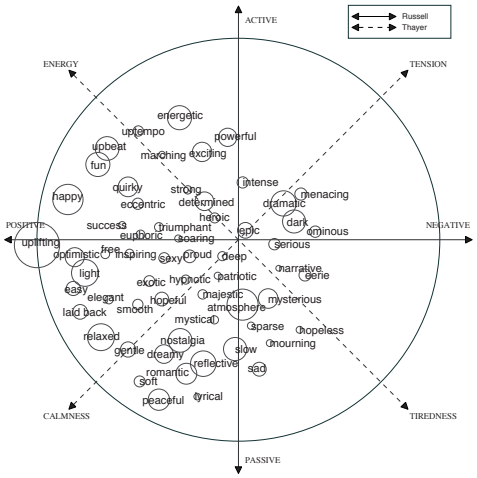
3. EVALUATION FRAMEWORK

3.1. Cross-evaluation Protocol

The system is evaluated using four methods outlined in Fig. 2. We use ACT_L and ACT_I hereafter to denote the semantic models obtained by applying ACT to Last.fm and ILM tags, respectively. Our four methods can be summarised as follows: (1) Using ACT_L for predicting mood in the Last.fm test set as in [7]; (2) Using ACT_I for predicting moods in the Last.fm test set; (3) Using ACT_L for predicting moods in the ILM production music set; and (4) Using ACT_I for predicting mood in the ILM production music test set.



(a) Last.fm.



(b) ILM.

Fig. 1. ACT with rank $k = 16$ based on Last.fm tags (a) and ILM production music tags (b). Only the most frequently applied tags are shown for each part of the AV-space. Tag frequencies are reflected in the bubble sizes.

Mood ratings obtained from two listening tests, one using 600 tracks from Last.fm (see [7] for details), and one using a set of ILM tracks (see Section 3.2) were used as ground-

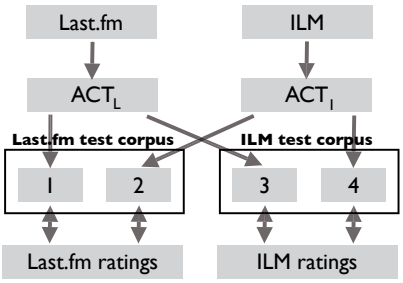


Fig. 2. Cross-evaluation framework for semantic musical mood models using two different sources of metadata and tracks.

truth data to evaluate the accuracy with which the system described in Section 2 predicts perceived mood in music. We apply ACT on the tag data, project the tracks used in the listening test onto the AV-space, and compute non-parametric Spearman's rho coefficients of correlation between the track positions and the ratings. Track positions along the valence and arousal axes are directly correlated with the ratings for the corresponding scales, and the model estimates for tension are obtained by projecting tags along the direction proposed by Thayer [13].

The evaluation of the ACT performance across annotation types (methods 2 and 3 in Fig. 2) is achieved following three principles. First, corresponding mood terms in the two sets are identified and only matching terms are used in further computation. A total of 251 co-occurring terms were found in both sets, which reduces the Last.fm data (357 terms) more than the ILM data (288 terms). Second, *TF-IDF weighting* of the rated tracks in one test set is computed based on the tag frequencies in the other test set. Third, mood term positions based on the ACT of the other set are used to project the tracks.

3.2. Listening Test with ILM Tracks

3.2.1. Track Corpus

A corpus of 205 production music tracks was sampled from the ILM dataset. The sampling was made in a semi-random fashion based on several criteria to ensure that the resulting set well covers the MDS mood space, as well as the main genres prevalent in the analysed production music data. To this end, *k*-means clustering was applied to group tracks according to genres based on an MDS genre representation, and tracks were then sampled from these clusters in a stratified manner. Our analyses suggested that *six* distinct genre clusters were enough to represent a large part of the ILM dataset. We present hereafter the three most prevalent genre tags within each cluster, the most prevalent in italic, and the other two within brackets: *jazz* (swing, lounge), *dance* (pop, house), *rock* (pop, alternative), *electronic* (urban, ambient), *folk* (country, pop), and *orchestral* (classical, choral).

3.2.2. Procedure

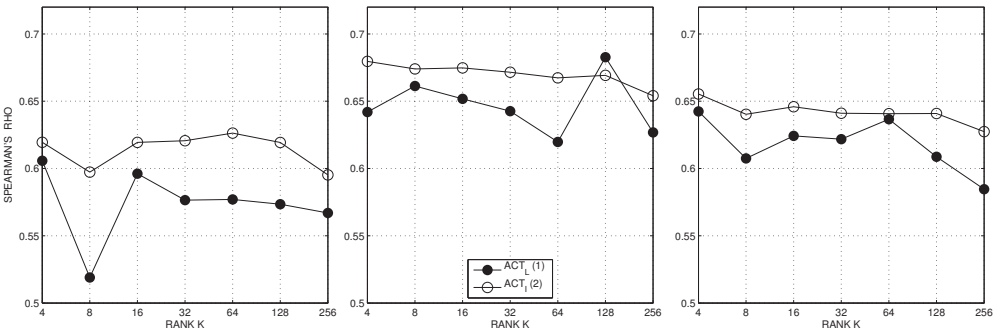
A procedure similar to that proposed in [7] (Last.fm corpus) was followed with the ILM corpus. An online annotation interface was used to ask participants to annotate 30 second audio excerpts of tracks from the corpus in terms of the perceived moods in music. Annotations for six mood scales were done using nine-step bipolar Likert scales: calm/energetic (*arousal*), negative/positive (*valence*), relaxed/tense (*tension*), submissive/dominant (*dominance*), cold/romantic (*romance*), and serious/funny (*humour*).

4. RESULTS AND DISCUSSION

4.1. Listeners' Ratings

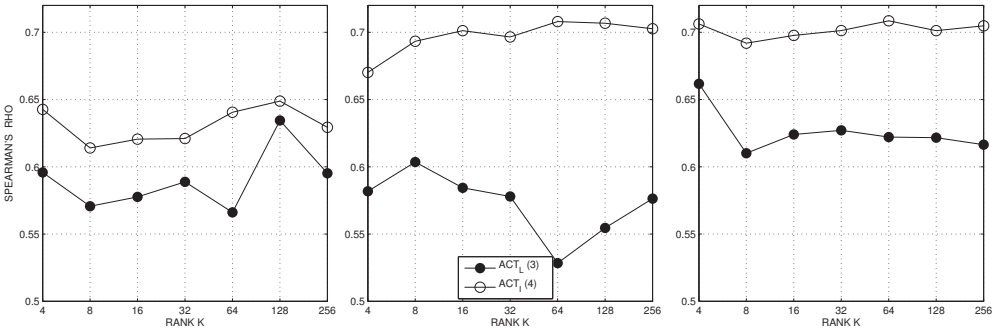
A total of 46 participants (mean age 32.3 years, SD = 9.0 years, 30 males) from 20 countries (mostly Europeans, 13 participants from the United Kingdom) took part in the experiment. Musical expertise of the participants spanned from listeners (N=14) to musicians (N=20), and trained professionals (N=20). For the sake of rating consistency between participants, we selected participants who had rated more than 20% of the tracks for further analyses. This resulted in 8.9 ratings per track on average (SD = 0.90 ratings). Cronbach's α, a widely used measure representing the inter-subject agreement, was computed for each mood scale to assess the reliability of the obtained data. This yielded acceptable values ($\alpha \geq 0.70$ [15]) for *valence*, *arousal*, and *tension*, and slightly lower values for the other scales ($\alpha > 0.64$). In the remainder of this article, we focus on the *valence*, *arousal*, and *tension* mood dimensions and characterise the mood expressed by each track with the mean values computed across participants.

Based on the ratings, no correlation between *arousal* and *valence* was found ($r = 0.06, p = 0.41$), which supports the two dimensional model proposed by Russell [9]. *Tension* is positively correlated with *arousal* ($r = 0.57$) and negatively correlated with *valence* ($r = -0.67$). In fact, almost all variance in *tension* ($R^2 = 0.81$) can be explained by a linear combination of *arousal* and *valence*, which in turn supports Thayer's [13] projection of tension diagonally against positive arousal and negative valence. These correlations were in line with those found with the Last.fm ratings. The ratings of *tension* showed high positive correlation with *dominance* ($r = 0.85$) and high negative correlation with *romance* ($r = -0.85$), whereas *valence* showed high correlation with *humour* ($r = 0.81$).



(a) Valence (b) Arousal (c) Tension

Fig. 3. Correlations between the semantic estimates and listener ratings for the Last.fm test set.



(a) Valence (b) Arousal (c) Tension

Fig. 4. Correlations between the semantic estimates and listener ratings for the ILM test set.

4.2. Fit between Mood Model Projections and Ratings

The system evaluation was performed using different values of the rank parameter k for the LSA technique employed prior to the ACT. The results with Last.fm test set (methods 1 and 2) are shown in Fig. 3, demonstrating that the correlations obtained with ACT_L ($0.52 < r_s < 0.61$ for valence, $0.62 < r_s < 0.68$ for arousal, and $0.58 < r_s < 0.64$ for tension) are generally lower than the correlations obtained with ACT_I ($0.60 < r_s < 0.63$ for valence, $0.65 < r_s < 0.68$ for arousal, and $0.63 < r_s < 0.66$ for tension). The only exception is the correlation obtained with the arousal dimension for $k = 128$.

A paired sample Student's t-test was applied to evaluate the difference in correlations obtained with ACT_L and ACT_I across k. The test revealed a highly significant difference between ACT_L and ACT_I for *valence* ($t(6) = -5.03, p = 0.00237$) and *tension* ($t(6) = -4.75, p = 0.00315$), and a significant difference ($t(6) = -3.15, p = 0.0197$) for *arousal*, all in favour of ACT_I. These results suggest that the semantic model derived from curated editorial mood annotations of production music is better in predicting moods than the semantic model derived from crowd-sourced data.

The results with the ILM test set (methods 3 and 4) are shown in Fig. 4. Applying ACT_I gives the highest performance of all four methods ($0.61 < r_s < 0.65$ for valence, $0.67 < r_s < 0.71$ for arousal, and $0.69 < r_s < 0.71$ for tension). Moreover, ACT_I again outperforms applying ACT_L ($0.57 < r_s < 0.63$ for valence, $0.53 < r < 0.60$ for arousal, and $0.61 < r_s < 0.66$ for tension). The difference between ACT_L and ACT_I and is highly significant (valence: $t(6) = -5.98, p = 0.00098$; arousal: $t(6) = -10.08, p = 0.00006$; tension: $t(6) = -13.53, p = 0.00001$) for all mood scales using the ILM test set.

Applying ACT_L on the ILM test set rather than the

Last.fm test set doesn't significantly affect the performance, except for the arousal dimension, for which the drop in the correlation coefficient (from $r_s \approx 0.65$ to $r_s \approx 0.57$) is highly significant ($t(6) = -7.28, p = 0.00034$). This shows that the semantic models derived from crowd-sourced annotations of commercial music can be used in a reliable manner to predict the moods expressed by production music tracks. In general, the results show that semantic models of moods based on ACT provide fairly robust generalizability across annotation types and music corpora.

5. CONCLUSIONS

In this study, we assess whether semantic mood models derived from the Last.fm and I Like Music (ILM) datasets can be used to predict mood expressed by music tracks (i) from the same corpora, and (ii) from different corpora. In summary, the results indicate the following conclusions:

- Data-driven semantic mood models are efficient to predict perceived mood in both data sets (Last.fm and ILM).
- The use of ILM editorial tags provide a statistically significant improvement compared to crowd-sourced data for the semantic modelling of mood expressed by music.
- Semantic model of moods can be built based on one corpus and efficiently applied to another, regardless of the difference in music styles and mood annotations.
- We claim that the overall quality of annotations is the most important factor determining the performance of the obtained models.

The results show promising ways to capitalise on large datasets of annotated music corpora to improve our understanding of how mood-related semantics can be reliably extracted from both crowd-sourced tags and editorial annotations. Future work includes adding other relevant semantic content (such as genre) and incorporating audio descriptors to tackle the challenge of predicting perceived mood in music in a robust fashion.

6. REFERENCES

[1] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models," in *Proc. of the 9th Int. Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012, pp. 492–507.

[2] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[3] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?," in *Proc. of the ACM Conference on Information and Knowledge Management (CIKM)*, 2008, pp. 193–202.

[4] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[5] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags," in *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 381–86.

[6] M. Sordo, Ò. Celma, M. Blech, and E. Guaus, "The quest for musical genres: Do the experts and the wisdom of crowds agree?," in *Proceedings of 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.

[7] P. Saari and T. Eerola, "Semantic computing of moods based on tags in social media of music," *IEEE Transactions on Knowledge and Data Engineering*, manuscript submitted for publication available at http://arxiv.org/, 2013.

[8] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.

[9] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[10] K. R. Scherer, *Emotion as a multicomponent process: A model and some cross-cultural data*, pp. 37–63, CA: Sage, Beverly Hills, 1984.

[11] P. N. Juslin and J. A. Sloboda, *Handbook of Music and Emotion*, chapter Introduction: aims, organization, and terminology, pp. 3–14, Oxford University Press, Boston, MA, 2010.

[12] J.B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.

[13] R. E. Thayer, *The Biopsychology of Mood and Arousal.*, Oxford University Press, New York, USA, 1989.

[14] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*, vol. 3, Oxford University Press Oxford, 2004.

[15] J. Nunnally, *Psychometric theory*, New York: McGraw-Hill, 1978.

IV

USING SEMANTIC LAYER PROJECTION FOR ENHANCING MUSIC MOOD PREDICTION WITH AUDIO FEATURES

by

Pasi Saari, Tuomas Eerola, György Fazekas & Mark Sandler 2013

In Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013), 722-728

USING SEMANTIC LAYER PROJECTION FOR ENHANCING MUSIC MOOD PREDICTION WITH AUDIO FEATURES

Pasi Saari and Tuomas Eerola
Finnish Centre of Excellence in Interdisciplinary Music Research
University of Jyväskylä, Finland
firstname.lastname@jyu.fi

György Fazekas and Mark Sandler
Centre for Digital Music
Queen Mary University of London
firstname.lastname@eecs.qmul.ac.uk

ABSTRACT

We propose a novel technique called Semantic Layer Projection (SLP) for predicting moods expressed by music based on audio features. In SLP, the predictive models are formed by a two-stage mapping from audio features to listener ratings of mood via a semantic mood layer. SLP differs from conventional techniques that produce a direct mapping from audio features to mood ratings. In this work, large social tag data from the Last.fm music service was analysed to produce a semantic layer that represents mood-related information in a low number of dimensions. The method is compared to baseline techniques at predicting the expressed Valence and Arousal in 600 popular music tracks. SLP clearly outperformed the baseline techniques at predicting Valence ($R^2 = 0.334$ vs. 0.245), and produced roughly equivalent performance in predicting Arousal ($R^2 = 0.782$ vs. 0.770). The difficulty of modelling Valence was highlighted by generally lower performance compared to Arousal. The improved prediction of Valence, and the increasingly abundant sources of social tags related to digital music make SLP a highly promising technique for future developments in modelling mood in music.

1. INTRODUCTION

The modern age of digital music consumption has brought new challenges in organising and searching rapidly expanding music collections. The popular appeal of music is often attributed to its striking ability to elicit or convey emotion. Therefore, managing large music collections in terms of mood has significant advantages that complement conventional genre-based organisation.

Social music services such as Last.fm [1] play an important role in connecting digital music to crowd-sourced semantic information. A prime advantage of using Last.fm data is in the large number of users worldwide applying semantic tags, i.e., free-form labels, to elements of the music domain, e.g. tracks, artists and albums. Tags are used in order to communicate users' music listening preferences that are also used for improving the service. The data is available to researchers through a dedicated API, which makes it possible to apply semantic computing to tags related to millions of tracks. Semantic computation of Last.fm tags has been found effective in characterising music information related to genre, mood, and instrumentation [1]. Parallel to analysing crowd-sourced tags, a tag set dedicated to music research purposes has also been collected in [2]. The importance of mood tags has been highlighted in several studies, including [3], claiming that mood tags account for 5% of the most commonly used tags. Applying semantic computation to tags can therefore yield effective mood-related semantic models for music.

The prominence of mood in music is reflected by the large number of studies modelling expressed or induced emotion. To this end, two prevalent techniques emerged: *i)* the dimensional model of Valence, Arousal and Tension; and *ii)* the categorical model of basic emotions such as happiness, sadness and tenderness. On one hand, these models have been found mutually inclusive to a large degree [4]. On the other hand, more general models of emotion have also been proposed, and refined using a taxonomy specifically designed for musically induced emotion [5].

These types of representations have been widely used in computational systems for predicting mood from audio. Feature extraction methods have been developed, for instance, in [6] and [7], providing a good basis for modelling and predicting perceived moods, genres and other characteristics of musical audio. The typical approach in most previous studies involves the use of computational algorithms, such as supervised machine learning, to predict perceived moods directly from audio features. For a more detailed overview of the advances of mood modelling and recognition, see e.g. [8].

Achieving high efficiency of these models, however, relies heavily on good quality ground-truth data. Due to the expense of human annotation, ground-truth is laborious to collect, and therefore typical data sets are limited to a few hundred tracks. This leads to challenges in mood prediction emerging from the high dimensionality of audio feature data and from the need for complex model parameter optimisation, often resulting in the lack of generalizability of the predictions to novel tracks [9]. One way of overcoming these challenges and increasing the efficiency of mood prediction is to utilise audio content related to a large number of tracks and associated crowd-sourced semantic tags.

In this work, we use multivariate techniques in a novel way to predict listener ratings of mood in 600 popular mu-

[1] Last.fm: http://www.last.fm/

sic tracks, using an intermediate semantic layer created from tag data related to a substantially large collection of tracks. This demonstrates how a large collection of tracks and associated mood tags can be used to improve prediction quality. The new technique involves mapping audio features (audio level) to a semantic mood space (semantic layer) first, and then mapping the semantic mood space to listener ratings (perceptual level). This differs from conventional methods that map audio directly to the perceptual level. Instead, we use direct mapping as baseline to assess the efficiency of the proposed technique.

2. RELATED WORK

This section summarises past research on connecting audio, as well as semantic and perceptual levels to represent music. Figure 1 illustrates how previous studies relate to the approach presented here.

2.1 Mapping from Audio Features to Semantic Layer

The challenge of auto-tagging music tracks can be considered analogous to our task. Gaussian Mixture Modelling (GMM) was used in [10], whereas [11] employed Support Vector Machines (SVM) for this purpose. Bertin-Mahieux et al. [12] proposed a boosting-based technique. This provided higher precision (0.312) and overall F-score (0.205) with somewhat lower recall (0.153) compared to hierarchical GMMs proposed in [10], when a set of general tag words were considered. In the context of mood tags, the authors reported 0.449, 0.176, 0.253 precision, recall and F-score, respectively, noting that, due to the specific experimental conditions, the results are bounded at a value lower than one. Miotto and Lanckriet [13] found that using semantic modelling of music tags improves auto-tagging compared to the conventional approach of treating each tag individually without any tag similarity information. The proposed Dirichlet mixture model (DMM) captured the broader context of tags and provided an improved peak precision (0.475) and F-score (0.285) compared to previous results using the same data set, when combining DMM with different machine learning techniques.

2.2 Mapping from Audio Features to Perceived Mood

Yang et al. [14] modelled moods represented in the Arousal-Valence (AV) plane using Support Vector Regression (SVR) with LIBSVM implementation [15] trained on audio features. Reported performance was lower for Valence ($R^2 = 0.281$) than for Arousal ($R^2 = 0.583$). Eerola et al. [16] compared various linear regression models at predicting multidimensional emotion ratings with acoustical features. A set of film soundtrack excerpts collected in [4] were used in this experiment. The best models based on Partial Least Squares Regression (PLS) showed high performance at predicting listener ratings of Valence, Arousal, and Tension ($R^2 = 0.72, 0.85, 0.79$). Especially for Valence, the performance was strikingly higher than in [14]. The same soundtrack data was utilised in classification of music to four basic emotion categories in [9], showing the maximum accuracy of 56.5%. Audio features related to tonality

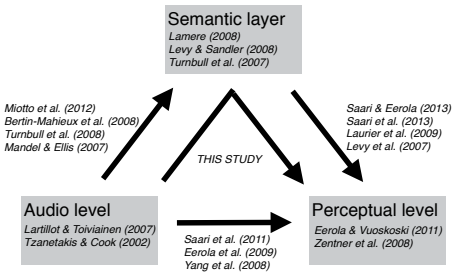


Figure 1. The difference of the present and past studies in mapping between audio features, semantic layer, and perceptual level. Selected past research is cited for each sub-task.

(average majorness of the mode and key clarity), as well as to the average slope of the onset attacks were found to be the most effective predictors of the perceived mood. SVM has been particularly popular in the annual MIREX mood classification challenge [2] representing the state-of-the-art in the field. Moreover, SVM together with ReliefF feature selection produced competitive results [17].

2.3 Mapping from Semantic Layer to Perceived Mood

The studies of Laurier et al. [18] and Levy et al. [19] compared semantic models of mood based on social tags to emotion models proposed by research in affective sciences, as well as expert-generated mood categories used in the MIREX challenge. The accuracy of tag-based semantic models at predicting listener ratings of musical mood was assessed in [20], proposing a technique called Affective Circumplex Transformation (ACT) for the task, based on previous research in affective sciences [21, 22].

ACT was used to predict perceived mood in 600 popular music tracks. The results showed promising performance ($R \approx 0.60$) for the ratings related to the dimensional emotion model as well as separate mood terms. Similar analysis across separate sources of curated editorial annotations for production music, and crowd-sourced Last.fm tags for commercial music, was performed in [23]. The results suggested that semantic models of mood based on tags can be used interchangeably to predict perceived mood across different annotation types and track corpora.

To apply the approach taken in [20] and [23] to new track corpora, semantic annotations need to be available for the corresponding tracks. In order to predict mood in unannotated track corpora, one must rely on other type of information, such as audio features. In the present study, we show how semantic tag data that was found to be promising and relevant in previous work can be used to enhance audio-based mood prediction.

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

	# Tracks	# Terms	# Terms / track
Mood set	259,593	357	4.44
SET10K	9,662	357	5.53

Table 1. Statistics of the mood term sets.

3. METHODOLOGY

3.1 Semantic Computing of Mood in Music

The following procedures were applied to uncover a semantic space of mood in music. More detailed account on the analysis and data collection is given in [20].

3.1.1 Vector-Space Modelling

First, a mood vocabulary was collected by aggregating and lemmatising mood term lists from several research papers in affective sciences, music psychology and Music Information Retrieval (MIR), and term lists in the Allmusic.com web service (see [20] for details). Synonyms and inflected forms of the vocabulary terms were identified and aggregated or added manually (e.g., happy $\approx$ happiness), resulting in 568 unique terms.

Semantic computation was applied to audio tracks and mood tags collected in [20]. Mood vocabulary terms were identified in tags using a bag-of-words approach similar to [1], and terms were applied to associated tracks accordingly. We excluded tracks with less than 2 mood annotations, as well as terms associated to less than 100 tracks, to avoid working with overly sparse information. Table 1 shows the resulting data (mood set) (SET10K is described in Section 3.2). Finally, the mood data set was normalised by computing Term Frequency - Inverse Document Frequency (TF-IDF) weights: $\hat{n}_{i,j} = (n_{i,j}+1)\log(\frac{R}{f_i})$, where $n_{i,j}$ is the original frequency weight related to term w_i and track t_j, R is the total number of tracks, and f_j is the number of tracks term w_i is associated to.

3.1.2 Latent Semantic Modelling

A low-rank approximation of the TF-IDF matrix was computed by Singular Value Decomposition (SVD) and Multidimensional Scaling (MDS). SVD decomposes a sparse matrix N so that $N = USV^T$, where matrices U and V are orthonormal and S is the diagonal matrix containing the singular values of N. Rank k approximation of N is computed by $N^k = U^k S^k (V^k)^T$, where the i:th row vector U_i^k represents a term w_i as a linear combination of k dimensions. Similarly, V_j^k represents track t_j in k dimensions. Based on a rank k approximation, dissimilarity between terms w_i and $w_{\hat{i}}$ is calculated by using the cosine distance between $U_i^k S^k$ and $U_{\hat{i}}^k S^k$.

To represent mood terms explicitly in a low-dimensional space, non-metric MDS [24] with Kruskal's stress-1 criterion was applied on the term dissimilarities, obtained by the rank k approximation of mood TF-IDF using SVD.

Next, we used the Affective Circumplex Transformation (ACT) proposed in [20] to conform the MDS configuration to the space of Arousal and Valence (AV), using AV values of 101 mood terms given in [21, p. 1167] and [22, p. 54].

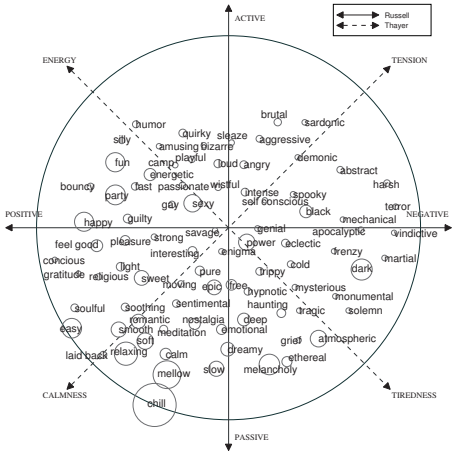


Figure 2. ACT with rank $k = 16$. Only the most frequently applied tags are shown for each part of the AV-space. Tag frequencies are reflected by the size of circles.

This technique is used here to *(i)* increase the interpretability of the MDS configuration; and *(i)* allow us to directly predict mood from the semantic layer. The first two dimensions of the resulting space represent Valence and Arousal as shown in Fig. 2 (with $k = 16$). The size of the circles reflects the frequencies of tags in the mood set, ranging from 110 tracks ("vindictive") to 79,524 tracks ("chill").

Finally, to represent a track in the MDS term space, we applied projection based on the positions of the associated terms. Given an MDS term configuration $y_i = (y_{i1}, y_{i2}, y_{i3})$, $i \in (1, ..., |w|)$, position of a track represented by a sparse term vector q is computed by the center-of-mass:

$$\hat{t} = \frac{\Sigma_i q_i y_i}{\Sigma_i q_i}. \quad (1)$$

3.2 Data set Description

Two data sets were used in our analysis: a 9,662 track subset of the mood set (SET10K), and a set of 600 tracks (SET600) collected in [20]. The audio tracks in both sets are non-overlapping.

SET10K was sampled from the mood set in a balanced manner by optimising mood variance in terms of track projections in the semantic space and including only unique artists. We use this set in successive analysis for mapping audio features to the semantic layer of mood. Audio content of the SET10K consists of 15-30s preview clips obtained from Last.fm. The clips are typically samples of full tracks in 128kB/s mp3 format starting from 30s-60s into the audio. Arguably, these samples contain relevant material that, up to a certain limit, characterise the full tracks.

SET600 was annotated in a listening test [20], where 59 participants rated 15s excerpts of 600 popular music tracks from Last.fm in terms of perceived mood expressed by music. Moods were rated in nine point Likert-scales for

Valence ("negative" / "positive"), Arousal ("calm" / "energetic"), Tension ("relaxed" / "tense"), Atmospheric, Happy, Dark, Sad, Angry, Sensual and Sentimental. The excerpts were sampled from full tracks corresponding to positions in the Last.fm previews. SET600 consists of 15s clips using 320kB/s mp3 format.

3.3 Audio Feature Extraction

Audio features describing dynamics, rhythm, pitch, harmony, timbre and structure were extracted from SET10K and SET600 using the MIRtoolbox [6]. Statistical means and standard deviations over features extracted from various short 50% overlapping time frames were computed to obtain song-level descriptors. The resulting set of 128 features is presented in Table 2. For the features describing rhythmic repetition (127-128) and zero crossing rate (43-44), we used long frame length of 2s, whereas for chromagram-based features such as the repetition of register (125-126), key clarity (19-20), centroid (17-18), mode (21-22), HCDF (23-24), and roughness (25-26) we used a frame length of 100ms. For other features the frame length was 46.4ms except for low-energy ratio (3), which was extracted directly from the full extent of the signal.

Features from SET10K were normalised using the z-score transform. All feature values more than 5 standard deviations from zero were considered outliers and truncated to the extremes $[-5, 5]$ (0.1% and 1.3% of the values in SET10K and SET600 respectively). SET600 was then normalised according to the means and standard deviations of SET10K. In particular, we discovered a slight discrepancy in mean RMS energy (1) between SET10K and SET600. The energy was generally higher in SET600, perhaps due to the use of different MP3 encoders. However, this was ignored in our study for simplicity.

3.4 Regression Techniques and Model Evaluation

3.4.1 Semantic Layer Projection

We propose a novel technique for mood prediction in music termed Semantic Layer Projection (SLP). The technique involves mapping audio features to perceived mood in two stages using the semantic mood level as a middle layer, instead of the conventional way of mapping audio features directly to the perceived mood. SLP may be implemented with several potential mapping techniques. We choose to use PLS for the first mapping, due to its higher performance demonstrated in previous research, and linear regression for the second.

First, we apply PLS to the SET10K to produce a mapping from audio features to the 10-dimensional semantic mood representation obtained using ACT. We compare two variants of the semantic mood layer: (SLP_{10D}) track projections in all 10 dimensions of the mood space, and (SLP_{1D}) track projections in separate dimensions corresponding to Valence (1st dim.), and Arousal (2nd dim.). To map from audio features to the semantic layer, we apply PLS to each dimension separately. Then, we project the audio features of SET600 to the semantic layer using the obtained mappings. Finally, we apply linear regression between the 10-dimensional (SLP_{10D}) and 1-dimensional (SLP_{1D}) layer representations and the listener ratings.

Table 2. Extracted feature set. Feature statistics (m = mean, d = standard deviation) are computed across sample frames.

Category	No.	Feature	Stat.
Dynamics	1-2	RMS energy	m, d
	3	Low-energy ratio	–
	4-5	Attack time	m, d
	6-7	Attack slope	m
Rhythm	8-9	Fluctuation (pos., mag.)	m
	10	Event density	m
	11-12	Pulse clarity	m, d
	13-14	Tempo	m, d
Pitch	15-16	Pitch	m, d
	17-18	Chromagram (unwr.) centr.	m, d
Harmony	19-20	Key clarity	m, d
	21-22	Key mode (majorness)	m, d
	23-24	HCDF	m, d
	25-26	Roughness	m, d
Timbre	27-28	Brightness (cutoff 110 Hz)	m, d
	29-30	Centroid	m, d
	31-32	Flatness ($<$ 5000 Hz)	m, d
	33-34	Irregularity	m, d
	35-36	Skewness ($<$ 5000 Hz)	m, d
	37-38	Spectr. entropy ($<$5000 Hz)	m, d
	39-40	Spectr. flux	m, d
	41-42	Spread	m, d
	43-44	Zerocross	m, d
MFCC	45-46	1st MFCC	m, d
	⋮	⋮	⋮ ⋮
	69-70	13th MFCC	m, d
	71-96	1st -13th Δ MFCC	m, d
	97-122	1st-13th $\Delta(\Delta)$ MFCC	m, d
Structure	123-124	Repetition (spectrum)	m, d
	125-126	Repetition (register)	m, d
	127-128	Repetition (rhythm)	m, d

We optimise the number of components used in the PLS mappings using $50 \times$ 2-fold cross-validation. In each fold, we divide SET10K into training and test sets, and estimate how well the PLS mapping based on train set fits the test set. To decide on the number of components, we apply (50, 100)-fold cross-indexing proposed in [9]. Cross-indexing is a technique developed to tackle model over-fitting in choosing the optimal model parameterisation from several candidates. Finally, we use the selected number of components to form a model based on the whole SET10K.

3.4.2 Baseline Techniques

In this study, two baseline techniques – PLS and Support Vector Regression (SVR) – were compared with SLP. These techniques were chosen since they represent regression methods that were already found efficient in previous MIR studies. Baseline techniques were applied in the usual way, mapping audio features of SET600 directly to the ratings of perceived mood.

We use PLS in a conventional way with 2 components as in [16]. In SVR, we use the Radial Basis Function (RBF) kernel and apply grid search to optimise the cost ($C = 2^l$, $l \in [-3, ..., 3]$) and gamma ($\gamma = 2^l$, $l \in [-13, ..., 8]$) model parameters. Moreover, we optimise the set of audio features used in SVR by feature subset selection. To

this end, we apply the ReliefF [25] feature selection algorithm adapted for regression problems. ReliefF produces relevance weights $\tau \in [-1, 1]$ for the individual features by taking into account their prediction potential and redundancy. To choose a subset of the features, we use a relevance weight threshold $\tau_0 = 0$ and include all features with $\tau > \tau_0$.

3.4.3 Cross-Validation Procedure

For validating the performance of the techniques, we use 50×2-fold cross-validation corresponding to 2-fold cross-validation run 50 times, and report the mean and standard deviation over the 100 performance estimates for each technique. All model optimisation and feature selection is based solely on the training set at each run.

4. RESULTS AND DISCUSSION

In SLP_{10D} and SLP_{1D} we use the rank $k = 16$ for SVD computation. This choice of k was found effective in [20], while other values had no consistent effect on the performance and did not improve the results.

Fig. 3 shows the performance of each technique at predicting the ratings for Valence and Arousal. For Valence, it is evident that SLP outperformed the baseline techniques. SLP_{10D} gave the highest performance ($R^2 = 0.334 \pm 0.035$), outperforming SLP_{1D} ($R^2 = 0.252 \pm 0.032$). SLP_{10D} performed at significantly higher level ($t(99) = 17.994, p = 5.63 \times 10^{-33}$)[3] than SVR ($R^2 = 0.245 \pm 0.048$), while the difference between SLP_{1D} and SVR was not significant. Conventional PLS was the least efficient with a performance of $R^2 = 0.152 \pm 0.045$.

Cross-indexing to optimise the number of PLS components in mapping from audio features to the semantic space yielded 7 components for SLP_{10D} and 13 components for SLP_{1D}. The number of components for SLP_{10D} is the average across 10 dimensions, while the latter relates to the first dimension of SLP_{10D}. The regression model used in the second-stage mapping of SLP_{10D} relied heavily on the first semantic dimension related to Valence: the first dimension showed an average significance of $p \approx 10^{-4}$ across cv-folds. SLP_{10D} model therefore bears a strong similarity to the SLP_{1D}. ReliefF feature selection to optimise the set of audio features used in SVR yielded on average 43 features ($SD = 11$).

In general, the fact that SLP_{1D} outperformed SVR shows the efficiency of SLP. In SLP_{1D} tracks are explicitly projected to Valence already in the first-stage mapping from the audio features to the semantic layer. Therefore minimal learning is required within SET600 for the second-stage mapping to perceived mood. This contrasts to the extensive adaptation to SET600 in SVR, which involves feature selection, cost and gamma optimisation, as well as support vector optimisation.

The overall performance for predicting Valence was at a significantly lower level than the performance of $R^2 = 0.72$ reported in [16]. Most notably, the PLS technique that was successful in [16] did not give convincing performance here. Since the set of audio features used in these studies is similar, the difference in performance is possibly due to the variety of genres covered by SET600. This is in contrast with the previous study using only film soundtracks. Film music is composed to mediate powerful emotional cues [4], which may provide higher variance in feature values so that better representations can be learnt. However, the performance in the present study is in line with other past research such as [14] ($R^2 = 0.281$).

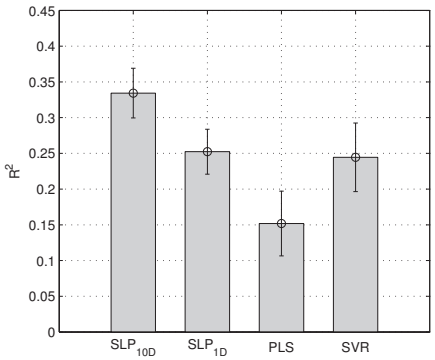
All techniques gave notably higher performance for Arousal than for Valence. In this case, SLP_{10D} again yielded the highest values ($R^2 = 0.782 \pm 0.020$), but outperformed SVR ($R^2 = 0.770 \pm 0.028$) only marginally. PLS gave the third highest performance ($R^2 = 0.751 \pm 0.027$) outperforming SLP_{1D} ($R^2 = 0.745 \pm 0.019$). For Arousal, SLP_{1D} used five PLS components, while the performance of SVR was obtained with 37 features on average ($SD = 9$). Again, the second-stage regression model in SLP_{10D} relied mainly on the 2nd dimension ($p \approx 2 \times 10^{-9}$) related to the Arousal dimension used in SLP_{1D}. Despite more complex training within SET600, SLP_{10D} gave only slight, although highly significant ($t(99) = 5.437, p = 5.4 \times 10^{-7}$) performance gain over SVR. In fact, all techniques performed better than $R^2 = 0.7$, which corroborates past findings that audio features provide a robust basis for modelling perceived Arousal in music.

Similar patterns in the general performance levels between techniques were found in modelling ratings in the other seven scales related to individual mood terms. In general, moods that are characterised by high or low arousal, such as Angry and Atmospheric, performed at similar, yet slightly lower level than Arousal, whereas moods such as Happy and Sad – characterised by positive and negative valence – produced performance similar to Valence.

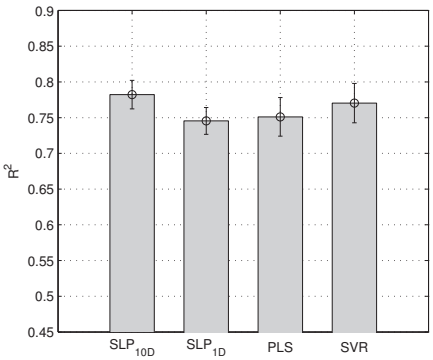
Since SLP_{10D} produced clearly the highest performance for Valence, while outperformed SVR by a more modest margin for Arousal, it is worth to compare the potential of these techniques in future approaches to mood prediction. SVR represents a sophisticated state-of-the-art technique that is efficient in learning characteristics of the training data relevant to the target mood, but requires complex optimisation of multitude of model parameters. Robust learning of SVR, and any method that could be used as baseline is solely dependent on high quality training data, which is typically laborious to collect. This also means that generalizability of these models to unknown music tracks, and possibly to new music genres, can not be guaranteed, as found in [26]. On the other hand, the efficiency of SLP is primarily based the first-stage mapping from audio to the semantic layer, and require only minimal adaptation to test data. This is suggested by the promising results of SLP_{1D} that produced explicit mood estimates already at the first-stage.

Semantic data required to built the semantic layer can be collected from online services by crowd-sourcing. Some services already make available data related to millions of tracks. Therefore, the cost of collecting training data for SLP is related mostly to obtaining the audio representation of the training set. Larger data for the semantic layer

[3] Pairwise Student's t-test across cv-folds.



(a) Valence.



(a) Arousal.

Figure 3. Performance ($R^2 \pm sd$) for each technique in predicting the perceived mood.

enables more delicate learning and would presumably increase the model performance. We therefore claim that the potential of SLP in future mood prediction approaches is higher than that of SVR. Note, however, that as SLP in general can be implemented with any prediction model, SVR can in fact be implemented in the future as the mapping technique within SLP.

Finally, we seek to gain understanding of what audio features are the most useful for modelling Valence and Arousal. We apply SLP_{10D} using each audio feature category described in Table 2 separately. Table. 3 shows the results. Eight harmony-related features including Mode and Key clarity were found to be the most useful in predicting Valence ($R^2 = 0.186$), and in fact, the model using only these 8 features would have outperformed PLS using all features. Features describing timbre, structure, and MFCC showed modest potential for predicting Valence ($R^2 > .10$), whereas rhythm features were largely redundant in this particular task. Prediction of Arousal was on the other hand highly efficient with most feature categories. Timbre ($R^2 = 0.687$) and MFCC ($R^2 = 0.649$) features performed the best. Prediction with harmony-related features was also competitive ($R^2 = 0.653$), while even the four pitch-related features could predict Arousal at moderate level ($R^2 = 0.471$).

In general, these results support previous findings that harmony-related features are useful in mood prediction [9], and that timbre-related features are more useful for predicting Arousal. The results also highlight the need to either optimise existing harmony-related features, or to uncover and investigate a wider variety of audio descriptors for Valence prediction.

Table 3. Performance ($R^2 \pm sd$) of SLP_{10D} using different audio feature categories. Number of features in each category are presented in brackets.

	Valence	Arousal
Dynamics (7)	0.092 ± 0.031	0.536 ± 0.034
Rhythm (7)	0.056 ± 0.044	0.583 ± 0.028
Pitch (4)	0.074 ± 0.034	0.471 ± 0.031
Harmony (8)	$\mathbf{0.186 \pm 0.035}$	0.653 ± 0.030
Timbre (18)	0.141 ± 0.037	$\mathbf{0.687 \pm 0.027}$
MFCC (78)	0.123 ± 0.030	0.649 ± 0.026
Structure (6)	0.127 ± 0.043	0.547 ± 0.025

5. CONCLUSIONS

In this study we developed a novel approach to predict the perceived mood in music called Semantic Layer Projection (SLP). By introducing a two-stage mapping from audio features to semantic layer and finally to mood ratings, SLP provides a way to exploit semantic information about mood learnt from large music collections. It also facilitates building predictive models for disparate music collections. The proposed technique outperformed SVR, a sophisticated predictive model on the Valence dimension, and produced prediction performance roughly at the same level on the Arousal dimension.

The results highlight the difficulty of modelling the Valence dimension in music. However, SLP provides clear advantage compared to baseline techniques specifically in this task, which signifies its high potential that can be developed further in more general audio and semantics-based mood recognition models.

Future direction of the present study includes using more efficient collection of tracks to represent the semantic layer, and improving the prediction of Valence via an extension of the audio feature set. Moreover, a version of the proposed technique that takes musical genre into account – possibly by introducing a genre layer – will be developed to further generalise our model to many different types of music collections.

6. REFERENCES

[1] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.

[2] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Towards musical query-by-semantic-description using the CAL500 data set," in *Proceedings of the 30th international ACM SIGIR conference on information retrieval*, 2007, pp. 439–446.

[3] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[4] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music." *Psychol. Music*, vol. 39, no. 1, pp. 18–49, 2011.

[5] M. Zentner, D. Grandjean, and K. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.

[6] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects, Bordeaux, France*, September 2007.

[7] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.

[8] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models," in *Proc. of the 9th Int. Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012, pp. 492–507.

[9] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 6, pp. 1802 –1812, aug. 2011.

[10] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[11] M. I. Mandel and D. P. Ellis, "Multiple-instance learning for music information retrieval," in *Proceedings of 9th International Conference of Music Information Retrieval (ISMIR)*, 2008, pp. 577–582.

[12] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *Journal of New Music Research*, vol. 37, no. 2, pp. 115–135, 2008.

[13] R. Miotto and G. Lanckriet, "A generative context model for semantic music annotation and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1096–1108, 2012.

[14] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.

[15] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[16] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *9th International Conference on Music Information Retrieval*, 2009, pp. 621–626.

[17] R. Panda and R. P. Paiva, "Music emotion classification: Dataset acquisition and comparative analysis," in *n 15th International Conference on Digital Audio Effects (DAFx-12)*, 2012.

[18] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags," in *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 381–86.

[19] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[20] P. Saari and T. Eerola, "Semantic computing of moods based on tags in social media of music," *IEEE Transactions on Knowledge and Data Engineering*, manuscript submitted for publication available at http://arxiv.org/, 2013.

[21] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[22] K. R. Scherer, *Emotion as a multicomponent process: A model and some cross-cultural data*. Beverly Hills: CA: Sage, 1984, pp. 37–63.

[23] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of mood expressed by music: Comparison between crowd-sourced and curated editorial annotations," in *IEEE International Conference on Multimedia and Expo (ICME 2013): International Workshop on Affective Analysis in Multimedia (AAM)*, In press 2013.

[24] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.

[25] M. Robnik-Sikonja and I. Kononenko, "An adaptation of relief for attribute estimation in regression," in *Proceedings of the Fourteenth International Conference on Machine Learning*, ser. ICML '97. San Francisco, CA: Morgan Kaufmann Publishers Inc., 1997, pp. 296–304.

[26] T. Eerola, "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*, vol. 40, no. 4, pp. 349–366, 2011.

V

THE ROLE OF AUDIO AND TAGS IN MUSIC MOOD PREDICTION: A STUDY USING SEMANTIC LAYER PROJECTION

by

Pasi Saari, Tuomas Eerola, György Fazekas, Mathieu Barthet, Olivier Lartillot & Mark Sandler 2013

In Proceedings of the 14th International Society for Music Information Retrieval Conference (ISMIR), 201-206

THE ROLE OF AUDIO AND TAGS IN MUSIC MOOD PREDICTION: A STUDY USING SEMANTIC LAYER PROJECTION

Pasi Saari*, Tuomas Eerola*, György Fazekas†, Mathieu Barthet†, Olivier Lartillot*, Mark Sandler†
*Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland
†Centre for Digital Music, Queen Mary University of London, United Kingdom
*{firstname.lastname}@jyu.fi, †{firstname.lastname}@eecs.qmul.ac.uk

ABSTRACT

Semantic Layer Projection (SLP) is a method for automatically annotating music tracks according to expressed mood based on audio. We evaluate this method by comparing it to a system that infers the mood of a given track using associated tags only. SLP differs from conventional auto-tagging algorithms in that it maps audio features to a low-dimensional semantic layer congruent with the circumplex model of emotion, rather than training a model for each tag separately. We build the semantic layer using two large-scale data sets – crowd-sourced tags from Last.fm, and editorial annotations from the I Like Music (ILM) production music corpus – and use subsets of these corpora to train SLP for mapping audio features to the semantic layer. The performance of the system is assessed in predicting mood ratings on continuous scales in the two data sets mentioned above. The results show that audio is in general more efficient in predicting perceived mood than tags. Furthermore, we analytically demonstrate the benefit of using a combination of semantic tags and audio features in automatic mood annotation.

1. INTRODUCTION

Our daily experiences with music, together with strongly corroborated research evidence [1], suggest that music has a remarkable ability to induce as well as to express emotions or moods. For this reason, the mood associated with a musical piece is often a key aspect in music listening. This provides clear motivations for creating Music Information Retrieval (MIR) systems to organize, navigate or access music collections based on mood. These systems typically rely on mood models and appropriately selected machine learning techniques [2, 3]. Among several models proposed for emotions, the Circumplex model [4, 5] connecting mood terms to underlying emotion dimensions of valence (positive / negative) and arousal (active / passive) is one of the most popular [6]. On the other hand, Thayers variant [7] of this model suggests dimensions of tension and energy diagonal to arousal and valence. However, training machine learning models that automatically associate musical pieces with moods require high quality human mood annotations that are laborious to create, hence typically limited in amount.

Mood-related tags, i.e., free-form labels applied to artists, albums, tracks, etc., are abundantly available from popular online services such as Last.fm [1], while editorial track-level mood tags are vital in large production music catalogues. However, due to issues related to noise and ambiguity in semantic relations between tags, uncovering reliable mood representations from tag data requires typically filtering and semantic analysis [8, 9]. Previous research showed that semantically processed information using track-level Last.fm tags is congruent with listener ratings of valence, arousal, tension and various mood terms [10]. In a test set of 600 popular music tracks, moderate to high ($.47 < r < .65$) correlation was found using the Affective Circumplex Transformation (ACT) technique, that is based on Latent Semantic Analysis (LSA) and the circumplex model of emotions. These results outperformed several conventional semantic analysis techniques, and notably, raw tag frequency scores ($.16 < r < .47$). The robustness of ACT was also demonstrated in [11], by applying the technique to editorial tags from a production music library of about 250,000 tracks.

In a wider context, modelling mood, and thus estimating mood tags may be seen as a specific form of auto-tagging, which is a popular research topic in MIR. A system is typically trained using audio features extracted from a collection of tracks and their associated tags. Then, the trained model is utilised to label new untagged tracks automatically given their features. Typical auto-tagging studies have trained models independently for each tag [12–14], omitting semantic associations between tags, while results in [15] and [16] showed that post-processing auto-tags according to their semantic similarity increases the performance. These techniques have produced promising results for mood tags, possibly due to the use of cleanly-labeled tag data collected for research purposes. As shown in [10], a considerable semantic gap exists between raw crowd-sourced mood tags and verified listener ratings. However, semantic computing provides a promising direction, not yet exploited to the full extent in auto-tagging, for capturing reliable information from large tag collections.

Previous studies in auto-tagging have compared predict-

[1] http://www.last.fm

ed tags with human-labeled tags as a way to assess performance. By considering listener ratings as ground-truth rather than tags, in this paper we analytically compare auto-tags to actual tags as predictors. The two representations relate to two distinct assumptions in tag estimation: either human-labeled tags or audio is available for each new track. Our aim is to challenge these assumptions by highlighting the benefit of semantic computing in the context of music mood auto-tagging. Semantic Layer Projection (SLP) proposed in [17] provides a robust method for projecting the audio feature space to multi-dimensional semantic layer based on ACT. SLP followed by linear regression with the projected feature components outperformed state-of-the-art regression models in predicting listener ratings of valence in 600 tracks from Last.fm. In this paper we evaluate the benefits of SLP in auto-tagging using two corpora: tracks and crowd-sourced mood tags from Last.fm tags as well as tracks and curated editorial mood tags obtained from the I Like Music (ILM) production music catalogue. We predict listener ratings of moods in separate test sets extracted from these corpora.

The rest of the paper is organised as follows: Section 2 describes the tag data and the ACT technique for building the semantic space of moods based on the tags, the set of audio features, and the SLP technique for predicting mood in new tracks based on ACT and the features. Section 3 gives a detailed account of the experimental setup, the data sets used for SLP evaluation, baseline techniques, and the method for comparing mood prediction based on tag or audio information of new tracks. Section 4 shows the results of the experiments and conclusions are drawn in Section 5.

2. METHODOLOGY

2.1 Affective Circumplex Transformation

We used two sources of tracks and tags in the analysis: 259,593 tracks from Last.fm and 226,344 tracks from I Like Music (ILM) production music catalogue, associated with 357 and 288 mood terms, respectively. To create these data sets, tags associated to track sets from the two sources were first lemmatized and identified from a vocabulary of 560 mood terms, aggregated from mood words obtained from selected research papers in affective sciences, music psychology and MIR, as well as from the Allmusic.com web service. In both data sets, tracks with only one tag, and tags associated with less than 100 tracks were then excluded. Finally, the tag data was normalised using term frequency-inverse document frequency (TF-IDF) weights. A detailed account of the data sets and the above process is given in [10, 11].

The following process was applied to Last.fm and ILM sets separately. To uncover semantic similarity between individual mood terms, a low-rank approximation of the TF-IDF matrix was computed using Singular Value Decomposition (SVD) and Multidimensional Scaling (MDS) as in [10]. SVD decomposes a sparse TF-IDF matrix N into orthogonal matrices U and V, and a diagonal matrix S with singular values in decreasing order, such that $N = USV^T$. A rank k approximation of N is then computed by $\tilde{N}_k = U^k S^k (V^k)^T$, where each row vector U_i^k represents the terms w_i with k relative weights for each dimension. Similarly, V_j^k represents track t_j as k relative weights. Based on the rank k approximation, dissimilarity between terms w_i and $w_{\hat{i}}$ can be computed using the cosine distance between the $U_i^k S^k$ and $U_{\hat{i}}^k S^k$ vectors. To represent mood terms explicitly in a low-dimensional space that resembles the arousal-valence space, MDS was applied on the term distances to obtain a three-dimensional configuration. The choice of using three dimensions instead of two is motivated by the debate around whether two dimensions is enough to capture relevant variance in moods. Past research have proposed various candidates for the third dimension, such as dominance, potency, or movement.

Next we applied the Affective Circumplex Transformation (ACT) to conform the MDS configuration to the space of *arousal* and *valence* (AV), using AV values of 101 mood terms given in [4, p. 1167] and [5, p. 54]. This technique takes advantage of the Procrustes transformation [18] involving translation, reflection, orthogonal rotation, and isotropic scaling using sum of squared errors as goodness-of-fit. The motivation for this is to *i)* increase the interpretability of the MDS configuration, and *ii)* enable direct prediction of arousal and valence from the semantic space. The technique yields a mood term configuration $x_i = (x_{1,i}, x_{2,i}, x_{3,i}), i = 1, ..., nterms$. A subset of Last.fm and ILM mood term configurations are visualised in Fig. 1 (with $k = 16$). The frequencies of the terms across tracks (co-occurence counts) range from 110 ("vindictive") to 79,524 ("chill") for Last.fm, and 346 ("narrative") to 39,892 ("uplifting") for ILM. Each track j was projected onto the resulting space by taking the Euclidean mean of the term positions, weighted by the sparse TF-IDF vector q_j of the track:

$$t_j = (\Sigma_i q_{j,i} x_i)/(\Sigma_i q_{j,i}). \quad (1)$$

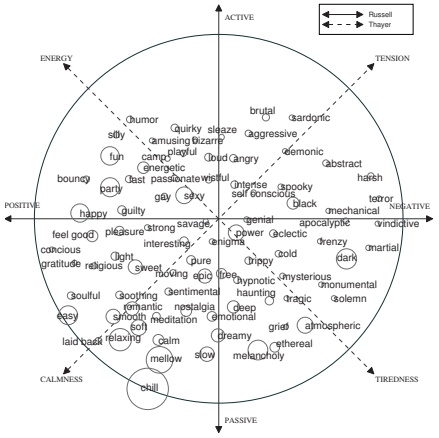
Finally, explicit mood-term-specific weights for the track with position t_j were computed using:

$$P_{j,i} = (x_i/|x_i|) \cdot t_j, \quad (2)$$

whereas arousal and valence for a track was estimated directly by using the positions along corresponding dimensions. Tension was obtained by projecting tracks along the direction (-1,1,0) as suggested in [7] (see Fig. 1). The analysis in [10] showed that the value of the rank parameter k in SVD computation has minor effect on ACT performance. Therefore we chose to use a heuristically selected value $k = 16$ in our analysis.

2.2 Audio Feature Extraction

Audio features describing dynamics (RMS energy, Low-energy ratio, Attack time, Attack slope), rhythm (Fluctuation pos. & mag., Event density, Pulse clarity, Tempo), pitch (avg. pitch, Chromagram unwrapped centroid), harmony (Key clarity, Mode [majorness], Harmonic change, Roughness), timbre (Brightness, Irregularity, Zerocrossings, Spectral Centroid, Flatness, Skewness, Entropy, Flux



(a) Last.fm.

(b) ILM.

Figure 1. Two first dimensions (valence–arousal) of the three-dimensional mood term configurations obtained with ACT ($k = 16$) for (a) Last.fm and (b) ILM.

and Spread), and structure (Spectral, Rhythmic and Registral repetition) as well as 13 MFCCs, Δ MFCCs, and $\Delta(\Delta)$ MFCCs were extracted from the data sets presented in Table 1 using the MIRtoolbox [19]. To characterise tracks using audio features, statistical means and standard deviations were computed for each feature extracted over short 50% overlapping time frames, yielding a 128 element feature vector for each track. For the features describing the rhythmic repetition and zero crossing rate, we used longer frame lengths of 2s, whereas for chromagram-based features such as the repetition of register, key clarity, centroid, mode, harmonic change, and roughness we used a frame length of 100ms. For other features the frame length was 46.4ms, except for low-energy ratio which is a track-level feature by definition.

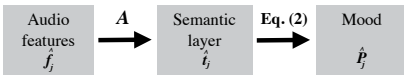


Figure 2. Mapping process in SLP for a novel track represented by audio features $\hat{f}_j$.

	Last.fm		ILM	
	SET10K	SET600	SET5K	SET205
# Tracks	9,662	600	4,692	205
# Terms	357	357	288	288
Term density (%)	1.59	2.43	1.86	2.38

Table 1. Statistics of the mood term sets.

2.3 Semantic Layer Projection

Semantic Layer Projection (SLP), originally proposed in [17], is a technique for the automatic annotation of audio tracks with mood using audio features. SLP is trained on a large collection of audio features and associated tag information. The difference between SLP and conventional auto-tagging is that audio features are not directly mapped to individual tags, but to three-dimensional semantic representation of the mood space obtained by ACT. The mapping is determined by a training stage using the Partial Least Squares (PLS) method. PLS has been found efficient at handling high dimensional and collinear input variables, and it is shown to be robust when using a large number of observations [20].

Given a set F of m audio features related to n tracks $F^{n\times m} = (f_1, f_2, ..., f_n)$, and a semantic layer representation $T^{n\times 3} = (t_1, t_2..., t_n)$ for the corresponding tracks (see Eq. 1), the mapping matrix A between F and T is determined using PLS so that $T \approx AF$. We optimize the number of components in PLS by applying (50, 100)-fold cross-indexing [21]. Cross-indexing tackles problems of model overfitting when choosing the optimal parameterisation from several candidates.

The explicit mood of a previously unseen track represented by audio features $\hat{f}_j$ are estimated by first $\hat{t}_j = A\hat{f}_j$, and then by $\hat{P}_{j,i} = (x_i/|x_i|) \cdot \hat{t}_j$ as in Eq. 2. This process is summarised in Fig. 2.

3. EXPERIMENTAL SETUP

3.1 Data Sets

For both data sources, Last.fm and ILM, the semantic models are built from the full set of tracks (250,000 approx.), whereas mappings between audio features and the semantic space in SLP are trained on subsets of these large corpora, namely SET10K and SET5K. The performance of the model is evaluated using listener ratings of perceived moods in separate test sets: SET600 and SET205. Statistical measures of these data sets in terms of semantic mood content are summarised in Table 1.

SET10K consists of 9,662 tracks and was also used in [17]. The set was sampled from the Last.fm corpus in a balanced manner by *i)* optimising mood variance in terms

of track projections in the ACT space, *ii)* favouring tracks with many listeners according to Last.fm, and *iii)* including only unique artists. The audio content of SET10K consists of 15-30s Last.fm preview clips. The clips are typically samples of full tracks in the 128kB/s mp3 format, starting from 30s-60s into the beginning. We assume that these samples are sufficiently representative of the whole tracks. SET600 collected in [10] consists of 15s excerpts of 600 popular music tracks containing no overlapping artists with SET10K, and no tracks overlapping with the large Last.fm corpus. The set was fetched from Last.fm in a similar balanced manner as SET10K, with additional balancing across multiple popular music genres (jazz, pop, rock, electronic, folk and metal), and favouring tracks with many associated mood tags. SET600 was annotated in a listening test [10], with 59 participants rating the excerpts in terms of perceived mood expressed by music. Moods were rated in nine point bipolar Likert-scales for the mood dimensions of valence (negative / positive), arousal (calm / energetic), and tension (relaxed / tense), as well as in unipolar scales for individual mood terms atmospheric, happy, dark, sad, angry, sensual, and sentimental.

The 4,692 tracks from SET5K were picked up randomly from the ILM production music catalogue by *i)* keeping tracks with a duration of at least 60s (in order to discard short instances of the tracks), and *ii)* discarding instrumental stems, i.e. individual tracks from multitrack recordings. Six main genres were represented (jazz, dance, rock, electronic, folk and orchestral). 30s audio clip versions of the tracks were produced in the 128kB/s mp3 format. SET205, described in [11], consists of 205 clips of 30s duration from SET5K. The tracks were sampled in a similar fashion as for the Last.fm test set, but without taking listener statistics into account. The set was annotated by 46 participants in a similar manner as SET600, but for bipolar scales of valence (negative / positive), arousal (calm / energetic), tension (relaxed / tense), dominance (submissive / dominant), romance (cold / romantic), and humour (serious / funny) [11].

Features extracted from SET10K and SET5K were normalised using the z-score transform. All feature values with more than 5 standard deviations from zero were considered outliers and truncated to the extremes $[-5, 5]$. The features associated with SET600 and SET205 were then normalised according to the means and standard deviations of the larger feature sets.

3.2 Modelling Techniques

To show the efficiency of the mappings from audio features to the semantic layer, we compare SLP to two baseline techniques (BL1 and BL2) aiming at predicting mood ratings of e.g. valence, arousal, and tension in the test corpora. Prediction rates are computed as squared correlation coefficients (R^2) between the estimates and ratings over the test sets. The difference between the three techniques lies in how the semantic relationships between mood terms are exploited in the modelling. **BL1** uses mappings between audio features and individual mood terms directly, in order to predict mood ratings for the corresponding terms in the test corpora. This is analogous to the techniques used in [12–14]. **BL2** uses mappings between audio features and individual mood terms to predict each (term-track) pair in the test corpora. Test tracks are then projected using Eq. 1 and Eq. 2 based on the inferred tags. This is analogous to the techniques presented in [15, 16]. The **SLP** technique has been described in Section 2.3.

In short, BL1 does not use information about mood term relationships at all, while BL2 exploits the semantic information after producing a mapping from audio features to mood terms. SLP, on the other hand, maps audio features directly to the semantic layer.

Mappings in BL2 were trained for terms appearing at least ten times in SET10K and SET5K, amounting to 287 and 201 terms, respectively. Since valence, arousal, or tension are not explicitly modeled by BL1 (and no tags "valence" or "arousal" exist in either of the tag corpora), we use terms corresponding to the bipolar labels of the mood scales in the listening tests for modelling these ratings. Tags "positive", "energetic", and "relaxing" / "relaxed" were applied more often than tags "negative", "calm", and "tense" in both SET10K and SET5K, so we use the aforementioned tags to model the corresponding mood dimensions. Similarly, for dominance, romance, and humour that were rated in bipolar scales in SET205, we use tags "powerful", "romantic", and "funny".

Evaluating the role of tags and audio in predicting moods is achieved by comparing SLP and ACT prediction rates. While both of these techniques rely on the same semantic representation of moods, for each novel track, SLP uses only audio features and automatically inferred moods. ACT however uses actual tags associated with the track. We use these techniques in conjunction by computing the weighted mean of these two estimates for each track, and comparing that to the mood ratings. We vary the weights $[w, 1 - w]$ ($w \in [0, 1]$) for the techniques so that the case $w = 0$ corresponds to using ACT, whereas the case $w = 1$ corresponds to using SLP.

4. RESULTS AND DISCUSSION

4.1 Evaluation of SLP

Table 2 presents the comparison of SLP with the baseline methods. In case of Last.fm, prediction rates of SLP span from moderate ($R^2 = 0.248$ for happy) to considerably high ($R^2 = 0.710$ for arousal). SLP consistently outperforms both baseline methods, except in one case, where BL1 gives marginally higher performance for sad ($R^2 = 0.313$). The differences between the baseline techniques and SLP are however small for the arousal, angry, and sensual dimensions. We also note that valence and related moods (happy, sad, and angry) are the most difficult to predict with all of the models, and in turn, arousal is the easiest to predict. This is consistent with past studies in music emotion recognition [22]. Although BL1 suffers from the lack of explicit tags for valence, arousal, and tension to infer explicit predictions, results for the seven mood

		BL1	BL2	SLP
Last.fm	Valence	0.045	0.244	**0.322**
	Arousal	0.693	0.662	**0.710**
	Tension	0.198	0.469	**0.560**
	Atmospheric	0.075	0.541	**0.581**
	Happy	0.073	0.183	**0.248**
	Dark	0.264	0.314	**0.370**
	Sad	**0.313**	0.295	0.310
	Angry	0.475	0.465	**0.497**
	Sensual	0.505	0.523	**0.546**
	Sentimental	0.218	0.354	**0.390**
	Mean	0.286	0.405	**0.453**
ILM	Valence	0.156	0.330	**0.486**
	Arousal	0.680	0.672	**0.718**
	Tension	0.478	0.501	**0.588**
	Dominance	**0.461**	0.376	0.352
	Romance	0.274	0.301	**0.351**
	Humour	0.209	0.362	**0.502**
	Mean	.376	.424	**.499**

Table 2. Prediction rates (R^2) for the Last.fm and ILM test sets using SLP and two baseline methods (BL1 and BL2). For each dimension, best scores are reported in bold.

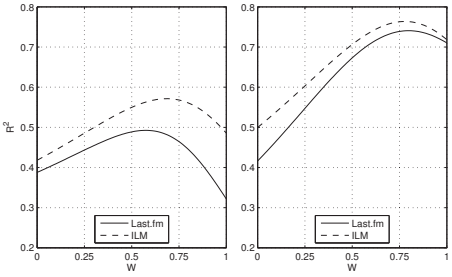
terms show that exploiting semantic associations between tags is highly beneficial. Moreover, as SLP outperforms BL2 for all mood dimensions, mapping tags to the semantic layer directly rather than projecting individual auto-tags to the layer is efficient.

In the case of the ILM data sets, our results show patterns that are highly consistent with those of Last.fm – in general SLP outperforms the baseline methods, while BL1 obtains the lowest performance, on average. However, the performance for valence is considerably higher ($R^2 = 0.486$) than for the Last.fm data set. A clear exception to this pattern is the higher performance of BL1 for dominance ($R^2 = 0.461$) compared to the other techniques. Since dominance is not directly captured either by the tags or the semantic layer dimensions, using other tags than "powerful" would have changed the modelling. In fact, tags "airy", "intimate", and "soft" yielded the highest performance for SLP ($R^2 > 0.57$), the tag "relaxed" yielded the highest performance for BL1 ($R^2 = 0.493$), and the tag "airy" yielded the highest performance for BL2 ($R^2 = 0.543$).

Overall, the results show the advantage of mapping audio features directly to a semantic layer to train predictive models for moods. This solution provides increased performance over methods not exploiting semantic associations at all, or projecting auto-tags to the semantic layer in a later stage, after mapping from audio features to mood tags.

4.2 Tags vs. Audio Features in Mood Prediction

To assess the importance of tags and audio in conjunction, systematic evaluation of using SLP and ACT separately or in conjunction using the weights was carried out. Overall, the results of such comparisons (see Fig. 3 and Table 3) first suggest that the predictions driven by audio features alone yield better performance. However, the combination of audio features and tags lead to a notable increase, especially for moods that are the most difficult for SLP



(a) Valence (b) Arousal

Figure 3. Prediction rate obtained when relying on different information in Last.fm and ILM test sets: tags ($w = 0$), audio ($w = 1$), or combination ($0 < w < 1$).

(valence, happy, and sad). For Last.fm, the mean of the maximum performance when using audio and tags in conjunction is higher ($R^2 = 0.531$) compared to the individual use of tags ($R^2 = 0.334$) and audio ($R^2 = 0.453$). Similar patterns can be observed with the ILM data, for which audio content-based methods outperform tag-based methods for all mood scales (0.062 and 0.164 increases in mean R^2 when both audio and tags are used in conjunction, compared to audio and tags alone, respectively). As can be seen on Fig. 3, the optimal weight for the combination varies to a small degree in both data sets, but lies around 0.70 (mean). In other words, the best prediction of mood is achieved when the acoustic features are attributed a higher weight and are supplemented by tag data, both projected via a semantic space.

However, there are significant exceptions to the conclusions drawn from simply tallying up the prediction rates across the models and data sets. In the Last.fm data set, audio features are actually worse than tags in explaining the ratings of the valence and happy dimensions. This is in line with a number of previous studies in mood prediction with audio features, such as [22], and may have to do with the fact that valence is an elusive concept in music, and maybe particularly dependent on music genres. Further research that extends the mutual patterns between mood and genre is required to untangle such specific results.

5. CONCLUSIONS

In this study, we demonstrated that mood prediction is efficient when relying on large-scale music tag data and audio features, and is boosted by exploiting semantic modelling. The results suggest that higher prediction rates are achievable using the semantic layer projection (SLP) technique when compared to baseline techniques related to conventional auto-tagging that do not incorporate semantic modelling into mappings from audio features.

We conclude that building large-scale predictive models for moods in music can be done more efficiently for certain mood dimensions by relying on audio features rather than

		Tags	max(R^2)	w	Audio
Last.fm	Valence	0.388	0.492	0.57	0.322
	Arousal	0.416	0.741	0.80	0.710
	Tension	0.392	0.618	0.71	0.560
	Atmospheric	0.298	0.607	0.83	0.581
	Happy	0.357	0.429	0.53	0.248
	Dark	0.328	0.506	0.71	0.370
	Sad	0.300	0.393	0.58	0.310
	Angry	0.221	0.518	0.84	0.497
	Sensual	0.371	0.584	0.73	0.546
	Sentimental	0.271	0.422	0.72	0.390
	Mean	0.334	0.531	0.70	0.453
ILM	Valence	0.418	0.571	0.69	0.486
	Arousal	0.500	0.764	0.78	0.718
	Tension	0.497	0.667	0.69	0.588
	Dominance	0.271	0.386	0.73	0.352
	Romance	0.261	0.386	0.75	0.351
	Humour	0.437	0.590	0.67	0.502
	Mean	0.397	0.561	0.72	0.499

Table 3. Prediction rate for the Last.fm and ILM test sets using tags (ACT), audio (SLP), or a weighted combination.

associated tags. This is supported by the higher overall performance of audio compared to tags, and by the overall stable performance of the predictions between the models in two different data sets, crowd-sourced tags from Last.fm and a curated production music corpus (ILM). These data sets consisted of nearly 250,000 tracks each, out of which different subsets were carefully utilized in model training and evaluation. The results also imply that mood tags for novel tracks are not crucial for the automatic annotation of tracks along most mood dimensions. However, for moods related to valence, the use of tags yields a considerable increase in the predictive performance when combined with audio feature-based estimations. In the future we will factor in music genre to the approach presented here.

Acknowledgements This work was partly funded by the Academy of Finland (The Finnish Centre of Excellence in Interdisciplinary Music Research) and the TSB project 12033 - 76187 Making Musical Mood Metadata (TS/J002283/1).

6. REFERENCES

[1] J. A. Sloboda and P. N. Juslin. *Music and Emotion*, chapter Psychological Perspectives on Music and Emotion, pages 71–104. Oxford University Press, New York, 2001.

[2] Z. Fu, G. Lu, K. M. Ting, and D. Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303 –319, 2011.

[3] M. Barthet, G. Fazekas, and M. Sandler. Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models. In *Proc. of the 9th Int. Symposium on Computer Music modelling and Retrieval (CMMR)*, pages 492–507, 2012.

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

[5] K. R. Scherer. *Emotion as a multicomponent process: A model and some cross-cultural data*, pages 37–63. CA: Sage, Beverly Hills, 1984.

[6] T. Eerola and J. K. Vuoskoski. A review of music and emotion studies: Approaches, emotion models and stimuli. *Music Perception*, 30(3):307–340, 2012.

[7] R. E. Thayer. *The Biopsychology of Mood and Arousal.* Oxford University Press, New York, USA, 1989.

[8] C. Laurier, M. Sordo, J. Serra, and P. Herrera. Music mood representations from social tags. In *Proceedings of 10th International Conference on Music Information Retrieval (ISMIR)*, pages 381–86, 2009.

[9] M. Levy and M. Sandler. A semantic space for music derived from social tags. In *Proceedings of 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.

[10] P. Saari and T. Eerola. Semantic computing of moods based on tags in social media of music. *IEEE Transactions on Knowledge and Data Engineering*, In press 2013.

[11] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler. Semantic models of mood expressed by music: Comparison between crowd-sourced and curated editorial annotations. In *IEEE International Conference on Multimedia and Expo (ICME 2013): International Workshop on Affective Analysis in Multimedia*, 2013.

[12] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[13] M. I. Mandel and D. PW Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[14] M. I. Mandel and D. P. Ellis. Multiple-instance learning for music information retrieval. In *Proceedings of 9th International Conference of Music Information Retrieval (ISMIR)*, pages 577–582, 2008.

[15] R. Miotto and G. Lanckriet. A generative context model for semantic music annotation and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1096–1108, 2012.

[16] T. Bertin-Mahieux, D. Eck, F. Maillet, and P. Lamere. Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135, 2008.

[17] P. Saari, T. Eerola, G. Fazekas, and M. Sandler. Using semantic layer projection for enhancing music mood prediction with audio features. In *Sound and Music Computing Conference*, 2013.

[18] J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*, volume 3. Oxford University Press Oxford, 2004.

[19] O. Lartillot and P. Toiviainen. A matlab toolbox for musical feature extraction from audio. In *Proceedings of the 10th International Conference on Digital Audio Effects*, 2007.

[20] S. Wold, M. Sjöström, and L. Eriksson. Pls-regression: a basic tool of chemometrics. *Chemometrics and intelligent laboratory systems*, 58(2):109–130, 2001.

[21] P. Saari, T. Eerola, and O. Lartillot. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. *IEEE Transactions on Speech and Audio Processing*, 19(6):1802 –1812, 2011.

[22] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen. A regression approach to music emotion recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):448–457, 2008.

VI

GENRE-ADAPTIVE SEMANTIC COMPUTING ENHANCES AUDIO-BASED MUSIC MOOD PREDICTION

by

Pasi Saari, György Fazekas, Tuomas Eerola, Mathieu Barthet, Olivier Lartillot & Mark Sandler

submitted

Genre-adaptive Semantic Computing Enhances Audio-based Music Mood Prediction

Pasi Saari, György Fazekas, Tuomas Eerola, Mathieu Barthet, *Member, IEEE,* Olivier Lartillot, and Mark Sandler, *Senior Member, IEEE,*

Abstract—This study investigates multiple genre-adaptive techniques for audio-based music mood prediction. A novel technique ACT+SLPwg is proposed that employs semantic computing based on social tags and audio-based modeling in a genre-adaptive manner. In the experimental evaluation various techniques are compared at predicting listener ratings of core affects and mood terms related to a set of 600 popular music tracks spanning multiple genres. The results show that the ACT+SLPwg outperforms other genre-adaptive alternatives and general models that do not exploit genre information. In particular, improvements in the prediction rates are obtained for the valence dimension that is typically the most challenging core affect dimension for audio-based prediction. This study also demonstrates that normative data from affective sciences does not improve on the semantic modeling of the mood space based on music-specific social tag data. Moreover, the study presents analytical insights into inferring concise music genre representation based on tag data.

Index Terms—Music information retrieval, mood prediction, social tags, semantic computing, music genre.

1 Introduction

Musical genre and mood are linked together in an intriguing manner. People tend to use particular genres for mood regulation [1], different genres are able to induce distinct emotional responses [2], and mood and genre terms are often combined to express musical qualities (e.g. 'smooth jazz' and 'dark ambient'). In the field of Music Information Retrieval (MIR), musical genre and mood have received considerable attention [3]. Audio-based methods provide good performance in predicting genre labels [4], whereas mood prediction has remained more elusive in general [5]. The inherent link between genre and mood may provide a way of improving the audio-based music mood prediction. It has been shown that genre-specific mood prediction models, i.e., models trained on a set of tracks from a particular genre, give more accurate predictions within the corresponding genre than across genres [6]. Moreover, genre-specific mood auto-tagging models trained on tracks drawn from the corresponding genre have been shown to perform better than a general model trained on tracks from all genres [7].

Genre is a useful and popular way of describing musical content in most everyday contexts (music retailers, radio channels, libraries) and also forms one of the most important categories of social tags in online services for music consumption such as Last.fm[1], accounting for nearly 70% of the tags [8]. Genres are also easily identified from extremely brief excerpts [9] and yet genres are generally difficult to define in an unambiguous fashion [10] since they are partly cultural constructs and thus not purely defined by their musical properties.

Psychological studies have shown that music can be organized according to expressed or elicited emotions[2], and the ability of music to convey and affect moods is an important factor explaining why people are attracted by music [11]. For these reasons, developing reliable and scalable automatic mood prediction models has great potential. Audio-based models are particularly beneficial since they do not rely on pre-existing existing annotations of novel tracks. Most audio-based mood prediction approaches have utilized either categorical models (e.g., happiness, sadness and anger) [12], [13] or dimensional models of emotion [14], [15], [16]. A well-known example of the latter is the affective circumplex model [17], that represents different moods in the underlying dimensions of valence, reflecting positive vs. negative emotions, and arousal, relating to the activity or intensity of emotion. These dimensions, as well as tension, have been described as *core affects* [18]. Valence and arousal have been considered as independent dimensions, whereas tension has been inferred as the product of negative valence and positive arousal [19]. A number of studies have employed the dimensional model to audio-based music mood prediction. Eerola et. al [15] employed Partial Least Squares (PLS) regression to predict core affects in a set of film soundtracks clearly expressing distinct basic emotions, and achieved high performance for all affects

- *P. Saari is with the Department of Music, University of Jyväskylä, 40 014 Jyväskylä, Finland. E-mail: pasi.saari@jyu.fi*
- *G. Fazekas, M. Barthet and M. Sandler are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, E1 4NS London, U.K. E-mail: g.fazekas@qmul.ac.uk; m.barthet@qmul.ac.uk; mark.sandler@qmul.ac.uk*
- *T. Eerola is with the Department of Music, Durham University, DH1 3RL Durham, U.K. E-mail: tuomas.eerola@durham.ac.uk*
- *O. Lartillot is with the Department for Architecture, Design and Media Technology, Aalborg University, Rendsburggade 14, DK-9000 Aalborg, Denmark. E-mail: olartillot@gmail.com.*

Manuscript received Xxxx 00, 0000; revised Xxxxxx 00, 0000.

1. http://www.last.fm/home
2. We employ the words emotion and mood interchangeably in the present paper.

($R^2 = 0.70, 0.77, 0.71$ for valence, arousal and tension, respectively). However, performance levels have changed drastically when dealing with music drawn from multiple genres. Sufficiently high performance have still been reported for arousal ($R^2 = 0.58$ [14] and $R^2 = 0.71$ [20]), but the performance for valence has been far less satisfying (e.g., $R^2 = 0.21$ [14] and $R^2 = 0.32$ [20]) [21]. This suggests that further effort is needed to propose models that operate across genres in a robust manner for both valence and arousal.

In the previous studies, the relationship between genre and mood has only been broadly described. For instance, [22] statistically tested the significance of associations between genre and mood tags drawn from AllMusic[3] web service. The results indicated that while all genres can be characterized by certain moods, mood content also varies within genres to a large degree. Therefore, these two domains can potentially be used in a complementary fashion. Along these lines, Lin et al. [23] trained audio-based mood prediction models separately within different genres, and predicted mood successfully by a genre-weighted combination of the model outputs.

The present study offers an analytical exploration of genre-adaptive music mood prediction. Large social tag data and associated audio tracks are exploited in the analysis, and the prediction models are evaluated on a separate corpus of tracks from various popular music genres, annotated according to the core affects and mood terms in a listening test. The dimensional model of emotion is inferred from tags using the Affective Circumplex Transformation (ACT) technique deemed successful in [24], and the Semantic Layer Projection (SLP) technique proposed in [25] and [20] is employed to map audio features to the semantic space. Several techniques that exploit genre information in mood prediction are compared, and a novel technique is proposed that adapts both the semantic space of moods and audio-based models to different genres. Performance gains of the novel technique is compared to the other genre-adaptive techniques and a general SLP model that does not exploit genre information.

The following section discusses how this work relates to the previous studies in MIR. Section 3 describes the data covered in the study, while Sections 4 and 5 delineate the techniques to represent mood and genre of music tracks. Section 6 details the used audio-based mood and genre prediction techniques, and Section 7 introduces the genre-adaptive mood prediction techniques. Finally, Sections 8 and 9 report the results and conclude the paper.

2 Related Work

2.1 Tag- and Audio-based Mood Prediction via Semantic Computing

Social tags are free-form labels collaboratively applied to content by a online user communities. Due to the free-form nature, social tagging produces rich source of information about large sets of content items, but exhibits problems related to user errors, subjectivity, synonymy, polysemy and sparsity [26]. Several techniques have been applied to tags to successfully alleviate these problems and to infer useful semantic information only scarcely represented by raw tag data. In particular, the Latent Semantic Analysis (LSA) [27] has been used to infer low-dimensional semantic spaces from the co-occurrences of tag pairs within content items. Semantic computing akin to the LSA has been applied to music tag data in several studies [24], [28], [29], [30]. In the paper the most relevant to the present study, several techniques to infer semantic space of musical mood were compared [24]. The proposed technique called the Affective Circumplex Transformation (ACT) based on the LSA and the dimensional model of emotion outperformed typical semantic analysis techniques and the use of raw tags to represent music mood. The training data consisted of 357 unique track-level mood tags associated to 260k tracks from Last.fm, and the techniques were evaluated on listener ratings of moods in a separate set of 600 tracks from various popular music genres. Correlations between the track-level mood estimates and ratings were $r = 0.58, 0.64, 0.62$ for the core affects valence, arousal and tension, and between $r = 0.50 < r < 0.60$ for seven mood terms. In a follow-up study, [31] applied ACT models trained in [24] on a separate set of curated editorial tags associated to 205 production music tracks. Vice versa, ACT models were trained on 250,000 production music tracks and their performance was estimated on the aforementioned 600 popular music tracks. High performance rates obtained in this evaluation across music corpora and annotation types demonstrated the generalizability of ACT at capturing mood information from tag data.

In the studies above, the dimensional Valence and Arousal (VA) space was inferred with the ACT by conforming the semantic mood space with reference VA-configuration of mood terms given in [17], [32]. However, other reference configurations are equally promising, such as the affective norms in [33] related to a large set of English lemmas, or even ad-hoc subsets of these configurations specifically tuned for music data. We will also examine these options in this paper.

The Semantic Layer Projection technique (SLP) was proposed in [25] as an extension to ACT to enhance audio-based music mood prediction. The SLP involves projecting tracks to moods via a two-stage process. In the first stage, a corpus of tracks is mapped to the VA-space obtained with the ACT based on associated tags, and regression models are trained between audio features and the mood space dimensions. In the second stage, track positions in the mood space are projected to infer the core affects and mood term estimates. In an attempt to predict listener ratings of mood, variants of SLP trained on Last.fm data outperformed regression techniques trained directly on the ratings. In a subsequent study [20] the SLP outperformed baseline models that excluded the ACT-based projection in the first stage, which highlighted the robustness of the technique.

2.2 Musical Genre Representation

Genre is the most widely used semantic facet to describe music, as almost all large music libraries and record shops categorize music into genres. A common debate between music enthusiasts is what genre an artist or track represents. Several studies have highlighted the challenges of

3. http://www.allmusic.com/

representing the genre of music, relating in particular to the optimal resolution of genres [34] and the fuzziness of genre categories [35]. In studies attempting to identify the underlying factors of music preferences based on genres, four [36] and five [37] underlying factors have typically been singled out. In other studies, 10 [38], 13 [39] 14 [40] and 16 [41] genres have been employed to characterize the typical range of music. On the other hand, the fuzziness of genre categories has been highlighted with artist tags at Last.fm [35]. For instance, 56% of artists tagged with 'pop' were also tagged with 'rock', and 87% of 'alternative' music overlaps with 'rock'. Still these three genres are considered as separate categories in typical music catalogues, such as iTunes. The evidence from social tags indicates that a single genre describing a track is not inclusive enough, but perhaps a (weighted) combination of several genre labels would better describe genre information.

2.3 Contextual Information in Music Mood Prediction

The aim of the present study is to exploit genre as context in music mood prediction. Context is typically associated with different music-listening situations [2], whereas the MIR-oriented view relates context to all types of information about music, not represented by the content, i.e., the audio [3]. Context therefore includes multimodal information such as lyrics, album covers, music videos, tags and other online material. However, in music auto-tagging [42], audio-based prediction of tags has not automatically been described as contextual. Instead, contextual techniques have typically involved two-stage approaches that exploit the relationships between individual tags [43], [44]. As an example of a two-stage technique, [43] trained Support Vector Machine (SVM) models with probabilistic outputs in two stages. Audio-based SVMs were first trained for each tag. Then, stacked probabilistic outputs for all tags was used as input to second-stage SVMs, thus exploiting correlations between tags. Contextual techniques have systematically outperformed non-contextual models, and in particular, stacked SVMs have yielded state-of-the-art performance[4]. As the set of tags in these studies typically includes both genres and moods, genre has been implicitly been incorporated into mood tag prediction, but this aspect has not been investigated further in previous research.

Genre has been given emphasis in only few music mood prediction analyses in the past. Schuller et al. [41] compared genre to other sources of information for music mood classification. In other studies by Lin et. al [7], [23], a two-stage approach to audio-based mood prediction was proposed. In [7], several genre-specific mood auto-tagging models were trained, and for an unknown track, the model corresponding to the track genre was applied. This approach was taken further in [23], where instead of inferring the genre for an unknown track from available metadata, also genre was predicted from audio. The present study follows the footsteps of the above studies and finally increases the genre-adaptivity of mood prediction by allowing the semantic mood space to adapt to different genres. This is challenging, since inferring a taxonomy for both mood and genre is nontrivial. Moreover, the semantic link between mood and genre makes it difficult to devise a protocol for combining these semantic facets. The next sections presents ways to deal with these issues.

4. http://www.music-ir.org/mirex/wiki/2012:MIREX2012_Results

TABLE 1
Data sets used in the study.

			nterms per track (avg.)	
	ntracks	*nartists*	Mood	Genre
TRAIN100k	118,874	5,470	3.23	5.38
TRAIN10k	10,199	5,470	3.56	5.52
TEST600	600	600	7.22	8.53

3 Data Collection

In this section we explain how we obtained our data sets comprising mood- and genre-related social tags and associated audio tracks from the I Like Music (ILM) library, and how we filtered and sampled the tracks and tags to produce the main training sets TRAIN100k dedicated for semantic computing and TRAIN10k dedicated for audio-based modeling. We also introduce the TEST600, a set of tracks reserved for model evaluation. Table 1 summarizes the statistics of the data sets.

3.1 Training Data Sets

We used the social tag data collected in [24] using the Last.fm API as a starting point for further analyses. This data consists of 1,338,463 tracks and 924,230 unique tags after lemmatizing the tags and removing non-alphanumeric characters. Each track-tag association in the data is represented by normalized count spanning from weak to strong. Mood- and genre-related tags were identified by string matching against 560 mood and 865 genre terms gathered from various sources including the AllMusic.com service and several studies in affective sciences [24]. For moods, each tag that included a term as a separate word was associated to the corresponding term, whereas genre terms were matched with the full tags. This approach was chosen since genre is more prevalent tag category than mood. In the case when several tags associated to a track matched the same mood term, the association with the highest normalized count was retained. The obtained set was further filtered by sorting mood and genre terms according to the number of associated tracks and kept the top 100 mood and top 100 genre terms. Tracks associated to both mood and genre terms, and from artists not appearing in TEST600, were included in further analysis.

TRAIN10k including audio for 10,199 full tracks was subsampled from the corpus. We obtained audio for these tracks using exclusive access to the I Like Music (ILM) catalogue. Tracks were first paired with the catalogue using exact string matching between the artist names and song titles. From the initial corpus, we found 218,000 tracks in the ILM database that matched one of the Last.fm tracks. However, this dataset included a large proportion of duplicates. We then applied controlled track sampling to arrive to the tracks used in subsequent analyses.

Track sampling was done by following several potentially conflicting criteria: First, including tracks with close

matches between metadata entries (artist name, song title, album title) in Last.fm and ILM based on Levenshtein string distance and less than 0.5s difference between reported track durations. Second, balancing the set according to broad genre categories following expert classification available from ILM. Third, limiting the maximum number of songs sampled from the same artist. Finally, mood-balancing the tracks within each genre based on track positions in three-dimensional mood space obtained by ACT model trained in [24]. The resulting TRAIN10k includes tracks from 5,470 unique artists.

Since subsampling excluded most of semantic information included in the original set of tracks, we formed the larger training set TRAIN100k for semantic computation purposes. This was done by augmenting TRAIN10k with all the tracks in the corpus that were performed by any artists in TRAIN10k. This produced a set of 118,847 tracks.

As seen in Table 1, the average number of terms associated to each track is higher for genres than for moods, even after different string matching criteria between tags and terms. This obviously reflects the overall higher prevalence of genre tags than mood tags in social tag data. Within TRAIN10k (TRAIN100k, respectively), a median of 162 (1,687) tracks is associated to a mood term, and of 329 (3,869) to a genre term. The most prevalent mood terms are Chillout (2,569), Party (1,638) and Mellow (1,569), whereas the least prevalent mood terms are Pleasant (51), Bliss (51) and Spooky (52). The most prevalent genre terms are Rock (3,587), Pop (3,091) and Alternative (2,047), whereas the least prevalent are Root reggae (147), Jazz fusion (150) and Electropop (153). The relative term prevalences are roughly the same within TRAIN100k.

For experimental evaluations, 10 training partitions, each comprising 80% of tracks in TRAIN100k and TRAIN10k, were randomly subsampled from the training data. Evaluations were carried out by performing the model training separately on each training partition and applying the models thus obtained on the full TEST600 set. We will denote the partitions as T (within TRAIN100k) and T' (within TRAIN10k), $T' \subset T$.

3.2 Test Data Set

The TEST600 data set reserved for evaluation is the same as the one described in [24]. The set consists of Last.fm tags, listener ratings of mood, and audio for 600 tracks by unique artists. The tracks represent six main genres: *Metal*, *Rock*, *Folk*, *Jazz*, *Electronic* and *Pop*. The listener ratings of the expressed mood, given in nine-step Likert scales, were averaged from 59 participants. Three core affects (Valence, Arousal and Tension) and seven mod terms (Atmospheric, Happy, Dark, Sad, Angry, Sensual and Sentimental) were evaluated. A more detailed description of TEST600 can be found in the original publication [24]. Although the listener ratings were given for 15 second clips, we use the full tracks in the present study, relying on the claim made in [24] that the ratings describe the tracks well overall. The ratings and links to the audio and tag data is publicly available[5].

The tag data associated to TEST600 was subjected to the same process applied to the training corpora, by associating each track to the 100 mood and 100 genre terms, and excluding tracks not associated to any mood or genre term. As a result, 12 tracks were excluded.

3.3 Audio Features

62 audio features related to dynamics, rhythm, harmony and timbre, were extracted from the full-length tracks of TRAIN10k and TEST600 using the MIRtoolbox[6] [45]. Table 2 summarizes the feature set. The features were aggregated over time to generate song-level descriptors. Finally, each song was represented by a 178-dimensional feature vector.

The audio material was first converted to mono and cut into overlapping analysis frames with feature-specific lengths and degrees of overlap. A frame length of 50ms with 50% overlap was used for the low-level spectral features, MFCCs and their 1st (Δ) and 2nd-order ($\Delta\Delta$) instantaneous derivatives, and for features related to dynamics. Audio onsets were detected from the temporal amplitude curves extracted from 10-channel filterbank decomposition. *Event Density* was calculated by taking the number of onsets in 10s, 50% overlapping frames. Features related to meter were derived from the onsets using the autocorrelation function with 3s frames with 90% overlap (33.3% overlap for *Tempo*). Furthermore, chromagrams were computed from 750ms, 50% overlapping frames, from which several high-level features related to tonality, such as *Mode* (majorness) and *Key Clarity* were calculated.

All features with different frame lengths where brought to the same time granularity by computing the Mean (*m*) and Standard deviation (*s*) over 1s, 50% overlapping texture windows. However, only the Mean was computed for *Event Density* and chromagram-related features, since they were extracted from longer frames to begin with. This was also done for the MFCC derivatives, since their Means already describe temporal change. Finally, 178 song-level descriptors were computed by taking the Mean (*mm* and *ms*) and Standard deviation (*sm* and *ss*) of these features over song lengths. This process is motivated by the approach presented in [43]. Typically the song-level representation of audio features is obtained by taking the Mean and Standard deviation over the whole song length without first aggregating within the texture windows. The approach taken here incorporates the temporal dynamics of features at both short and long time span in more sensitive fashion than the typical song-level averaging approach.

4 Mood Representation from Tag Data

The ACT technique [24] was applied on the training partitions of TRAIN100k set to enable representing the mood of the tracks based on associated tags. Initially, associations between mood terms and tracks are represented in a standard Vector Space Model (VSM) matrix form.

4.1 ACT Training

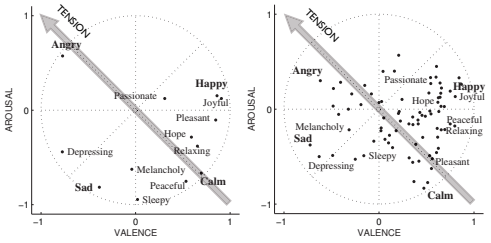
As in [24], M is first normalized by computing Term Frequency-Inverse Document Frequency (TF-IDF) scores, and then transformed to a three dimensional mood space by

5. http://hdl.handle.net/1902.1/21618

6. MIRtoolbox version 1.5.

TABLE 2
Audio features, aggregated to $*$: *mm*, *ms*, *sm* and *ss*; †: *mm* and *sm*. The "Frame" column reports the window lengths and overlaps ($\star$: 50ms length with 50% overlap).

Category	Feature	Stats	Frame
Dynamics	RMS, Zerocrossing rate	$*$	$\star$
Onsets	Attack (time, slope, leap)	$*$	Onset-based
	Event density	†	10s, 50%
Autocorrelation	Pulse clarity, Novelty	†	3s, 90%
	Tempo	†	3s, 33.3%
Chromagram	Mode, HCDF, Key Clarity, Centroid, Novelty	†	750ms, 50%
Spectrum	Novelty, Brightness, Centroid, Spread, Flux, Skewness, Entropy, Flatness, Roughness	$*$	$\star$
	13 coef. MFCC, Δ, $\Delta\Delta$	$*$	$\star$



(a) Russell (**Russell4**) (b) Norms

Fig. 1. Reference mood term configurations from a) Russell [17] and Scherer [32] and b) Affective norms [33].

applying the non-metric Multi-Dimensional Scaling (MDS). Notice, that in [24] dimension reduction was employed in two stages by applying Singular Value Decomposition (SVD) prior to MDS, which gave slight performance improvement compared to the dimension reduction with MDS only. The SVD stage is excluded in the present study to reduce the number of alternative model parameterizations (e.g., the number of dimensions in SVD). The next stage of the ACT process conforms the mood space to a reference configuration of mood terms in a VA-space. Such configuration can be extracted for example from Russell's and Scherer's studies [17], [32] similar to the approach presented in [24]. This is done via Procrustes transformation [46] which performs a linear mapping from a space to another, while retaining the relative distances between objects in the original space. This stage yields a configuration $\tilde{X}$ of all mood terms in the VA-space.

A track represented by a VSM vector q_i of mood term associations is projected to the VA-space by the following operations: First q_i is normalized to $\hat{q}_i$ according to the learned TF-IDF scores, and $\hat{q}_i$ is projected to the VA-space by computing the centre-of-mass:

$$\tilde{q} = \frac{\sum_i \hat{q}_i \tilde{x}_i}{\sum_i \hat{q}_i}. \quad (1)$$

Consequently, the estimates for Valence and Arousal for $\tilde{q}$ are inferred directly from the track positions along the first and second dimensions, whereas the weights related to mood terms i are computed by

$$\frac{\tilde{x}_i}{|\tilde{x}_i|_{L2}} \cdot \tilde{q}. \quad (2)$$

Moreover, the estimate for Tension is obtained by projecting the track positions along the direction $(-1, 1, 0)$ corresponding to negative valence–positive arousal, as suggested by Thayer [19].

4.2 Alternative Configurations for ACT

Of all 101 mood terms present in Russell's and Scherer's reference configuration [17], [32], 13 terms could be matched with the 100 mood terms used in the present study. These terms are plotted onto the VA-space in Fig. 1a. We denote this configuration by *Russell*. Most of the matched terms are located in the low Arousal – high Valence quadrant. Due to this imbalance, we formed a more simple alternative reference configuration by including only one mood term for each VA-quadrant: Happy, Calm, Sad and Angry, indicated in boldface in Fig. 1a. This configuration is denoted by *Russell4*. These terms were chosen since they are frequently cited in the past music and emotion research [47] and their prevalence within TRAIN100k was above the median (*10,459*, *3,554*, *9,306* and *1,921* for Happy, Calm, Sad and Angry, respectively).

We also explored how well the norm data from [33] would perform as a direct alternative to the mood term positions inferred using the ACT. For this model the MDS and Procrustes stages in the ACT were skipped. In addition to Valence and Arousal, the data includes Dominance as the third dimension. 81 mood terms, summarized in Fig. 1b, could be matched between the norm data and tags. We will refer to this model as *Norms*. 339 tracks had to be excluded from TRAIN10k when training the model, since those tracks were not associated to any of the matched mood terms.

5 Genre Representation from Tag Data

In order to exploit genre information as a context in mood prediction, we sought for a concise genre representation. Many of the genre terms present in TRAIN100k are similar; for instance 'alternative' with 'indie' and 'electronic' with 'dance'. We therefore applied genre term clustering to reduce the number of distinct genres.

5.1 Genre Clustering Techniques

Formally, given the genre tags in TRAIN100k expressed as a VSM matrix $G = g_{i,j}$, where $j \in T$ (tracks in a training partition of TRAIN100k) and i relates to the i:th genre term, the rows g_i are grouped into clusters $C = \{C_1, C_2, ..., C_K\}$ such that $C_k \bigcap C_l = \emptyset$ and $\bigcup_{k=1}^{K} C_k = G$. The clustering techniques were used with the following specifications:

K-means: G was first normalized to a unit Euclidean length by

$$\hat{g}_{i,j} = g_{i,j} / (\sum_{j \in T} {g_{i,j}}^2)^{1/2}, \quad (3)$$

after which the algorithm was run using the cosine distance.

Agglomerative hierarchical clustering: G is first normalized according to the TF-IDF to produce $\hat{G}$. The cosine distance between $\hat{g}_i$ and $\hat{g}_{i'}$ was then used as the distance measure, and the agglomeration was done based on the average link criterion.

Spectral clustering: G was normalized according to the TF-IDF. The affinity matrix used was the cosine similarities (1 – cosine distance) between $\hat{g}_i$ and $\hat{g}_{i'}$. Clustering was then done following the method described in [48], similar to [23] where the technique was applied to emotion tags.

5.2 Genre Clustering Survey and Evaluation

To assess the quality of the obtained genre clusterings, we arranged an online genre grouping survey. This also provided insight into deciding how many genre clusters would optimally represent the data. The task in the survey was to organize the 100 genre terms into any number of clusters between 2-16 considered most appropriate by each participant. The range for the candidate number of genres was selected to acknowledge the typical number of genres assessed in past studies. It was specified in the instructions that the clusters should share characteristics based on musical, social, cultural or other factors. The participants were asked to be objective in their assignments, and the instructions allowed using external web resources for checking unfamiliar genre terms. 19 participants, predominantly engineering and musicology students knowledgeable of different music genres, took part in the survey.

No clear optimal number of genre clusters arose from the survey results. The number of clusters ranged between 6 and 16, with a peaks around 9, 10 and 16 clusters (M $=$ 11.34, SD $=$ 3.17). To question this finding, the conventional Davies-Bouldin technique [49] was applied on the genre tag data to infer the optimal number of clusters, but this analysis did not yield a clear optimum either. This may reflect the more general difficulty of defining the genre granularity that would satisfy all purposes.

We computed genre clusterings with the three aforementioned techniques separately on the training partitions of TRAIN100k using $K = \{2, 4, 6, ..., 16\}$ clusters. These clusterings were compared to those obtained from the survey using the Mirkin metric [50], which can be used to assess the disagreement between two clusterings $C = \{C_1, C_2, ..., C_K\}$ and $C' = \{C'_1, C'_2, ..., C'_{K'}\}$ by:

$$d_M(C, C') = \sum_k n_k^2 + \sum_{k'} n_{k'}^2 - 2 \sum_k \sum_{k'} n_{kk'}^2, \quad (4)$$

where n and n_k are the numbers of genre terms in G and cluster C_k, respectively, and $n_{kk'}$ is the number of terms in $C_k \bigcap C'_{k'}$. This metric can be used to compare clusterings with different K. For identical clusterings $d_M = 0$ and $d_M > 0$ otherwise. We computed d_M separately between the tag-based clusterings and the 19 clusterings obtained from the survey, and averaged it across the participants and training partitions. The results are shown in Fig. 2. One can see that all clustering techniques compare similarly to the survey data, except that the hierarchical clustering performed poorly with $K = 2$. In general, K-means outperformed the other techniques by a slight margin. Therefore we will use K-means in further analyses.

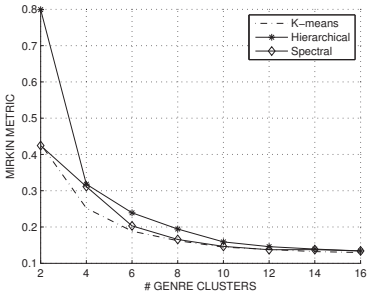


Fig. 2. Mirkin metric of each genre clustering technique averaged over participants and training partitions.

TABLE 3
Genre clusters obtained using K-means with $K = \{2, 4, 6, ..., 16\}$.

K	Most prevalent genre term
2	Pop, Rock
4	Soul, Rock, Hard rock, Electronic
6	Hard rock, Singer songwriter, Electronic, Jazz, Rock, Pop
8	Electronic, Rnb, Soul, Instrumental, Pop, Singer songwriter, Rock, Hard rock
10	Soul, Hip hop, Rock, Electronic, Singer songwriter, Reggae, Alternative, Jazz, Metal, Lounge
12	Electronic, Downtempo, Country, Soul, Hard rock, Punk, Rnb, Singer songwriter, Rock, Jazz, Classic rock, Pop
14	Hip hop, Rock, Singer songwriter, Pop, Pop rock, Jazz, Country, Soul, Metal, New wave, Hard rock, Classic rock, Instrumental, Electronic
16	Jazz, Rnb, Instrumental, Reggae, Ambient, Pop rock, Rock n roll, Experimental, New wave, Classic rock, Pop, Soul, Electronic, Hard rock, Singer songwriter, Rock

As the survey did not give clear indication of the optimal number of genre clusters, the subsequent analyses will primarily be conducted with $K = 6$. This was the minimum number obtained in the survey, but the main reason for the choice was that it corresponds well with TEST600 that was sampled to contain six genres: *Metal*, *Rock*, *Folk*, *Jazz*, *Electronic*, and *Pop*.

In order to be thorough about the genre clustering results, we computed the final clusterings with $K = \{2, 4, ..., 16\}$ based on the full TRAIN100k set. Fig. 3 shows in detail the discrepancy between the clustering with $K = 6$ and the survey data. We computed for each pair of genre terms the of number participants that assigned both terms to the same cluster. Six genre terms most prevalent in the TRAIN100k are shown for each cluster in the order of prevalence. The six clusters correspond well with the main genres in the TEST600 set since each of these terms are in different clusters. Therefore we label the clusters with these genres. One can see from the figure that genre terms in *Metal*, *Folk* and *Electronic* were mostly grouped together also by participants, whereas terms in *Jazz*, *Rock* and *Pop* were not as consistently grouped.

Table 3 shows the most prevalent genre tag for the genre clusters obtained with different K.

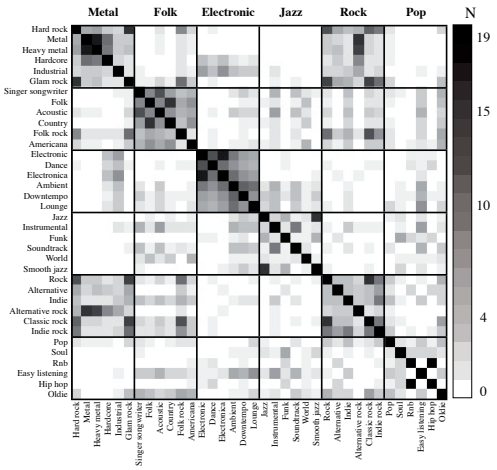


Fig. 3. Discrepancy between the six genre clusters obtained with K-means, and cluster co-occurrences of genre terms obtained from the survey.

TABLE 4
The percentage of tracks in the data sets belonging to each of the six genre clusters.

Genre cluster	TRAIN100k	TRAIN10k	TEST600
Metal	18.2	14.0	27.0
Folk	28.7	34.0	40.3
Electronic	30.3	31.8	46.2
Jazz	33.6	41.3	44.0
Rock	60.1	55.3	80.2
Pop	49.2	57.1	66.2

5.3 Representing the Genre of a Track

The genre of a track j, given its associated genre tags and a clustering C, is represented by a weighted combination $H = h_{k,j}$ ($k \in \{1, 2, ..., K\}$) of the associated genre clusters:

$$h_{k,j} = \frac{\sum_{i \in C_k} \hat{g}_{i,j}}{n_k} \left[\sum_k \frac{\sum_{i \in C_k} \hat{g}_{i,j}}{n_k} \right]^{-1}, \quad (5)$$

where $\hat{g}_{i,j}$ is computed with (3) based on the full TRAIN100k set.

We also consider that a track j *belongs to a genre cluster* k, if any of its associated genre tags belong to the corresponding cluster, i.e., if $h_{k,j} > 0$. This "hard" assignment will later be used to sample the training tracks for genre-adaptive mood prediction models. Note, that following this definition, a track can belong to more than one genre cluster. Table 4 shows the percentage of tracks in in the data sets belonging to each genre cluster. It can be seen that the clusters are very broad, since 80.2% and 66.2% of TEST600 tracks belong to *Rock* and *Pop*, respectively. High prevalence of tags related to *Pop* and *Rock* reflects the fuzziness of these genres, discussed in [35].

6 Audio-based Prediction of Mood and Genre

We use the SLP technique as the main framework for audio-based music mood prediction. This section gives details on the general form of the SLP that predicts moods without exploiting genre information, whereas Section 7 describes how the SLP is used in genre-adaptive mood prediction. This section also introduces two baseline methods that are based on the stacked SVM auto-tagger, and describes our technique for audio-based genre prediction.

6.1 Mood Prediction with SLP

The SLP involves training a set of regression models to map audio features to the VA-space dimensions and to predict moods in music tracks. In [25] and [20] Partial Least-Squares (PLS) was employed as a regressor for the SLP, whereas in the present paper we use the LIBSVM implementation of Support Vector Regression (SVR) [51] to allow for more direct comparison to the baseline SVM auto-tagger.

The SLP receives as an input the tag data associated to the training partitions of TRAIN100k and TRAIN10k and audio features extracted for the training partition of TRAIN10k. The following stages are then applied:

Audio feature pre-processing: All features are z-score-transformed to a zero mean and unit standard deviation, and extreme feature values are considered outliers and truncated to the extremes $[-5, 5]$. After this, highly correlated audio features are removed to reduce the SVR training time. This is done by agglomerative hierarchical clustering with correlation distance function. Clustering is done using the complete linkage criterion with a cutoff correlation distance of 0.1. The first feature of each obtained cluster in the list presented in Table 2 is kept.

Projecting tracks to VA-space: As in Section 4, mood term configuration in VA-space is learned, and tracks in T' expressed in the VSM form are projected to the VA-space based on the associated tags.

Regressor training: Mappings from the pre-processed audio feature set to each VA-space dimension separately are trained with the SVR. In a preliminary analysis we tested the SVR with linear and Radial Basis Function (RBF) kernels, and found that the linear kernel produced results comparable to the RBF with a shorter training time. We also found that setting the cost parameter c to 0.001 gave consistently high performance compared to several candidates $c = 10^y, y = [-4, -3, ..., 1]$.

SLP prediction: When applying SLP to a novel track, first the preprocessing is applied to the associated audio features, and the track is mapped to the VA-space using the learned regressors. Finally, the estimates for Valence and Arousal are represented by the first and second dimension of the space, whereas the estimates for Tension and mood terms are obtained by projecting the tracks along the corresponding directions, as described in Section 4.

6.2 Baseline Mood Prediction with SVM Auto-tagger

We employ the two-stage stacked SVMs [43] to compare the SLP performance to a conventional auto-tagger. Our implementation of the technique mainly follows that in [43].

First, the input audio features related to TRAIN10k partition are pre-processed as described in Section 6.1, and mood tags are transformed into binary classes. The first-stage SVM classifiers with linear kernel and probabilistic outputs are then trained separately for each term, and applied on the training tracks. Obtained positive class probabilities for all terms are then served as input for the second-stage SVM classifiers, that again map the input to the binary classes. When predicting moods of a novel track, the first- and second-stage models are applied consecutively, producing a vector of probability estimates which we finally normalize to sum to one. Notice, that the stacked SVMs are not capable of directly producing estimates for the core affects, since Valence, Arousal and Tension are not explicitly represented by any of the mood terms.

The binary tag data fed to the SVMs is highly imbalanced, as the term prevalence in TRAIN10k varies from 0.5% to 26%. In the past, taking into account the class imbalance has yielded positive results for SVM-based music mood auto-tagging [7]. Therefore we employed cost-sensitive learning found successful in [52] by setting different misclassification error cost for the positive and negative class for each term. We set a cost value $c_i^+ = 1$ for the positive class and a lower cost value $c_i^- = \frac{n_i^+}{n_i^-}$ for the negative class, where n_i^+ and n_i^- are the numbers of positive and negative tracks within the training data for a mood term i.

To form another baseline technique, we project tracks to the VA-space based on the probabilistic outputs of the second-stage SVMs by employing the ACT operations of TF-IDF-weighting, projection to VA-space (1), and mood weight estimation (2). This technique, as opposed to the original stacked SVM, inherently produces estimates for both the core affects and mood terms.

We refer to these two baseline techniques as *SVM-orig* and *SVM-ACT*. Similar baseline techniques were implemented also in [20] using PLS regression to predict the normalized tag counts, but using instead the stacked SVMs is more efficient, as the results will show.

6.3 Genre Prediction with SVM Auto-tagger

For audio-based genre prediction we train stacked SVMs on genre tags similar to *SVM-orig*. Given the audio features of a novel track, we first predict the vector of genre term probabilities, and then map the vector with (5) to genre clusters obtained by K-means.

7 Genre-Adaptive Mood Prediction Techniques

As the main contribution of the present paper, we compare several genre-adaptive mood prediction techniques. We use the SLP as the general framework for all of these techniques, and assess their efficiency compared to the general form of the SLP described in Section 6.1. The techniques either exploit genre by using genre information as input features to the SLP (*SLPg* and *SLPga* models), or train a collection of mood prediction models within different genres (*SLPwg* and *ACT+SLPwg* models). Moreover, we compare these techniques when genre information of a novel track is either derived from tag data or predicted from audio. When used with audio-based genre-prediction, these models can be applied to any track irrespective of the availability of tag data.

In the training phase the techniques receive as input the training partitions T and T', the mood VSM M, genre VSM G, audio features A, and a clustering of genre terms $C = \{C_1, C_2, ..., C_K\}$. Tag-based genre representation $h_{k,j}, j \in T$ is then computed with (5). In the prediction phase (for tracks in TEST600), tag-based genre representation is computed again with (5), whereas audio-based genre representation is computed as described in Section 6.3.

7.1 Genre-based Prediction (SLPg)

To obtain indication as to how much variance of mood ratings can be attributed to mood prevalence differences between genres, we predict the mood of a novel track with the *SLPg*, relying on the associated genres. The *SLPg* differs from the general SLP in that tracks in regressor training are represented by genres $h_{k,j} = (h_{1,j}, h_{2,j}, ..., h_{K,j}), j \in T'$ instead of audio features. All other stages are the same as in the general SLP.

7.2 Genre- and Audio-based Prediction (SLPga)

SLPga is similar to *SLPg*, but appends audio-feature vector to the genre-vector to form the input to the regressors. Input tracks are therefore represented by $(h_{1,j}, h_{2,j}, ..., h_{K,j}, a_{1,j}, a_{2,j}, ..., a_{nfeatures,j}), j \in T'$. By exploiting both audio-to-mood and genre-to-mood associations in parallel, the *SLPga* represents a potential alternative to the techniques introduced next.

7.3 Audio-based Modeling within Genres (SLPwg)

The *SLPwg* trains regression models from audio to VA-space separately on tracks belonging to each of the K genres, and combines the predictions obtained with each genre-specific model to form the final predictions. The assumption underlying *SLPwg* is that different audio features and feature combinations relate to mood differently within different genres. Audio feature pre-processing and regressor training (cf. Section 6.1) are employed separately for each genre, yielding K genre-specific models. Training set T'_k for a genre k is here defined as $T'_k = t'_{k,j}, \{j : j \in T' \wedge h_{k,j} > 0\}$.

The prediction stage is applied as follows: A novel track is represented by genres h'_k, either derived from tags or predicted from audio, and audio features. All genre-specific models are first applied to map the track to the VA-space based on audio features, as described in Section 6.1. This yields genre-specific VA-space positions $\tilde{q}'_k$. The final predictions are then obtained by weighting the genre-specific estimates proportionately to h'_k:

$$\frac{1}{\sum_k h'_k} \sum_k h'_k \frac{\tilde{x}_i}{|\tilde{x}_i|_{L2}} \cdot \tilde{q}'_k. \quad (6)$$

Combining genre-specific audio-based mood models this way bears analogy to the emotion classification model proposed in [23].

7.4 Semantic and Audio-based Modeling within Genres (ACT+SLPwg)

ACT+SLPwg combines genre-adaptive semantic computing of moods with genre-adaptive regressor training. The underlying assumption is that both the semantic relationships between mood terms, and audio-to-mood associations vary between genres. *ACT+SLPwg* is similar to *SLPwg*, except that the VA-space is formed with the ACT separately for each genre, i.e., on sets T_k within TRAIN100k, $T_k = t_{k,j}, \{j : j \in T \wedge h_{k,j} > 0\}$. This produces K genre-specific VA-spaces $\tilde{x}_i^k$. Regressor training stage is employed separately for each genre, producing mappings from audio to the corresponding genre-specific VA-spaces. The final predictions are obtained by

$$\frac{1}{\sum_k h'_k} \sum_k h'_k \frac{\tilde{x}_i^k}{|\tilde{x}_i^k|_{L2}} \cdot \tilde{q}'_k. \quad (7)$$

Genre-adaptive semantic computing combined with audio-based prediction of moods has not been examined in the past.

8 Results and Discussion

The squared correlation coefficient (multiplied by the sign of the correlation) between the predictions and the listener ratings of mood is used as the evaluation metric. This corresponds to the coefficient of determination statistic (R^2). Each technique is trained on each of the 10 training partitions of TRAIN100k, and correspondingly TRAIN10k, and tested on TEST600. Median and median absolute deviation (MAD) across the partitions are reported.

8.1 Mood Prediction without Genre Information

First, we compare the performance of ACT using different mood term configurations. The most succesful configuration is then chosen for subsequent audio-based prediction analysis using SLP and the stacked SVMs.

8.1.1 Tag-based Prediction with Mood Term Configurations

The results for ACT obtained with *Russell*, *Russell4* and *Norms* configurations are shown in Table 5. In general, core affects were more easy to predict than mood terms, and the performance for Valence was lower than for Arousal. These findings are in line with those from past work evaluating ACT with TEST600 data [20], [24]. *Russell4* gave the highest performance for seven out of ten scales, and was clearly more efficient than *Russell* for Dark and Sad. These mood terms were also among the most difficult to predict. On the other hand, *Russell* was more successful at predicting Valence, Tension and Atmospheric. *Norms* yielded dramatically lower performance than the other configurations. This arguably supports exploiting music domain-specific tag data in forming the semantic mood space rather than using a mood space that describes affective connotations of mood words in general. Moreover, this indicates that the inclusion of Dominance as the explicit third dimension in the mood space does not provide evident benefits. When examining the average performance across mood scales, *Russell4* ($R^2 = 0.387$) outperformed Russell slightly ($R^2 = 0.371$). This suggests that ACT is not overly sensitive

TABLE 5
Prediction results for ACT with *Russell* and *Russel4* reference configurations and Norms.

		Russell	*Russell4*	*Norms*
Core affects	Valence	**0.425** 0.013	0.413 0.015	0.270 0.000
	Arousal	0.477 0.004	**0.486** 0.003	0.269 0.000
	Tension	**0.382** 0.015	0.378 0.014	0.219 0.001
Mood terms	Atmospheric	**0.424** 0.039	0.395 0.026	0.157 0.000
	Happy	0.384 0.016	**0.386** 0.011	0.300 0.000
	Dark	0.274 0.087	**0.348** 0.035	0.038 0.000
	Sad	0.201 0.015	**0.276** 0.013	0.166 0.000
	Angry	0.522 0.013	**0.531** 0.017	0.214 0.000
	Sensual	0.403 0.006	**0.416** 0.007	0.002 0.000
	Sentimental	0.220 0.020	**0.238** 0.023	0.061 0.000
Average		0.371	0.387	0.170

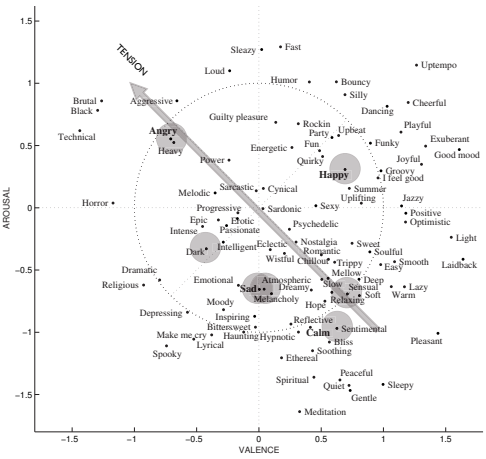


Fig. 4. Mood tag positions (the mean across training partitions) obtained with ACT using *Russell4* as the reference configuration.

to changes in the mood reference configuration, and that a simple reference configuration provides a strong enough reference to reliably represent mood terms in the VA-space. Therefore, *Russell4* is chosen for the subsequent audio-based analyses.

The AV-space obtained with the ACT using *Russell4* is shown in Fig. 4. The mood positions in the figure are computed as the average of those obtained for each training partition. The underlying dimensions of Valence and Arousal are easily distinguishable, and the obtained positions for the four reference terms correspond fairly well with the original positions. However, it is notable that Sad is close to neutral along the Valence dimension. This finding is in line with [53], where sadness was not found to indicate negative Valence in the context of music.

8.1.2 Audio-based Prediction

Table 6 shows the performance obtained with SLP (using *Russell4*) and the stacked SVMs. It is notable that SLP provided dramatically higher performance than ACT for all mood scales except Valence, Happy and Dark. The clearest

TABLE 6
Prediction results for SLP and SVM baseline techniques.

		SLP	SVM-orig	SVM-ACT
Core affects	Valence	0.359 0.019	–	**0.369** 0.020
	Arousal	**0.728** 0.004	–	0.714 0.005
	Tension	**0.485** 0.019	–	0.483 0.022
Mood terms	Atmospheric	**0.696** 0.014	0.069 0.004	0.684 0.020
	Happy	0.312 0.030	0.205 0.003	**0.314** 0.031
	Dark	0.235 0.023	**0.311** 0.004	0.248 0.020
	Sad	0.303 0.011	0.316 0.007	**0.323** 0.007
	Angry	0.589 0.016	**0.622** 0.008	0.618 0.017
	Sensual	**0.544** 0.004	0.252 0.026	0.535 0.010
	Sentimental	0.300 0.024	**0.436** 0.016	0.304 0.030
Average		0.455	0.316	0.459

difference between SLP and ACT was obtained for Arousal ($R^2 = 0.728$ vs. 0.477). This rather surprising result, congruent with that reported in [20], may be explained by the sparsity and inherent unreliability of tag data. The ACT maps tracks to the mood space based on only few tags, which may cause local inconsistencies. By contrast, mapping audio features to the mood dimensions using the SLP may tap into more global patterns and provide a way to "smooth out" these inconsistencies. The mean SLP performance across mood scales was similar to that reported in [20] using a larger training set ($R^2 = 0.455$ vs. 0.453). However, the prediction performance for Valence was clearly higher in the present study, $R^2 = 0.359$ compared to $R^2 = 0.322$, while the performance for Arousal was approximately the same: $R^2 = 0.728$ vs. $R^2 = 0.710$.

The *SVM-orig* auto-tagger performed inconsistently. The low performance for Atmospheric, Happy and Sensual suggests that the way these tags are applied in Last.fm is not well-explained in terms of musical characteristics, and that these terms may be better modeled by musical characteristics associated to more general patterns in the tag data evident in the low-dimensional mood space. The results of the *SVM-ACT*, obtained using *Russell4*, increased the overall performance of the *SVM-orig* and provided performance comparable to SLP. This shows that mapping audio to the semantic mood space provides similar results when carried out after training tag-specific classification models or before training dimension-specific regression models. However, the regression approach is much less computationally intensive, as it requires training only one model for each dimension. Therefore, we consider SLP the most promising technique as the basis for further genre-adaptive analyses.

8.2 Genre-adaptive Techniques

To evaluate the genre-adaptive techniques the performance obtained using the *ACT+SLPwg* is compared to those of the other techniques exploiting genre information and to the general SLP baseline. First, prediction rates are reported for the audio-based genre prediction using the SVM auto-tagger.

8.2.1 Audio-based Genre Prediction

The performance of the audio-based genre prediction was assessed by comparing the predictions to the tag data.

TABLE 7
Genre prediction performance in terms of median and MAD across training partitions.

	Precision	Recall	AP	AROC
Metal	0.776 0.010	0.569 0.003	0.826 0.004	0.841 0.002
Folk	0.642 0.010	0.531 0.015	0.734 0.006	0.755 0.003
Electronic	0.800 0.010	0.515 0.009	0.852 0.005	0.769 0.002
Jazz	0.698 0.013	0.577 0.004	0.795 0.006	0.766 0.003
Rock	0.918 0.004	0.524 0.003	0.939 0.001	0.731 0.001
Pop	0.850 0.004	0.629 0.011	0.888 0.004	0.781 0.003

Although the reliability of social genre tags can be questioned, subsidiary role of the audio-based genre prediction in the present paper renders this evaluation adequate. Table 7 shows the performance for each genre cluster in terms of standard evaluation metrics Precision, Recall, Average Precision (AP) and the area under the ROC curve (AROC). For each track the SVM auto-tagger produces probability estimates related to the association strength of each genre cluster. To compute Precision, Recall and AP, we consider for each track three genres with the highest probability as positive. AROC, on the other hand, is computed based on the raw probabilities[7].

The results show that genre prediction from audio is sufficient (cf. [42]), as seen for example from the high AROC for all genres. For the genre-adaptive mood models AROC is arguably the most relevant evaluation metric for genre prediction, since we rely on the raw probability estimates rather than the binary representations.

8.2.2 Mood Prediction Results

Results for genre-adaptive techniques are presented in Table 8. To assess the statistical significance of the performance differences, Wilcoxon rank sum tests across the training partitions were carried out between *ACT+SLPwg* and the other genre adaptive techniques ($p < .05$ denoted by $*$), and between *ACT+SLPwg* and the general SLP ($p < .05$ denoted by †). Genre-adaptive techniques relying on tag- and audio-based genres were evaluated separately this way.

For the core affects, ACT+SLPwg was the most successful technique with both tag- and audio-based genres. The clearest performance improvement was achieved for Valence, that was the most challenging core affect for SLP. Using tag-based genre information alone as predictors of Valence *SLPg* already outperformed the SLP slightly. *ACT+SLPwg* yielded $R^2 = 0.457$ with tag-based genres and $R^2 = 0.431$ with audio-based genres. Both of these results were higher by a large and statistically significant margin than that of the SLP. The performance difference was statistically significant also between *ACT+SLPwg* and the other genre-adaptive techniques, except using *SLPga* with tag-based genres. For Arousal and Tension, *ACT+SLPwg* with audio-based genres yielded the highest performance of $R^2 = 0.741, 0.520$, respectively.

The results for mood terms follow the same patterns as for core affects. The *ACT+SLPwg* with audio-based genres gave the overall highest performance, as it outperformed

7. See [42] for detailed explanation of these metrics

TABLE 8
Performance of the genre-adaptive techniques with genres inferred from tags and audio. Performance improvements over the SLP are marked in boldface.

		Tag-based genres				Audio-based genres			
		SLPg	SLPga	SLPwg	ACT+SLPwg	SLPg	SLPga	SLPwg	ACT+SLPwg
Core affects	Valence	**0.372** 0.003†	**0.453** 0.008 *	**0.434** 0.017†*	**0.457** 0.004 *	0.346 0.003†	**0.400** 0.020†*	**0.397** 0.018†*	**0.431** 0.004 *
	Arousal	0.146 0.004†*	0.702 0.004†*	0.722 0.006	**0.732** 0.003	0.234 0.004†*	**0.731** 0.005	0.725 0.006†	**0.741** 0.005 *
	Tension	0.278 0.007†*	0.463 0.012†*	**0.505** 0.017	**0.518** 0.004 *	0.349 0.009†*	**0.494** 0.018†	**0.497** 0.017†	**0.520** 0.005 *
Mood terms	Atmospheric	0.205 0.016†*	0.662 0.025 *	0.689 0.013†	0.631 0.035 *	0.294 0.013†*	**0.704** 0.021	**0.699** 0.014	0.689 0.024
	Happy	0.221 0.003†*	**0.398** 0.020 *	**0.367** 0.032 *	**0.389** 0.006 *	0.175 0.002†*	**0.332** 0.034†	**0.331** 0.031†	**0.369** 0.012 *
	Dark	**0.379** 0.011†*	**0.378** 0.025†*	**0.300** 0.025 *	**0.268** 0.041	**0.326** 0.016†*	**0.271** 0.019	**0.275** 0.021	**0.270** 0.037
	Sad	-0.000 0.000†*	0.271 0.023†*	0.288 0.012†	**0.330** 0.005 *	0.009 0.002†*	0.296 0.009†	0.291 0.011†	**0.338** 0.006 *
	Angry	0.501 0.004†*	**0.647** 0.009 *	**0.643** 0.016 *	**0.643** 0.006 *	0.571 0.008†*	**0.630** 0.019 *	**0.629** 0.017 *	**0.639** 0.004 *
	Sensual	0.341 0.015†*	0.532 0.026	**0.546** 0.009†	0.517 0.045	0.379 0.008†*	**0.546** 0.009	**0.545** 0.011	**0.546** 0.021
	Sentimental	0.058 0.003†*	0.246 0.021†*	0.282 0.027	**0.338** 0.017	0.096 0.003†*	0.296 0.028†	0.289 0.026†	**0.377** 0.017 *
Avg.		0.250	0.475	0.478	0.482	0.278	0.470	0.468	0.492

† $p < .05$ for performance difference between the *ACT+SLPwg*.

* $p < .05$ for performance difference between the SLP.

SLP for all moods except Atmospheric, and provided statistically significant performance improvement over SLP for Happy, Sad, Angry and Sentimental. It also outperformed the *SLPwg* for all moods except Atmospheric and Dark. Interestingly, SLPg was the most successful technique for Dark. Tag-based *SLPga* performed well for Happy, Dark and Angry, whereas audio-based *SLPga* provided more consistently high performance. Tag-based *SLPg* outperformed all other techniques for Dark, which shows that Dark correlates highly with genre information. This holds also for Angry, except that using audio features in conjunction with genre provides improvement over *SLPg*.

In summary, the *ACT+SLPwg* yielded the highest average performance across moods with both tag-based genres ($R^2 = 0.482$) and audio-based genres ($R^2 = 0.492$). These figures are considerably higher than those of the general SLP that does not exploit genre information ($R^2 = 0.455$). Notably, the *ACT+SLPwg* with audio-based genres gave statistically significant improvements over the SLP for seven out of ten mood scales, and importantly for all core affects. This technique also outperformed the *SLPwg* for eight scales. Since the *SLPwg* was itself more successful than SLP for most of the scales, this suggests that genre-adaptive mood prediction gives significant advantages over the general model, and that modeling the semantic mood space separately within genres is the most beneficial approach. Audio-based genre inference for mood prediction yielded performance comparable to tag-based genres. This indicates that relying solely on audio in making predictions for novel tracks is a potential approach when associated semantic data is not available.

To further assess the benefit of genre-adaptivity, the *ACT+SLPwg* was applied to the test tracks by first randomly rearranging the tag- and audio-based genre weights. If the performance thus obtained would not be affected, the high performance of *ACT+SLPwg* could be attributed to the benefit of ensemble modeling [54] and less to the genre-adaptivity. The analysis showed that this is not the case, as the genre randomization degraded the prediction performance consistently. The average performance across mood scales dropped to $R^2 = 0.424, 0.400$ using tag- and audio-based genres, respectively, and the performance difference was statistically significant at $p < 0.05$ for seven scales (tag-based genres) and for all scales (audio-based genres).

8.2.3 Comparison of ACT+SLPwg and Genre-specific Models

If the aim of an automatic music annotation is to provide mood information to a music collection drawn specifically from one genre, one could ask whether a genre-specific model corresponding to that particular genre would be more appropriate than the *ACT+SLPwg*. Here we test such hypothesis by comparing the prediction performance for core affects separately for subsets of TEST600 belonging to the six genre clusters. TEST600 tracks are assigned to the genre clusters based on the tag data. Notice, that the six subsets of tracks are partly overlapping, as indicated in Table 4. For each of the subsets, Table 9 shows the results obtained with SLP, the genre-specific *ACT+SLPwg* model corresponding to the subset genre, and the *ACT+SLPwg* with audio-based genre information.

The *ACT+SLPwg* yielded consistently higher performance than genre-specific model and SLP, with only few exceptions: Valence/Electronic, Arousal/Metal and Tension/Electronic. Genre-specific models alone were more successful at predicting Valence than the general model. This is further evidence that genre-specific aspects need to be taken into account when modeling Valence. The results for Arousal showed an opposite pattern. These results corroborate the findings of [6], where audio-based genre-specific models of Arousal generalized better across genres than those of Valence.

In overall, the genre-adaptive technique is clearly more successful at predicting mood than the genre-specific models alone. Since genre-specific models rely on training data from one genre, the models may suffer from low variance in mood content, and therefore might not tap into more general relationships between audio features and mood, only attainable from collections of tracks spanning multiple genres.

8.2.4 The Impact of the Number of Genres

To explore the role of the number of genre clusters on the performance of the *ACT+SLPwg* with audio-based gen-

TABLE 9
Prediction performance of the SLP, genre-specific model and the *ACT+SLPwg* separately for tracks from different genres.

		SLP	Genre-specific	ACT +SLPwg
Valence	*Metal*	0.387 0.020	0.407 0.011	**0.421*** 0.008
	Folk	0.199 0.019	0.127 0.037	**0.267*** 0.006
	Electronic	0.239 0.022	**0.339*** 0.009	0.316* 0.014
	Jazz	0.267 0.024	0.305 0.021	**0.360*** 0.012
	Rock	0.311 0.019	0.351* 0.003	**0.378*** 0.004
	Pop	0.225 0.020	0.299* 0.009	**0.306*** 0.006
Arousal	*Metal*	**0.720** 0.006	0.584 0.011	0.713 0.008
	Folk	0.703 0.006	0.674 0.006	**0.715*** 0.003
	Electronic	0.735 0.004	0.727 0.003	**0.748*** 0.003
	Jazz	0.671 0.009	0.642 0.008	**0.686** 0.008
	Rock	0.723 0.005	0.707 0.006	**0.733*** 0.006
	Pop	0.713 0.004	0.716 0.002	**0.723*** 0.004
Tension	*Metal*	0.541 0.014	0.499 0.022	**0.571** 0.004
	Folk	0.379 0.022	0.321 0.030	**0.415*** 0.007
	Electronic	0.372 0.019	**0.424*** 0.030	0.415* 0.007
	Jazz	0.358 0.020	0.336 0.011	**0.399*** 0.007
	Rock	0.473 0.018	0.469 0.005	**0.505*** 0.004
	Pop	0.426 0.025	0.443 0.009	**0.465*** 0.010

$^*p < .05$ for improvement over SLP.

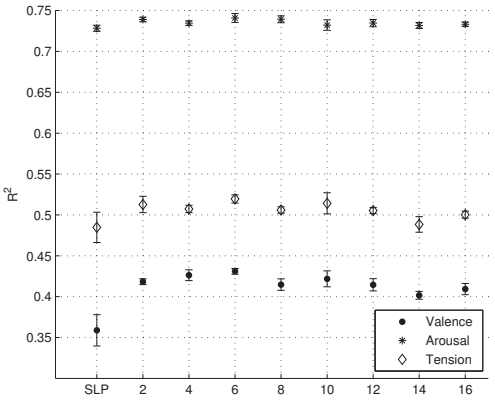


Fig. 5. The Median $\pm$ MAD performance of the genre-adaptive technique (*ACT+SLPwg*) with different numbers of genre clusters ($K = \{2, 4, ..., 16\}$). SLP is shown for comparison.

res, analysis was carried out with 2-16 genre clusters (cf. Table 3). The results, shown in Fig. 5, demonstrate that the *ACT+SLPwg* performance is not overly sensitive to the number of genres, and that the performance remains at a higher level than SLP with all genre clusterings. The optimal performance was found for all core affects at $K = 6$, which can probably be attributed to the fact that TEST600 is balanced according to the corresponding genres.

9 Conclusion

The present study has examined how genre information can optimally be incorporated into music mood prediction. As the general baseline model, SLP performed favorably when compared to a state-of-the-art auto-tagging technique, and comparison to genre-adaptive models showed that taking into account the genre information in mood prediction yields consistent improvements. The best performing novel genre-adaptive technique, the *ACT+SLPwg*, models both the semantic mood space and the audio-based models within different genres separately. Moreover, audio-based genre inference for a novel track performs favorably compared to tag-based inference, which has positive implications for applying the models to large unannotated datasets.

The study also offered survey results and analytical insights into inferring concise music genre representations based on large set of genre tags. Moreover, the study demonstrated that semantic modeling of mood space based on music-specific social tag data is not surpassed by a normative data obtained from a controlled laboratory survey.

Context-adaptive semantic modeling combined with content-based prediction could be transferred to other domains where context-adaptivity could be beneficial, such as object recognition from images, video auto-tagging or multimedia retrieval.

Acknowledgments

The authors would like to thank...

References

[1] T. Schäfer and P. Sedlmeier, "From the functions of music to music preference," *Psychology of Music*, vol. 37, no. 3, pp. 279–300, 2009.
[2] M. R. Zentner, D. Grandjean, and K. Scherer, "Emotions evoked by the sound of music: Characterization, classification, and measurement," *Emotion*, vol. 8, no. 4, pp. 494–521, 2008.
[3] M. Barthet, G. Fazekas, and M. Sandler, "Multidisciplinary perspectives on music emotion recognition: Recommendations for content- and context-based models," in *Proc. 9th Int. Symposium on Computer Music Modeling and Retrieval (CMMR)*, 2012, pp. 492–507.
[4] Y. Panagakis, C. Kotropoulos, and G. R. Arce, "Non-negative multilinear principal component analysis of auditory temporal modulations for music genre classification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 576–588, 2010.
[5] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Computer Science Review*, vol. 6, no. 2, pp. 89–119, 2012.
[6] T. Eerola, "Are the emotions expressed in music genre-specific? an audio-based evaluation of datasets spanning classical, film, pop and mixed genres," *Journal of New Music Research*, vol. 40, no. 4, pp. 349–366, 2011.
[7] Y.-C. Lin, Y.-H. Yang, and H. H. Chen, "Exploiting online music tags for music emotion classification," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, vol. 7, no. 1, p. 26, 2011.
[8] T. Bertin-Mahieux, D. Eck, F. Maillet., and P. Lamere, "Autotagger: A model for predicting social tags from acoustic features on large music databases," *Journal of New Music Research*, vol. 37, no. 2, pp. 115–135, 2008.
[9] R. O. Gjerdingen and D. Perrott, "Scanning the dial: The rapid recognition of music genres," *Journal of New Music Research*, vol. 37, no. 2, pp. 93–100, 2008.
[10] J. Aucouturier and E. Pampalk, "Introduction-From Genres to Tags: A Little Epistemology of Music Information Retrieval Research," *Journal of New Music Research*, vol. 37, no. 2, pp. 87–92, 2008.
[11] P. N. Juslin and J. A. Sloboda, *Handbook of Music and Emotion*. Boston, MA: Oxford University Press, 2010, ch. Introduction: aims, organization, and terminology, pp. 3–14.
[12] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 5–18, Jan. 2006.

[13] P. Saari, T. Eerola, and O. Lartillot, "Generalizability and simplicity as criteria in feature selection: Application to mood classification in music," *IEEE Transactions on Speech and Audio Processing*, vol. 19, no. 6, pp. 1802 –1812, aug. 2011.
[14] Y. H. Yang, Y. C. Lin, Y. F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, Feb. 2008.
[15] T. Eerola, O. Lartillot, and P. Toiviainen, "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proc. 9th International Conference on Music Information Retrieval*, 2009, pp. 621–626.
[16] E. M. Schmidt, D. Turnbull, and Y. E. Kim, "Feature selection for content-based, time-varying musical emotion regression," in *Proc. International Conference on Multimedia information retrieval*. ACM, 2010, pp. 267–274.
[17] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.
[18] J. A. Russell and L. F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant." *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999.
[19] R. E. Thayer, *The Biopsychology of Mood and Arousal*. Oxford University Press, New York, USA, 1989.
[20] P. Saari, T. Eerola, G. Fazekas, M. Barthet, O. Lartillot, and M. Sandler, "The role of audio and tags in music mood prediction: A study using semantic layer projection," in *Proc. 14th International Conference on Music Information Retrieval (ISMIR)*, 2013.
[21] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, 2012.
[22] X. Hu and J. S. Downie, "Exploring mood metadata: relationships with genre, artist and usage metadata," in *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.
[23] Y.-C. Lin, Y.-H. Yang, and H.-H. Chen, "Exploiting genre for music emotion classification," in *Proc. IEEE International Conference on Multimedia and Expo*, 2009, pp. 618–621.
[24] P. Saari and T. Eerola, "Semantic computing of moods based on tags in social media of music," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 10, pp. 2548–2560, 2014.
[25] P. Saari, T. Eerola, G. Fazekas, and M. Sandler, "Using semantic layer projection for enhancing music mood prediction with audio features," in *Proc. Sound and Music Computing Conference 2013 (SMC 2013)*, 2013, pp. 722–728.
[26] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of Information Science*, vol. 32, no. 2, pp. 198–208, April 2006.
[27] S. Deerwester, S. T. Dumais, G. W. Furnas, and T. K. Landauer, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, vol. 41, no. 6, pp. 391–407, 1990.
[28] M. Levy and M. Sandler, "A semantic space for music derived from social tags," in *Proc. 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.
[29] ——, "Music information retrieval using social tags and audio," *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 383–395, 2009.
[30] C. Laurier, M. Sordo, J. Serra, and P. Herrera, "Music mood representations from social tags," in *Proc. 10th International Conference on Music Information Retrieval (ISMIR)*, 2009, pp. 381–86.
[31] P. Saari, M. Barthet, G. Fazekas, T. Eerola, and M. Sandler, "Semantic models of mood expressed by music: Comparison between crowd-sourced and curated editorial annotations," in *2013 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, July 2013, pp. 1–6.
[32] K. R. Scherer, *Emotion as a multicomponent process: A model and some cross-cultural data*. Beverly Hills: CA: Sage, 1984, pp. 37–63.
[33] V. Warriner, Amy Bethand Kuperman and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, pp. 1–17, 2013.
[34] J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.
[35] P. Lamere, "Social tagging and music information retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.
[36] M. Delsing, T. ter Bogt, R. Engels, and W. Meeus, "Adolescents music preferences and personality characteristics," *European Journal of Personality*, vol. 22, no. 2, pp. 109–130, 2008.
[37] P. J. Rentfrow, L. R. Goldberg, and D. J. Levitin, "The structure of musical preferences: A five-factor model." *Journal of Personality and Social Psychology*, vol. 100, no. 6, p. 1139, 2011.
[38] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, Jul. 2002.
[39] M. Sordo, Ò. Celma, M. Blech, and E. Guaus, "The quest for musical genres: Do the experts and the wisdom of crowds agree?" in *Proc. 9th International Conference on Music Information Retrieval (ISMIR)*, 2008.
[40] R. Ferrer, T. Eerola, and J. K. Vuoskoski, "Enhancing genre-based measures of music preference by user-defined liking and social tags," *Psychology of Music*, vol. 41, no. 4, pp. 499–518, 2013.
[41] B. Schuller, H. Hage, D. Schuller, and G. Rigoll, "'mister d.j., cheer me up!': Musical and textual features for automatic mood classification," *Journal of New Music Research*, vol. 39, no. 1, pp. 13–34, 2010.
[42] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.
[43] S. R. Ness, A. Theocharis, G. Tzanetakis, and L. G. Martins, "Improving automatic music tag annotation using stacked generalization of probabilistic svm outputs," in *Proc. 17th ACM International Conference on Multimedia*. ACM, 2009, pp. 705–708.
[44] R. Miotto and G. Lanckriet, "A generative context model for semantic music annotation and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 4, pp. 1096–1108, 2012.
[45] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proc. 10th International Conference on Digital Audio Effects, Bordeaux, France*, September 2007.
[46] J. C. Gower and G. B. Dijksterhuis, *Procrustes Problems*. Oxford University Press Oxford, 2004, vol. 3.
[47] T. Eerola and J. K. Vuoskoski, "A review of music and emotion studies: Approaches, emotion models and stimuli," *Music Perception*, vol. 30, no. 3, pp. 307–340, 2013.
[48] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Advances in Neural Information Processing Systems*, vol. 14. Cambridge, MA: MIT Press, 2001, pp. 849–856.
[49] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, april 1979.
[50] B. Mirkin, *Mathematical Classification and Clustering*. Kluwer Academic Press, Dordrecht, 1996.
[51] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.
[52] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*. Springer Berlin Heidelberg, 2004, pp. 39–50.
[53] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music." *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.
[54] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

Pasi Saari received MSc degree in Computer Science in 2008 and MA degree in Musicology in 2010 from the University of Jyväskylä, Finland. He is currently working as a Doctoral Student at the Finnish Centre of Excellence in Interdisciplinary Music Research within the University of Jyväskylä, Finland. His research interests are in semantic computing of moods in music and content-based analysis of musical audio.

György Fazekas is a post doctoral research assistant at Queen Mary University of London, working at the Centre for Digital Music (C4DM), School of Electronic Engineering and Computer Science. He received his BSc degree at Kando Kalman College of Electrical Engineering, now Budapest Polytechnic University, Kando Kalman Faculty of Electrical Engineering. He received an MSc degree at Queen Mary University of London, and a subsequent PhD degree at the same institution in 2012. His thesis titled Semantic Audio Analysis–Utilities and Applications explores novel applications of semantic audio analysis, Semantic Web technologies and ontology-based information management in Music Information Retrieval and intelligent audio production tools. His main research interest includes the development of semantic audio technologies and their application to creative music production. He is working on extending audio applications with ontology based information management. He is involved in several collaborative research projects, and he is a member of the AES, ACM and BCS.

Tuomas Eerola received his MA and Ph. D degrees from the University of Jyväskylä, Finland, in 1997 and 2003. He is a Professor of Music Cognition at the Durham University, UK. His research interest lies within the field of music cognition and music psychology, including musical similarity, melodic expectations, perception of rhythm and timbre, and induction and perception of emotions. He is on the editorial boards of Psychology of Music, and Frontiers in Digital Humanities and is consulting editor for Musicae Scientiae and member of the European Cognitive Sciences of Music (ESCOM) and the Society for Education, Music and Psychology Research (SEMPRE).

Mathieu Barthet was born in Paris, France, in 1979. He received the M.Sc degree in Acoustics, Aix-Marseille II University, Ecole Centrale Marseille, France, and the Ph.D. degree from the Aix-Marseille II University in 2004 in the field of musical acoustics and signal processing. From 2007 to 2008, he was a Teaching and Research Assistant at Aix-Marseille I University. Since 2009, he has been a Postdoctoral Research Assistant at Centre for Digital Music, Queen Mary University of London, London, UK. He has taken part in several projects in collaboration with the BBC, the British Library, and I Like Music.

Olivier Lartillot is a researcher in computational music analysis at the Department for Architecture, Design and Media Technology, Aalborg University, Denmark. Formerly at the Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, he designed MIRtoolbox, a referential tool for music feature extraction from audio. He also works on symbolic music analysis, notably on sequential pattern mining. In the context of his 5-year Academy of Finland research fellowship, he conceived the MiningSuite, an analytical framework that combines audio and symbolic research. He continues his work as part of a collaborative European project called Learning to Create (Lrn2Cr8).

Mark Sandler (SM98) was born in 1955. He received the B.Sc. and Ph.D. degrees from the University of Essex, Essex, U.K., in 1978 and 1984, respectively. He is a Professor of Signal Processing at Queen Mary University of London, London, U.K., and Head of the School of Electronic Engineering and Computer Science. He has published over 350 papers in journals and conferences. Prof. Sandler is a Fellow of the Institute of Electronic Engineers (IEE) and a Fellow of the Audio Engineering Society. He is a two-time recipient of the IEE A. H. Reeves Premium Prize.