UNIVERSITY OF CALIFORNIA
Santa Barbara

Interactive Latent Space for Mood-Based Music Recommendation

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Media Arts and Technology

by

Ivana Andjelkovic

Committee in Charge:

Professor Curtis Roads, Chair

Dr. John O'Donovan, Associate Researcher, Lecturer

Professor Dennis Parra, PUC Chile

Professor Matthew Turk

Professor Gert Lanckriet, UC San Diego

March 2016

The Dissertation of
Ivana Andjelkovic is approved:

Dr. John O'Donovan, Associate Researcher, Lecturer

Professor Dennis Parra, PUC Chile

Professor Matthew Turk

Professor Gert Lanckriet, UC San Diego

Professor Curtis Roads, Committee Chairperson

December 2015

Interactive Latent Space for Mood-Based Music Recommendation

by

Ivana Andjelkovic

Acknowledgements

First, I would like to thank my advisors Curtis Roads and John O'Donovan for believing in me, inspiring me as people with admirable qualities and sharing their expertise throughout my graduate studies. Special thanks to Denis Parra for his generous guidance, coaching and always positive attitude. I would also like to thank Shlomo Dubnov for his patience in helping this research get off the ground. Thanks to Matthew Turk and Gert Lanckriet for their critical insights. I am very grateful for the opportunity to be part of Media Arts & Technology community and know extraordinarily brave and creative people. My horizons broadened and life changed drastically thanks to the interaction with them.

I would like to thank my family in Europe for teaching me independence and instilling sense of wonder and adventure in me. Thanks to Rose family, my second American family, without whom my highest education would not be possible. Thanks to Bogac Cetiner for countless hours on the phone, helping me face my biggest challenges during the PhD studies. A very special thanks to my Yoga & Meditation teacher, Hari Das, for her glowing grace and beauty, and for teaching me how to reach infinite compassion, faith and strength. Finally, thanks to all my friends, San Diego yogis, volleyballers and sailors for all the joyful moments and fun adventures!

Abstract

Interactive Latent Space for Mood-Based Music Recommendation

Ivana Andjelkovic

The way we listen to music has been changing fundamentally in past two decades with the increasing availability of digital recordings and portability of music players. Up to date research in music recommendation attracted millions of users to online, music streaming services, containing tens of millions of tracks (e.g. Spotify, Pandora). The main focus of up to date research in recommender systems has been algorithmic accuracy and optimization of ranking metrics. However, recent work has highlighted the importance of other aspects of the recommendation process, including explanation, transparency, control and user experience in general. Building on these aspects, this dissertation explores user interaction, control and visual explanation of music related mood metadata during recommendation process. It introduces a hybrid recommender system that suggests music artists by combining mood-based and audio content filtering in a novel interactive interface. The main vehicle for exploration and discovery in music collection is a novel visualization that maps moods and artists in the same, latent space, built upon reduced dimensions of high-dimensional artist-mood associations. It is not known what the reduced dimensions represent and this work uses hierarchical mood model to explain the constructed space. Results of two user studies, with over 200 participants each, show that visualization and interaction in a latent space improves acceptance and understanding of both metadata and item recommendations. However, too much of either can result in cognitive overload and a negative impact on user experience. The proposed visual mood space and interactive features, along with the aforementioned findings, aim to inform design of future interactive recommendation systems.

Contents

List of Figures

Chapter 1

Introduction

1.1 Overview

Recommender systems are popular tools for predicting content that a target user is likely to be interested in, and they are a key component in many online systems nowadays since they allow people to find relevant items under information overload. Despite the large amount of research into these systems in terms of algorithmic accuracy, there are still under-explored areas such as emotion-aware recommenders and rich interfaces beyond static ranked-lists. Due to experimental evidence linking music and emotion, this dissertation contributes to research in recommender systems by introducing *MoodPlay*, a hybrid recommender system that suggests music bands by combining audio content and mood-based filtering in a novel interactive interface.

The work presented here addresses two related research topics. First, music recommendation that considers a user's mood, and second, interaction mechanisms that allow for explanation and elicitation of mood information. There are several platforms that recommend music based on different types of listening context (daily activity [85], time of the day [5], music genre [53], etc.). However, given the strong experimental evidence showing that music modulates our emotions, which are further linked to attention and communication [48], this research focuses on user's moods when recommending new music. Furthermore, the importance of building interactive recommender interfaces that go beyond the static-ranked list paradigm has been studied in the past [23, 30, 11, 46, 84, 64, 57]. Results show that higher user satisfaction is not always correlated with small improvements in recommendation accuracy [56, 49], but may be correlated with interface enhancements. Accordingly, the goal of this dissertation project is to build a recommender system with an interactive interface that supports users on making music choices over a mood space. Specific research questions are:

Q1: How can metadata such as mood information be visually represented for a music recommendation system?

Q2: How can interaction, explanation and control be supported over such a visualization?

Q3: What are the effects of such interactive visualizations on the user experience with a recommender system? For example, how much interaction is too much?

Q4: How does knowledge of and interaction with mood metadata influence recommendation accuracy and user experience?

In the effort to answer the above questions, a web based, mood aware music recommendation system was produced. Up to date work resulted in the following key contributions:

- *A novel visual interface for mood-aware recommendation.* We introduce a visualization that maps moods and music artists in the same space, and allows users to explore the items along these dimensions. To our knowledge, this is the first mood-based recommender interface that integrates the Geneva Emotional Music Scales (GEMS) model [92].
- *Mood-aware recommendation algorithm.* We describe a novel hybrid recommendation algorithm for mood and audio content-based music recommendation.
- *Enhanced interaction techniques.* We introduce new interaction mechanisms for hybrid recommendation on a latent space. Trail-based mechanism supports influence of previous navigational steps on the recommendation set. Radius-based method allows users to control the ratio of mood and audio content hybridization in the recommendation algorithm.
- *Evaluation design and analysis.* We present design and results of an evaluation of the system. We conducted two user studies (N=240, N=279) over 4

different conditions of increasing complexity - from baseline without visualization to the full interactive interface with a hybrid mood and content-based recommender algorithm. Based on the results, we provide some lessons for interface design in the context of exploratory tasks in recommender systems.

To evaluate the proposed system, we conducted user studies over 4 different conditions, having the following features: (1) static recommendations in the form of ordered lists, generated based on user's taste profile, (2) static recommendations, highlighted in a latent mood space visualization, (3) dynamic recommendations generated via user interaction in the latent mood space visualization, using current mood preference and (4) dynamic, interaction driven, trail-based recommendations.

Following are the key results of the system evaluation: (1) User experience metrics generally improve in conditions with mood visualization and more so with increased interaction, (2) Visualization and interaction in the mood space help users understand artists' moods and how the system arrived at the recommendations, (3) Visualization of mood space increases diversity and perceived accuracy, without impacting predictive accuracy of the algorithm, (4) Interaction in the mood space increases diversity of artists that were discovered by the user, but interaction with trail-based recommendation (provenance of positions in the mood-space) decreased diversity, (5) The biggest significant changes in user ratings on recommended items were observed in the most interactive condition, however, direction of the rating shift varied across participants.

1.2 Definitions

Recommender (recommendation system): Software that analyzes available data to help users find desired items in a crowded information space.

Interactive recommender: Recommender that enables the user to steer the received recommendations in the desired direction through explicit interaction with the system [78]

Latent space: Space containing data entities after dimensionality reduction is performed. Numerical methods are often used to represent high-dimensional data points in 2 or 3 dimensions, meaning of which is not known in advance. Hence, we refer to the obtained low-dimensional space a latent space.

Mood-based music recommender: Software that recommends music with the goal to match user's desired mood.

1.3 Affective States in Music

1.3.1 Affect, Emotions and Moods

As the title of the dissertation indicates, this research is concerned with mood-aware recommendation. It is important to define and make a distinction between three related terms before further delving into research problems:

Affect: Colloquial term that encompasses both emotions and moods [26, 83].

Emotions: Intense, short lived feelings, speculated by most researchers to be directed at someone or something [24].

Moods: General, low intensity feeling states that often lack a contextual stimulus [38, 87]. While the duration of emotions is typically measured in minutes, moods may last several hours or days and cause us to think or brood for a while [38, 65].

According to Hume [38], affect is an umbrella term for emotions and moods, which can mutually influence each other. Furthermore, both emotions and moods are often described in terms of positive or negative affect [86]. Emotions become mood states when grouped into positive and negative categories because such grouping allows us to look at emotions more generally instead of in isolation [38]. Therefore, emotion models such as Circumplex model of affect [71] are often used to represent moods as well.

1.3.2 Emotions and Moods in Music Research

Music has a strong effect on our emotions and speculatively somewhat milder effect on our moods [83]. In fact, music has been efficiently used to induce emotions and moods and study them outside of realm of music psychology [83]. As described by Curtis Roads in *Composing Electronic Music: A New Aesthetic* [70]:

Music inevitably invokes psychological states that are bound up with emotional responses. Even "technical" music like the Bach Two and Three-part Inventions (BWV 772-801) evokes affective reactions. This tightly constrained and orderly *music is at turns comforting, exhilarating, wistful, dazzling, contemplative, and delightfully inventive.*

Unlike in theory, the distinction between emotions and moods is not always clear in practice and depending on the area, researchers choose to focus on one or the other. For example, music psychologists often study emotional responses to music, while scientists interested in computational classification of music focus more on moods [37]. Psychologists are interested in emotions, which are short lived and may vary greatly (e.g. from sad to happy) even in a single musical piece. In contrast, to automatically classify music, researchers benefit from metadata describing stable affective states, commonly experienced by many people during music listening [37]. This is probably the reason why large, publicly available music repositories (Last.fm[1], AllMusic[2]) use moods to characterize and organize music. Hence, the terminology adopted in this dissertation research is recommendation based on moods rather than emotions.

1.3.3 Expression and Perception of Affect in Music

Research has shown that listeners are rather good at recognizing the intended expression in music [83]. Juslin et al. [40] summarize characteristics of music that listeners use to identify an intended emotional expression. For example, they indicate that happy music is characterized by medium to high voice intensity/sound level and medium high-frequency energy (among other parameters), while anger is

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

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

expressed with high voice intensity/sound level and much high-frequency energy. However, even though the listener can recognize the emotion in musical expression, she does not necessarily experience it. This can be exemplified the best with "spooky" or "scary" songs. It is easy for us to recognize fear in the music, but we rarely become afraid while listening to it.

1.4 Scope and Limitations

The work presented in this document addresses research questions in several sub-fields, including music recommendation, affective computing and interactive interfaces. Its scope must be limited.

As justified in the previous section, the system developed in this research visualizes and recommends artists using mood metadata. Although there is variance from song to song in terms of moods for one artist, this research is based on a comprehensive, professionally curated database where a variety of moods associated with an artist are given weights to compensate for the fact that different songs have different moods. As a result, moods with the greatest weights have the most impact on the artist's positioning in the visualization.

Next, since the professionally curated database of artist-mood associations is used, the interactive system developed in this research does not address the differences between moods expressed through music and those perceived by listeners. Similarly, this research does not go in depth to determine whether perceived mood is also induced in the listener. Evaluation sessions during which the users listened to recommended music lasted at least 1.5 minutes and in very few cases 20 minutes. Such a short period of time may not be sufficient to alter listener's mood, but the ratings of recommendations were collected nevertheless. It is suspected that the ratings reflect user's opinion about how well the recommended items match music from the user profile, how well they match mood labels in the area of mood space where the user navigated to and how they align with user's taste in general.

At this stage of research, the user taste profile is built upon the list of artists that user enters into the system. This research is not yet concerned with developing a sophisticated method for acquiring information about user's current and desired mood - the system simply asks users to enter artists they like to listen to. Provided list of artists may represent user's general preference, current mood or desired mood, which is up to the user to decide. The recommender suggests music in the similar mood, but also allows the user to navigate to different areas in the mood space, therefore changing the target mood for recommendation.

Finally, the system is developed as a web application and is not optimized for usage on mobile devices. Due to the complexity, the current architecture and design limits its usage to situations where users can allocate sufficient time and attention.

1.5 Document Outline

This first chapter introduced the research questions and described the contribution of the dissertation. The remainder or the thesis is organized as follows:

- Chapter 2 gives an overview of related work, including research in music recommendation, affective computing, emotion models and interfaces for music recommendation.

- Chapter 3 describes proposed visualization of artist - mood associations. It argues there is a need to create a visual, music specific mood space for music recommendation and discovery, describes the research conducted in this area and implementation details of visual mood space.

- Chapter 4 introduces MoodPlay - a mood based recommendation system with rich interface, developed as a tool for answering research questions of this thesis. It includes details about interface design and hybrid recommendation algorithms.

- Chapter 5 describes the procedure and measures for evaluating the proposed system.

- Chapter 6 presents results of a preliminary user study (N=240). The evaluation of the system though this study was focused on user characteristics, interaction and experience, and less attention was placed on ratings-based analyses.

- Chapter 7 presents results on second, more comprehensive user study (N=279), with improved user interface and experiment design. In addition to more detailed investigation of user interaction and exploration and their effects on recommendation ratings and diversity, we also explore influence of mood on ratings.
- Chapter 8 summarizes the contributions and findings of this thesis, and discusses directions for future research.

Chapter 2

Background

The research covered by this thesis spans over several related areas: music recommendation, interfaces for recommendation and mood aware recommender systems. Yet, at its core, it is a multifaceted study of the interplay between music, mood and listener, with aspects of automated algorithms, UI, interaction design and user modeling. This chapter presents relevant studies in the above mentioned areas and describes where the thesis contributions stand in relation to them.

2.1 Music Recommendation

Recommendations in the music domain is a well-established field within recommender systems. Among many others, methods have been developed to recommend tracks [17, 52], albums [62], artists and music bands [11, 35], playlists

[53, 4, 34], music targeted at specific venues [41] and music targeted at daily activities [85]. Furthermore, many of the music recommendation algorithms are being developed and improved in commercial setting, due to availability of online streaming services that give listeners access to millions of songs. The remainder of this section describes common machine learning methods employed to calculate recommendations and main features of popular music recommendation platforms.

Current state of the art algorithms for music recommendation can be divided into four main categories, depending on the type of data they utilize [16].

1. *Collaborative filtering (CF)* method typically uses rating or purchase patterns of like minded individuals to make recommendations for active user. It commonly relies on music metadata – textual information about songs and artists that can be factual (title, release data, geographical location) or cultural (genre, mood, social tags[1]).

2. *Context based (CXB)* methods rely on web mining techniques and social tags, in order to derive context of music, based on which the similarity between songs or artists is computed.

3. *Content based (CNB)* methods utilize item content to compute the similarity and recommend new items. Most often this is audio content, which can be computed using signal analysis methods or it can be described manually by musicologists.

[1] Short textual descriptors of songs or artists provided by many users

4. *Hybrid (H)* methods can be any combination of previously listed methods. They help minimize the disadvantages of individual algorithms.

Many of the currently available commercial music information retrieval (MIR) systems rely heavily on collaborative filtering, possibly in the combination with context based techniques (e.g. Last.fm[2], iTunes Genius[3]). They work very well when large amount of data is available, but one of their major drawbacks is the ”long-tail” problem – less known items have less metadata and ratings associated with them, hence they are less likely to be found in queries and recommended as relevant items. Therefore, systems relying on collaborative data and metadata are not suitable for music discovery and need to be complemented by content-based methods [15].

On the other hand, purely content based retrieval methods do not capture some of the information that users can provide with metadata [18]. For example, music descriptors such as dreamy, signer-songwriter, vintage can hardly be extracted from acoustic content. Furthermore, some studies suggest that music similarity using acoustic features has reached the glass ceiling of around 50% for general audio similarity (Mirex[4] evaluations) and roughly estimating, up to 70% using various constraints (Mirex, [3, 8]). Consequently, hybrid systems, such as those combining content based and metadata based techniques can yield better recommendation results than individual methods [10].

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

[3] https://support.apple.com/kb/PH20373

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

To better understand the features of popular music recommendation systems, Table 2.1 summarizes their characteristics. In most cases, detailed information about recommendation algorithms they employ is not available. Hence, the table lists most prominent algorithm each of the systems use, which may be in reality complemented with other techniques to improve the recommendation. Social networking feature refers to the ability of users to connect and communicate with each other, and social tags are music descriptors provided by users themselves. Because the music listening context is of particular importance for this research project, any references to mood and context associated with songs or artists are examined as well.

It can be seen from the table that many services allow users to connect with each other, thus making music listening a social experience. This facilitates the creation of social tags – words describing genre, mood, context or general impression, that users associate with tracks and artists. However, it can be noted that current systems do not offer fully developed recommendation based on the varying contexts of music listening experiences. Among the listed systems, Musicovery is the only service that provides automatic recommendation based on the mood. Stereomood offers automatically created playlists based on moods and context, but it relies on user provided tags which are sparse and noisy. Remaining three systems, Songza, Spotify and All Music Guide offer mood and context based music suggestions but not in the form of interactive and continuous music listening experience.

System	Source of data	Recommend. algorithm	Social network	Social tags	References to music moods	References to music context
Pandora	Musicologist take surveys	CNB	x	-	-	-
Last.fm	Activity data, tags on artists and songs, acoustic analysis	CF	x	x	-	-
Spotify	Acoustic analysis, text analysis	CXB + CNB(by EchoNest)	x	-	Manually created playlists based on moods	Manually created playlists based on context
Songza	Editors or music fans make playlists	N/A	x	-	Manually created playlists based on moods	Manually created playlists based on context
iTunes Genius	Purchase data, activity data from iTunes	CF	-	-	-	-
Grooveshark	Unknown	Unknown	x	x	-	-
Stereomood	Unknown	Unknown	-	x	Automatically created playlists based on social tags that reflect moods	Automatically created playlists based on social tags that reflect context
Rdio	Acoustic analysis, text analysis	CXB + CNB(by EchoNest)	x	-	-	-
Musicovery	Unknown	Unknown	-	-	Automatic recommendation based on the mood	-
All Music Guide	Music editors & writers	Unknown	-	-	Editorial mood tags. Tracks and artists categorized by mood.	Editorial mood tags. Tracks and artists categorized by context.

Table 2.1: List of popular music recommendation systems and their characteristics

The general approach taken by popular music recommendation systems is to learn about user's taste from a seed song or listening history, and to suggest more of the same music. Recommendations are usually presented as an ordered list, either in text or by playing one song after another. User's interaction with a system is often limited to liking, disliking and skipping a song, which affects subsequent music suggestions. On the other hand, academic research of interaction and user controllability in recommendation systems is steadily advancing. Particularly relevant to this thesis is TasteWeights, a system that allows users to control different aspects of a hybrid recommendation algorithm through a visual interactive interface [11].

Compared to previous approaches outlined here, this thesis innovates by (1) using hybrid of artists' mood representation and audio content to compute similarity and recommend items based based on user's mood, (2) by introducing a novel recommendation interface and (3) by providing user controls to explore the artists dataset interactively.

2.2 Affective Computing and Recommendation

Research in affective computing has been gaining extensive attention in recent years. Proliferation of mobile and wearable computer devices makes it both necessary and possible to achieve natural and harmonious human-computer interaction. Such devices enable us to track a variety of sources that carry emotional content. For example, different aspects of bodily movement and gestures have been used to recognize emotions: head and hands motion [27], gout patterns [42], body posture [44], to name a few. In the speech domain, vocal parameters such as pitch, speaking rate, formants and modulation of spectral content have also been successfully used to classify emotions in [66, 90, 88]. Furthermore, currently largest data repository of face videos (2 million) owned by Affectiva[5] is efficiently used to train computers in detecting emotions from facial expressions in real time.

The important role of emotions on human decision-making and judgement [67, 58] has made mood an actively studied variable in context-aware recommender systems. For instance, Masthoff *et al.* [54] integrated affective state in a group

[5] http://www.affectiva.com

recommender system by modeling satisfaction as mood, while González *et al.* [28] incorporated the emotional context in a recommender system for a large e-commerce learning guide. More related to this thesis, Park *et al.* [61] developed probably the first context-aware music recommender that exploited mood inferred from context information. Other works followed their approach inferring the users' mood for music recommendation based on movements, temperature and weather [21] or from the music content [69]. For instance, Griffiths *et al.* [31] measured a variety of contextual and physiological indicators of mood (temperature, light, heart activity). Mapping of both users' mood and music on the same emotion map enabled them to recommend music in the detected mood. Zwaag *et al.* [82] took target mood as an input from user and then selected songs that direct the user towards the desired mood, while measuring skin conductance to verify the change. Skin temperature [39] and arm gestures [2] have also been used for inferring mood and querying music collections.

Tkalcic *et al.* [81, 33] discussed the role of emotions in recommender systems and introduced a framework to identify the stages where emotion can be used for recommendation. They identified four main areas of research (i) the use of emotions as context in the entry stage, (ii) modeling affective content user profiles, (iii) using affective profiles for recommending items and (iv) building datasets. Within these categories, system proposed in this thesis deals with (i), (ii) and (ii). Moodplay uses emotions as entry stage by allowing users to navigate artists in a mood space, it models the user profile as a set of artist which are represented in a mood-based vector model, and these mood-based profiles are used for recommending music artists. In addition, rich user interface is proposed to help users explore mood space and choose music in desired mood. In the future, the system would be greatly enhanced by incorporating a method for automatic mood detection, using sensors available on wearable devices, social media activity or contextual information.

2.3 Interactive Interfaces for Recommendation Systems

The importance of developing interfaces for recommender systems rather than focusing only on improving recommendation algorithms, a user-centric approach, has been highlighted by the work of MacNee *et al.* [56] and Konstan *et al.* [49], who showed that small improvements in recommender accuracy do not necessarily improve users' satisfaction with a system. While rankings and similarity are of great importance to information search and recommendation, a level of diversity, surprise and serendipity is often desirable during music listening [76, 73]. These latter aspects can be supported by carefully designed user interface. Yet, the development of interfaces that present recommended items in a visual model different than a static ranked list is rather scarce. Some examples include SFViz [29], a sunburst visualization that allow users to find interest-based recommendation in Last.fm, and Pharos [93], a social map visualization of latent communities. Other examples that, in addition to visualizations, include a richer user interaction are PeerChooser [59], SmallWorlds [30] – that focus in representing collaborative filtering, and TasteWeights [11] [46], an interactive system that represents a hybrid recommender of music bands. In a different domain, TalkExplorer [84] is a graph-based interface with facets that let users explore and find relevant conference talks by analyzing the connections of different entities. Other work on visual interfaces related to the academic domain is SetFusion [64], an interface for conference article recommendation that makes use of an interactive Venn diagram to let users control the importance of different recommendation approaches, and a range of systems that support dynamic critiquing of an algorithm, such as Pu *et al.* [68] and Chen *et al.* [19]. Finally, with a focus on making users aware of the filtering mechanisms on a social network, Nagulendra and Vassileva [57] created an interactive interface presenting groups of categories and people into *bubbles* with the purpose of providing users' with awareness and control of the personalization mechanism.

Building on concepts from these previous works, MoodPlay offeres novel interface that maps artists in a mood-space and allows user to navigate the space by moving an avatar. It also ncorporates the notion of trails to account for historical user preference data and allow the user more flexibility on an incremental process of obtaining recommendations.

2.4 Visualizations of Music Collections

Visualizations are invaluable tool for exploring large datasets and understand relationships between items. To be truly effective, visual layout is often accompanied with interface for data filtering, highlighting relevant items or displaying additional information. For example, Soriano *et al.* [77] developed a tool for exploring music collection based on musical structure of songs, while MusicBox [51] tool offers the ability to select features for dynamically computing song similarity and map songs in space as a result. Songrium [32] detects remixed tracks published or shared on the web and embeds the visualization of relationships with originals into a rich interactive interface. An intuitive, game-like navigation though 3D music landscape, enriched by aural experience of near-by songs in Neptune [45] elicited positive reactions from users. More related to our work, [22] and [25] additionally utilize visualizations to highlight recommended music, thus providing a degree of transparency to the recommendation algorithm. On the other hand, very few works attempt to visualize music dataset according to listenting context or visualize changes of user preference over time. Following two subsections present notable related research.

2.4.1 Visual Mood Models

Although mood-based music selection and recommendation are gaining popularity in both research and commercial settings, development of visual aids for mood information is still scarce. Nearly all existing mood based visualizations are built upon Russell's circumplex model of affect [71], derived from general research in psychology. First presented in the 1980, this model is now commonly used to represent emotions as a mixture of two dimensions, valence and arousal, and position them in the coordinate system (Figure 2.1). Valence, the degree of pleasure, ranges from positive to negative whereas arousal ranges from low to high psychological activity. Yang *et al.* [89] incorporate it into their music retrieval method, and commercial applications such as Habu[6] and Musicovery[7] use it as a platform for music selection based on mood. Habu and Musicovery label axes as Dark – Positive and Calm – Energetic, and position songs in space according to the moods associated with them (Figures 2.2 and 2.3). Habu categorizes songs from user's personal collection into 25 grouped moods, or 100 granular moods, mapped onto the emotion plot. Musicovery, on the other hand, positions songs from an online music collection in the emotion plot without revealing specific, associated mood words to the user.

However, it has been shown that many emotions cannot be uniquely characterized by valence and arousal values [20]. For example, fear and anger, two distinctive emotions, both have high arousal and negative valence, and are commonly placed close to each other in the circumplex model [74]. Similarly, disparate moods, such as *wistful/forlorn* and *casual groove*, are found in close proximity in the Habu visualizations. Furthermore, many mood words that we use to charac-

[6] http://habumusic.com
[7] http://musicovery.com

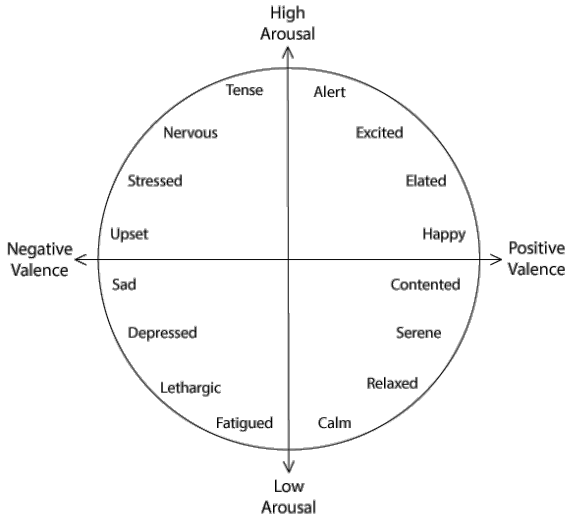


Figure 2.1: Russel's circumplex model of affect

terize music are not necessarily positive, negative, calm or energetic. Humorous, eccentric and philosophical are just a few examples of such words found in Rovi[8] - a popular database of comprehensive, professionally curated music descriptors. It is also important to note that models derived from general research in psychology, such as Russell's, may not be suitable for musical emotions. One reason being that music, unlike other life events, does not have goal implications, and thus possibly induces more contemplative range of emotions [91]. To address this problem, this thesis proposes a novel visual representation of music specific moods, built upon on a model derived from extensive psychological study by Zentner *et al.* [92].

[8] http://developer.rovicorp.com/docs

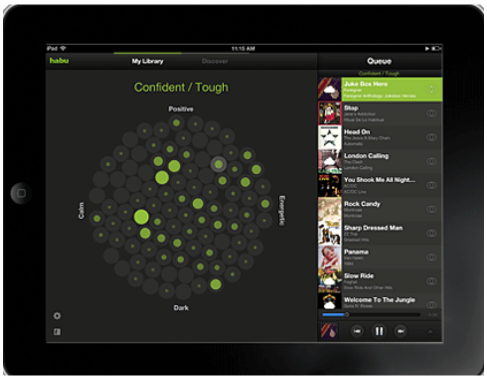


Figure 2.2: Habu interface

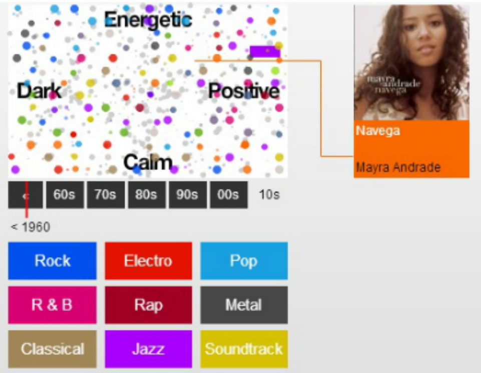


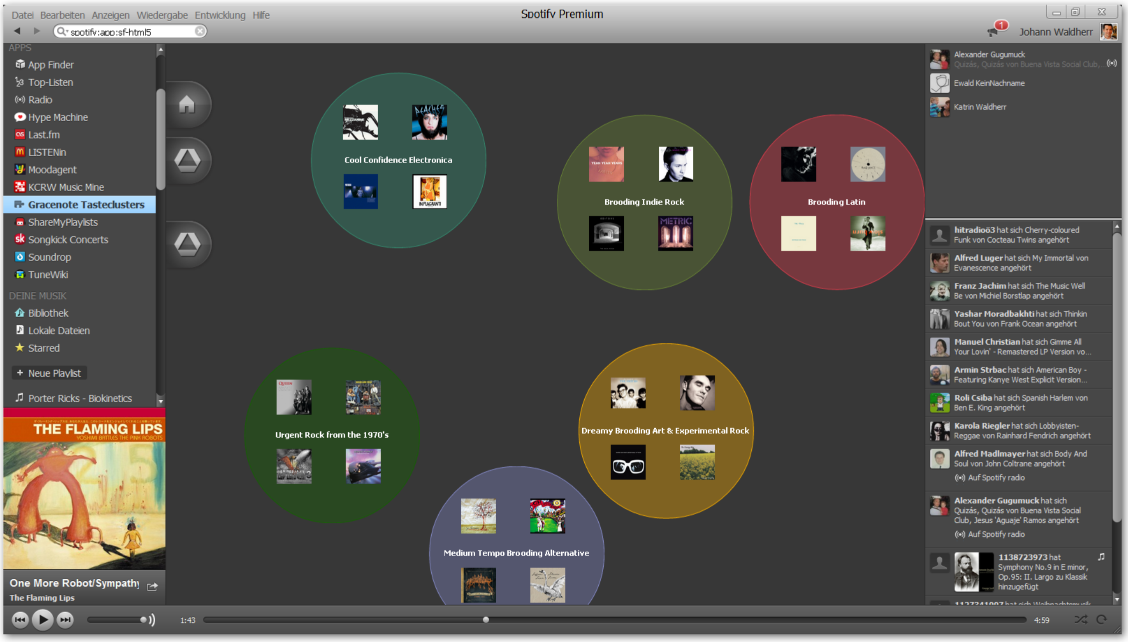
Figure 2.3: Musicovery interface

2.4.2 Visualizations of User's Preference

User's musical profile can be inferred from her listening history, which often consists of stylistically different tracks, conveying a variety of moods. Therefore, it would be difficult, and possibly undesirable to characterize a user with a narrow set of music and mood descriptors. This is likely the reason most visualizations of taste are elaborate, rather than concise, statistical representations of listening history. Publicly available user data on Last.fm inspired many to build tools for visualizing historical information. For instance, Baur et al. [6] plot artists on a time scale, color coded by genres, such that listening patterns can be observed throughout a day or over the years. Such visualizations reveal much about user's taste, but they do not aid the user in selection of music nor give greater control of the recommendation. On the other hand, research in companies Spectralmind and Gracenote resulted in a Spotify application Tasteclusters[9], which personalizes music based on taste and allows navigation using a graphic display (Figure 2.4).

[9] http://www.spectralmind.com/tag/music-visualization

The listening history is analyzed and grouped based on moods and genres. The results are intuitively arranged into personalized clusters, allowing the user to choose music from available groups. Compared to these approaches, MoodPlay models historical user preference by storing and displaying individual mood points, created as user navigates though the mood space.



hitradioö3 hat sich Cherry-coloured Funk von Cocteau Twins angehört

Alfred Luger hat sich My Immortal von Evanescence angehört

Franz Jachim hat sich The Music Well Be von Michiel Borstlap angehört

Yashar Moradbakhti hat sich Thinkin Bout You von Frank Ocean angehört

Manuel Christian hat sich Gimme All Your Lovin' - Remastered LP Version vo...

Armin Strbac hat sich American Boy - Featuring Kanye West Explicit Version...

Roli Csiba hat sich Spanish Harlem von Ben E. King angehört

Karola Riegler hat sich Lobbyisten-Reggae von Rainhard Fendrich angehört

Alfred Madlmayer hat sich Body And Soul von John Coltrane angehört
(•) Auf Spotify radio

Alexander Gugumuck hat sich Quizás, Quizás von Buena Vista Social Club, Jesus 'Aguaje' Ramos angehört
(•) Auf Spotify radio

1138723973 hat Symphony No.9 in E minor, Op.95: II. Largo zu Klassik hinzugefügt

Figure 2.4: Tasteclusters interface

Chapter 3

Music Specific Visual Mood Model

3.1 Overview

After reviewing the background research in relevant areas, this chapter aims to answer the first research question: *How can metadata such as mood information be visually represented in a music recommendation system?* As previously stated, to our knowledge, Russel's Circumplex model of valence and arousal is the only emotion model used to visually represent music artists and moods associated with them. However, due to its shortcomings described in 2.4.1, this dissertation proposes a novel approach to mapping artists and moods in the same space.

Established music services use vocabularies of various sizes to describe moods associated with different songs or artists. Songza[1] categorizes music based on 20 moods pre-defined by music professionals, Habu[2] uses 100 granular mood words provided by curators as well, while Stereomood's[3] growing dictionary contains over 100 words provided by users. Mood metadata is usually available on artist level and differs across different publicly available services (music repositories). Three most popular and widely used repositories in both commercial setting and academic research are EchoNest[4], Gracenote[5] and Rovi[6]. Comparing to EchoNest and Gracenote, Rovi offers the most comprehensive, professionally curated list of mood descriptors (289 unique words). It houses data pertaining to 3.5 million albums and comparable number of artists, most of which are tagged with a small subset of different moods. It can be argued that one artist plays songs in different moods and that mapping songs, rather than artists, in the mood space would yield a more accurate representation of a music collection. However, using moods on the artist level for the purpose of this research project has the following advantages:

- By visualizing artists, fewer data points are laid out in space, while it is possible to provide alternative ways to access individual artist songs.
- Comprehensive metadata on artist level is available via public APIs, unlike the metadata on song level.

[1] http://songza.com

[2] http://habumusic.com

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

[4] http://the.echonest.com

[5] http://www.gracenote.com

[6] http://www.rovicorp.com

Rovi compensates for the fact that different artist songs have different moods by assigning a variety of moods with different weights to artists. Therefore, moods with the greatest weights are the most representative of the artist's music. Given these advantages, the visual model described in this chapter is developed using Rovi database of moods on the artist level.

In order to build an easy to navigate structure around a large vocabulary of mood words (N=289), a data analysis technique is first used to identify relations between artists in a collection, based on the moods that describe them. Correspondence analysis (Appendix A), an exploratory technique, is particularly suitable for this task and results in a graphical representation of relations between artists and moods, and moods themselves. Next, the moods are categorized in order to reveal a hierarchical structure in the visual representation, by building upon the research of Zentner et al. [92] who suggest that music emotions generally fall into 3 main categories and 9 sub-categories, described by 45 emotion words. Numerical methods, including calculation of word similarity in WordNet[7], are used to place each of the 289 Rovi moods into one of the 9 Zentner sub-categories. As a result, the clusters of moods emerge in the space constructed by correspondence analysis. Thus, the music collection can be explored using a top – down, hierarchical approach, by focusing on one of the top 3 categories first, one of the sub-categories next, and finally a specific set of moods.

[7] https://wordnet.princeton.edu

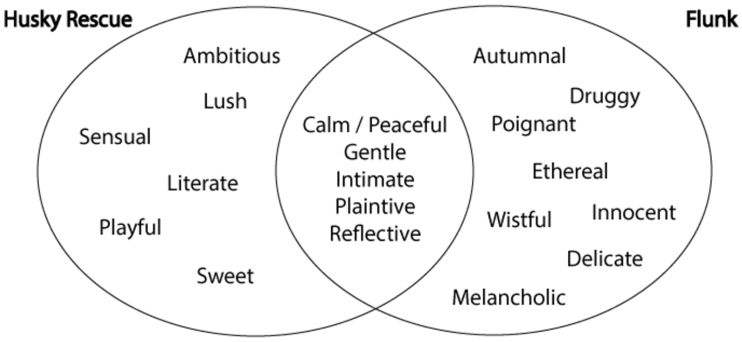


Figure 3.1: Moods associated with artists Husky Rescue and Flunk. Common moods are listed in the intersection of two circles.

3.2 Mapping Artist Mood Metadata to Music Mood Model

Music mood metadata used in this study is obtained via Rovi API. Most of the artists are described by 5-20 different moods which are weighted in order to distinguish between more and less relevant ones. The value of using several different moods to characterize an artist, and a need to create a mood map with dimensions other than valence and arousal can be observed from the following example.

3.2.1 Example Artist Comparison

Figure 3.1 shows moods associated with two artists, *Husky Rescue* and *Flunk*, some of which are common for both. The combination of words gives us the impression of their music, which is more encompassing than if it were based on a single mood descriptor. On one hand, the comparison of two artists positioned in the commonly used Russel's Circumplex model would be based on two criteria only – valence and arousal. On the other hand, by simply glancing over the mood words associated with *Husky Rescue* and *Flunk*, the similarities and differences between the two artists become more vivid and meaningful. For example, both are calm, but one is playful, sensual, lush and the other one melancholic, autumnal and druggy. In order to keep similar level of expressiveness when comparing large number of artists against each other, scales for comparison in addition, or alternative to valence and arousal would be beneficial.

3.2.2 Construction of Visual Mood Map

Pilot study to construct a music related visual mood model was conducted on a set of 3275 artists, randomly chosen from Million Song Dataset[8]. For each one of the artists in the dataset, associated Rovi moods and their weights were collected via Rovi API. As a result, each artist in the dataset is characterized by approximately 5 to 20 weighted mood words and represented with a vector $X \in \mathbb{R}^{289}$, where 289 is the number of unique moods available in the Rovi system. In order to visualize inter-relationships between artists and moods in a two-dimensional space, the first step is to apply a dimension reduction method. Given the categorical nature of the data, eigenanalysis-based ordination approaches (e.g. Principal Component Analysis - PCA, Correspondence Analysis - CA) are more suitable than gradient based (e.g. Multidimensional Scaling - MDS). When choosing be-

[8] http://labrosa.ee.columbia.edu/millionsong

tween commonly used techniques, PCA and CA, we opt for CA which allows us to plot both moods and artists in space based on relative distribution of moods, rather than on the exact mood weights for each artists.

While the resulting 2-D plots of moods and artists positioned in a coordinate system reveal relations between data points, the meaning of these relations is difficult to interpret. The map of 289 mood points is unintelligible and the challenge here is to detect axes or dimensions in it. Therefore, a small subset of 289 moods that could fall on an introvert – extrovert scale was manually chosen with a goal to explore possible clusters or infer the meaning of axes (Figure 3.2). By looking at this subset in the mood space obtained by Correspondence Analysis, it can be noticed that 3 mood groupings emerge, subjectively described as follows: (1) aggressive, hostile (left), (2) mellow, meditative (top right) and (3) uplifting, playful (bottom right). This suggests that relations between large number of moods in the obtained map could be understood by categorizing moods into discrete sets, thus possibly revealing a structure in the mood space.

For the purpose of identifying potential clusters in our mood space, we explore whether our visual map fits into a hierarchical music-specific emotion model proposed by Zentner *et al.* [92]. This model, from now on referred to as *Geneva Emotional Music Scales* or GEMS, consists of 3 main categories (*Vitality*, *Uneasiness*, *Sublimity*), 9 sub-categories and 45 music relevant emotion words distributed across different sub-categories. Figure 3.3 shows top two levels of this emotion

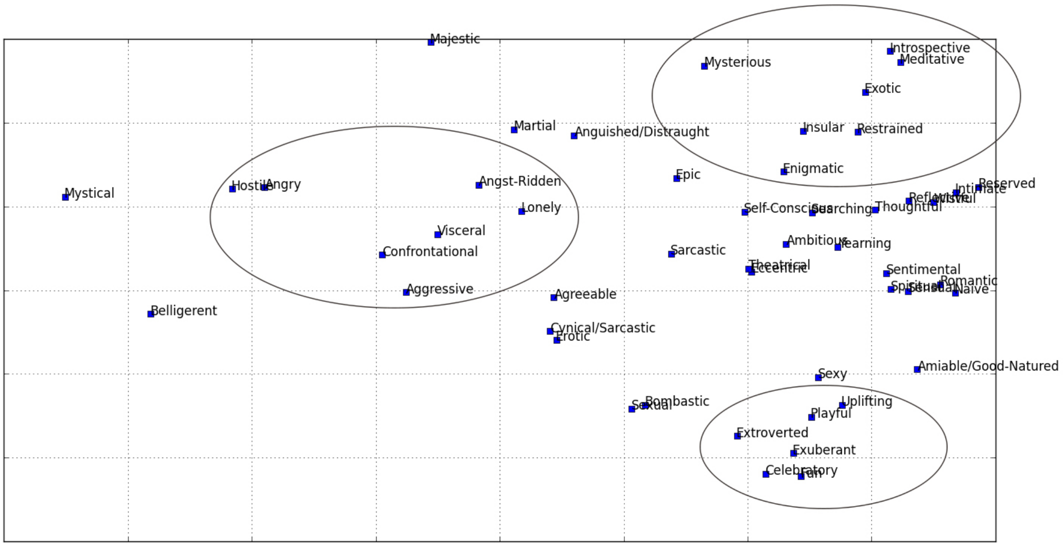


Figure 3.2: Mood map constructed using correspondence analysis. Only moods that could be placed on an introverted - extroverted scale are shown.

hierarchy. Our hypothesis was that such hierarchy should emerge in the visual mood space built upon professionally curated artist-mood associations.

The approach taken to classify Rovi moods was to place each one of them in a GEMS sub-category whose name is the most similar to the given Rovi mood (e.g. mood *exciting* was placed in the sub-category *Joyful Activation*). GEMS sub-categories contain 5 mood words on average, or 45 total. For example, emotions in sub-category *Wonder* are *happy*, *moved*, *allured*, *dazzled* and *amazed*. The computation of similarity between Rovi and GEMS moods was based on WordNet – a large lexical database of English words, grouped into sets of cognitive synonyms. WordNet module developed by Ted Pedersen[9] offers a number of methods for computing semantic similarity and relatedness between words. Three measures

[9] http://www.d.umn.edu/ tpederse/similarity.html

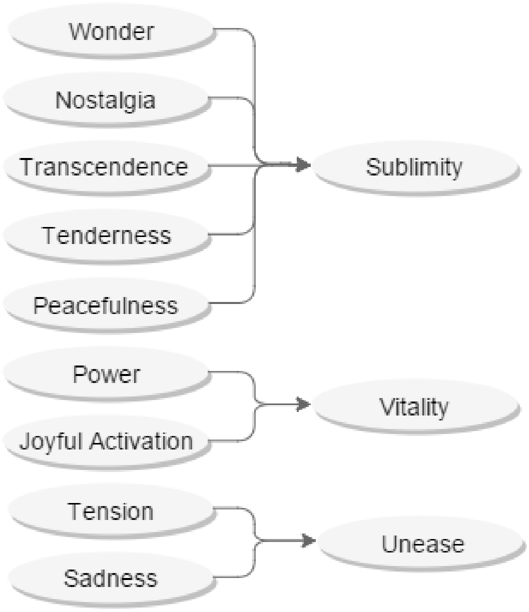


Figure 3.3: Emotion hierarchy proposed by Zentner et al.

in particular are suitable and used in this study for the task of comparing mood words in the form of adjectives: lesk, vector and vector pairs [10]. The main idea behind these three methods is that relatedness of words is derived from the degree of overlap between word definitions i.e. unique representations of the underlying concepts. After calculating pairwise similarity between Rovi and GEMS words using all three measures, the obtained values were normalized using Z-Score. The highest of three similarity values was chosen for each word as a representative measure. Next, selected similarity values between each Rovi mood and Zentner mood words on the lowest hierarchical level were averaged for each sub-category separately, thus resulting in similarity values between each Rovi word and each

[10] http://www.d.umn.edu/ tpederse/Pubs/AAAI04PedersenT.pdf

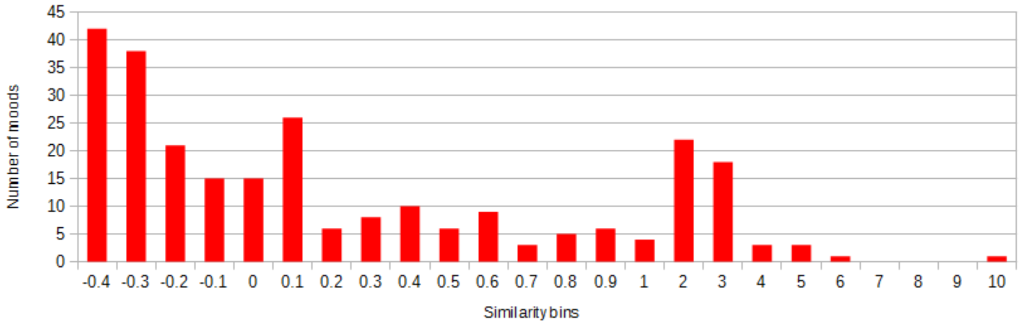


Figure 3.4: Histogram of similarity values between Rovi words and GEMS categories.

GEMS sub-category. Finally, Rovi moods were assigned a sub-category with the highest similarity value.

3.2.3 Subjective Evaluation

The evaluation of the mood classification was performed by subjective observation. More rigorous approach would be beneficial in the future, but the obtained results help lay the ground for the development of visual mood model. By looking at a histogram of final similarity values used to categorize moods (Figure 3.4), we notice that a large number of moods have low values. This means that the likelihood they truly belong to the assigned category is low. Indeed, some Rovi words did not fit into any of the GEMS sub-categories, and many were misclassified.

It is important to note that some of the Rovi moods are less frequently used to describe artists and 23 out of 289 mood words were not associated with any of the artists from our dataset. To verify that these 23 moods are indeed less relevant for describing artists in general, usage frequency was computed for all 289 moods on additional data from Rovi database. Specifically, this additional data was collected from web service AllMusic[11], which is built upon Rovi database and recommends to users the most representative albums and songs for each mood. The number of suggested albums and songs was scraped from the web for all moods and it was found that the average number of suggestions ranges from 1 to 55. The 23 moods not present in our dataset all had the average frequency below 17, and 19 of them (83%) had the frequency below 10. In comparison, 22% of all Rovi moods have the average usage frequency below 17, and 14% the average below 10. Given these numbers, Rovi moods not found in the dataset were discarded for the purpose of placing Rovi moods to Zentner categoris. By subjective assessment, it is estimated that only 35% of the 266 remaining moods were correctly categorized.

Although WordNet is the most comprehensive database of English words, with the greatest number of tools available for analysis, calculated similarities between words are based on their relatedness, and not strictly on synonymity. This has undesirable consequences to the outcome of mood classification. To illustrate with an example, word *volatile* was found to be more closely related to word *tender* than *tense*. As a result, it was placed to category *Tenderness* rather than *Tension*. Because of this, the categorization of moods using WordNet provides a good basis for further research, but it should not be considered conclusive. Therefore, misclassified words were subjectively assigned categories they are more likely to belong to. During this process it was found that 66 moods cannot be placed into any of the

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

GEMS sub-categories, and due to the scarcity of research in classification of music related moods, sub-categories were expanded to encompass uncategorized words. Although some of the misclassified words may be descriptive accounts of music rather than moods, the rational for expansion of GEMS sub-categories was to remain consistent with the data source, used by other researchers as well [75, 50]. Specifically, it was found that out of 66 moods, 28 describe the feeling of unease, and the remaining 38 pertain more to style of expression than to moods. Therefore, top category *Unease*, containing sub-categories *Tension* and *Sadness*, was supplemented with *Fear*, *Lethargy* and *Repulsiveness.* Descriptors that carry less or no mood content (e.g. *quirky*, *knotty*, *elegant*) were placed into new, distinct, top category *Other*, divided into sub-categories *Stylistic*, *Cerebral* and *Mechanical.* The final breakdown of moods across different categories is summarized in Table 3.1.

The categorization of Rovi moods in the 2D space obtained by CA can be observed in Figure 3.5. Three clusters indeed emerge, corresponding to three GEMS top categories: *Sublimity* (green), *Vitality* (red) and *Uneasy* (brown). Although there is slight overlap between the three clusters, the categorization offers a good foundation for building mood based visualization of music collection. Moods placed to the category *Other* (purple) are located mostly in the center of visualization – they spread over *Sublimity* and *Vitality*, and to lesser extent over *Unease.* This indicates that they are often used to describe artists along with more clearly defined moods in GEMS categories. Having constructed a hierarchical, visual

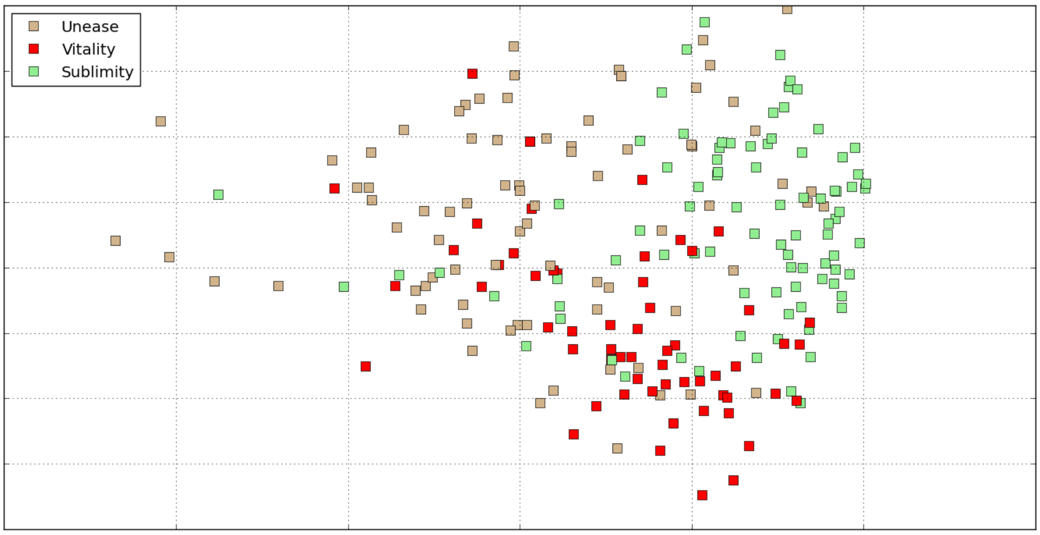


Figure 3.5: Map of mood space obtained using correspondence analysis. Moods are grouped in categories Sublimity (green), Vitality (red) and Unease (brown).

Category	Sub-category	No. of moods	Example moods	Total
Sublimity	Tenderness	24	Delicate, romantic, sweet	89
	Peacefulness	22	Pastoral, relaxed, soothing	
	Wonder	24	Happy, light, springlike	
	Nostalgic	9	Dreamy, rustic, yearning	
	Transcendence	10	Atmospheric, spiritual, uplifting	
Vitality	Power	29	Ambitious, fierce, pulsing, intense	61
	Joyful activation	32	Animated, fun, playful, exciting	
Unease	Tension	32	Nervous, harsh, rowdy, rebellious	78
	Sadness	18	Austere, bittersweet, gloomy, tragic	
	Fear *	10	Spooky, nihilistic, ominous	
	Lethargy *	8	Languid, druggy, hypnotic	
	Repulsiveness *	10	Greasy, sleazy, trashy, irreverent	
Other *	Stylistic *	19	Graceful, slick, elegant, elaborate	38
	Cerebral *	12	Detached, street-smart, ironic	
	Mechanical *	7	Crunchy, complex, knotty	
			Number of categorized moods	266

Table 3.1: Structure and description of *MoodPlay* mood hierarchy. Categories and sub-categories marked with * are the expansions of the original GEMS model.

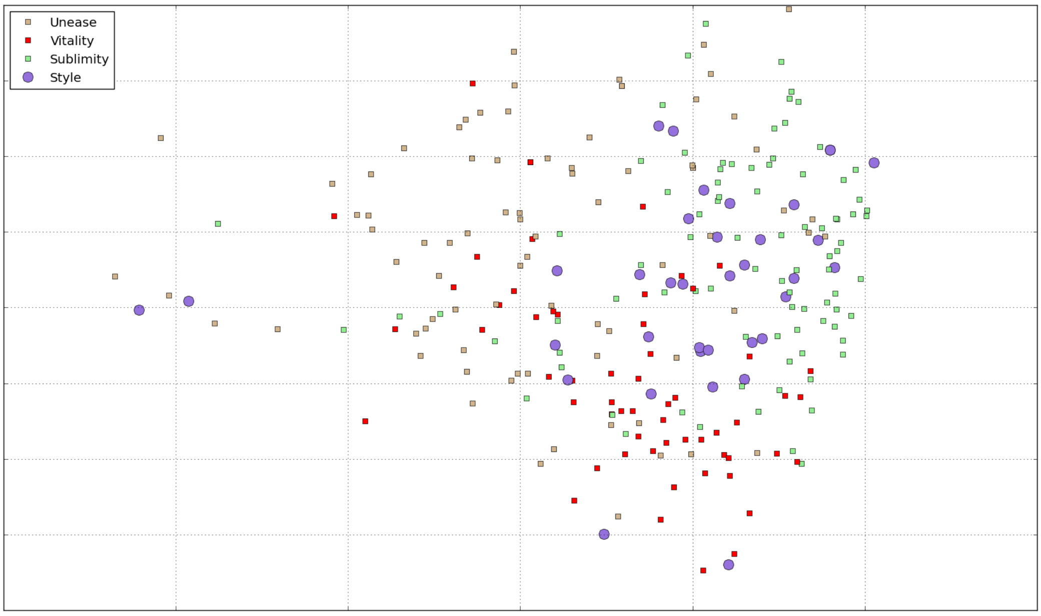


Figure 3.6: Mood map depicting moods in three GEMS categories (green, red, brown) and a new category Other (purple).

mood space, and shown that the results of our mood meta-data analysis align with those from physiological studies, following chapters document the research in the domain of interactive music recommendation.

Finally, as a demonstration of how artists can be compared across Zentner categories, Table 3.2 lists categorized moods associated with two artists Husky Rescue and Flunk, previously shown in Figure 3.1. Rows that correspond to sub-categories unique to either of the artists are highlighted in different colors. It can be observed that *Husky Rescue* is more filled with Wonder, Power and Joyful Activation, while *Flunk* is more Nostalgic and Sad. If we compare these two artist on the level of top-categories, we see the main difference is that *Husky Rescue* is considered more Powerful and *Flunk* is more Uneasy.

GEMS Category	*Husky Rescue* moods	*Flunk* moods
Wonder	Playful, literate	
Transcendence	Lush	Ethereal
Tenderness	Sensual	Innocent, delicate
Peacefulness		
Nostalgia		Wistful
Power	Ambitious	
Joyful Activation	Sweet	
Tension		
Sadness		Autumnal, melancholy, poignant, druggy

Table 3.2: Comparison of artists *Husky Rescue* and *Flunk* across nine GEMS categories. First column contains mood categories, while second and third contain moods associated with two artists. Rows in color highlight mood categories that distinguish between the two artists.

Chapter 4

MoodPlay: Interactive Music Recommender

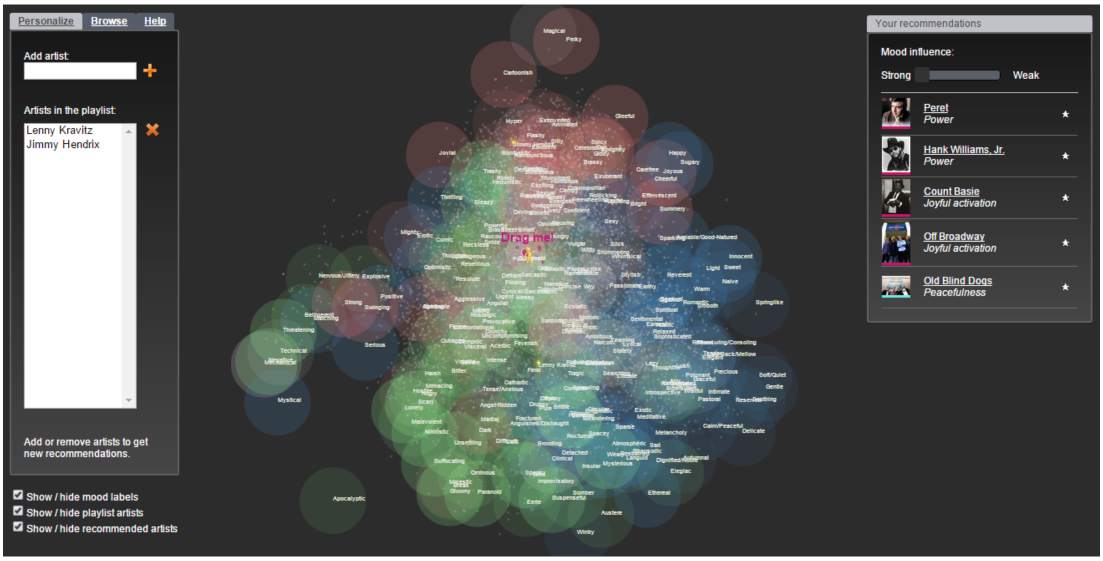
Previous chapter introduced a novel, visual mood model, targeted towards music specific moods. This chapter answers second research question of the thesis: *How can interaction, explanation and control be supported over a visualization of artist - mood associations?* The question is answered by designing and developing an interactive, mood based music recommender – *MoodPlay*, with the visual mood model as a central component.

4.1 System Overview

MoodPlay is a web application, currently available at `http://haze.mat.ucsb.edu/~ivana/recsys`. It was developed in two stages. A preliminary user study was conducted using the first version (Chapter 6), and more thorough user study (Chapter 7) was conducted using the second, slightly improved version. Both versions have the same recommendation algorithm and similar interface functionality. The main difference between the two is in the visual representation of mood hierarchy and the graphical design. This section describes a typical use-case and interface design decisions, with reference to the final version of the system.

4.1.1 Use Case

In a typical use-case, users enter the system via web browser, using a computer or mobile device. At the beginning of the session they are given step by step usage instructions and explanations of the numerous interface features. The instructions can be accessed again any time during the session. Once the system is loaded, users are presented with a three panel view, as shown in Figure 4.1, which reflects three main operating phases: (1) profile building, (2) positioning the user within a visual mood space, with the goal to facilitate transparency and control over recommendations and (3) suggesting new artists using hybrid recommendations over mood and audio content. A user can enter profile items on the left panel via predictive text list that appears when any input is given. With every such profile update, the system determines the overall mood associated with given artists and instantly (re)positions the user avatar in the mood space visualization, shown in the center panel. Simultaneously, the system updates a list of recommendations,



Lenny Kravitz
Jimmy Hendrix

Add or remove artists to get new recommendations.

☑ Show / hide mood labels
☑ Show / hide playlist artists
☑ Show / hide recommended artists

Peret
Power

Hank Williams, Jr.
Power

Count Basie
Joyful activation

Off Broadway
Joyful activation

Old Blind Dogs
Peacefulness

Figure 4.1: Screenshot of MoodPlay interface, divided into three sections: (left) pane for generating user profile by entering artist names, (center) mood-space visualization, (right) recommendation list, along with slider for adjusting mood influence

shown in the right panel, based on a hybrid model of mood and audio content. Artist data points are positioned within the mood space visualization. User can explore the music collection by zooming, panning, clicking on artist nodes and streaming music in real time. Furthermore, user can manually override their position in the mood-space by moving the avatar. Provenance of previous positions is maintained through a mechanism of interactive trails that can be modified or removed if desired. The user can also control the ratio of audio content to mood component in the hybrid recommendation algorithm via a slider that controls a dynamic radius around the current avatar position.

4.2 User Interface

Visualization of mood space and artist positions within it is central to solving the problem of navigation through music collection and explanation of recommendations. Hence, it occupies the largest portion of the interface. The mood space contains 266 moods visible to user, with similar moods appearing closer to each other than dissimilar ones. In order to help the user understand the mood space and navigate to desired areas, moods are organized in a hierarchy, having three primary categories at the top - *Vital*, *Sublime* and *Uneasy* (Subsection 3.2.2). This is portrayed on canvas by showing mood nodes in different categories in red, blue and yellow colors respectively. Mood nodes are semi-transparent, their size is equal and purposefully large enough to cause overlap. This produces an interplay of colors, thus forming the space with gradual transitions between mood categories. Artists from the MoodPlay database are placed within the mood space based on positions computed along multiple mood dimensions.

As described in Subsection 3.2.2, positions of moods and artists in space are computed using Correspondence Analysis. However, data points obtained as a result of the analysis were positively skewed for the final set of 4927 artists, more so than the data points shown in Figure 3.6. For aesthetic reasons the data was transformed to alleviate the skewness following the guidelines of [80] and [36]. The transformation changes the distances between the data points but preserves the order. Therefore, user's understanding of relations between artists and moods, and the music recommendation algorithm were not negatively affected by the transformation. Specific steps taken are as follows:

1. x coordinates were transformed using the formula:

$$x_{new} = \log_{10}(x_{old} + c) \tag{4.1}$$

   where c equals to the absolute of minimum x_{old} value, in this case 1.5.

2. Both x and y coordinates were transformed to range [-2, 2]

3. Sign was changed for y values

Finally, via the interface users can stream their music in real-time and see additional artist information by clicking on the nodes, which are grey in color and significantly smaller than mood nodes. In addition to common-place zooming and panning actions in the visualization, interaction in mood-artist space is supported by the features described in following sub-sections.

4.2.1 Mood Filtering

Mood hierarchy has been constructed to enable finding desired moods in a visual map in a more efficient way than browsing a long list of mood words. Such a design supports exploring music collection in a top-down manner, starting from a top mood category (vital, sublime or uneasy) and then narrowing down to sub-categories and specific moods. Figure 4.2 shows four different views of the

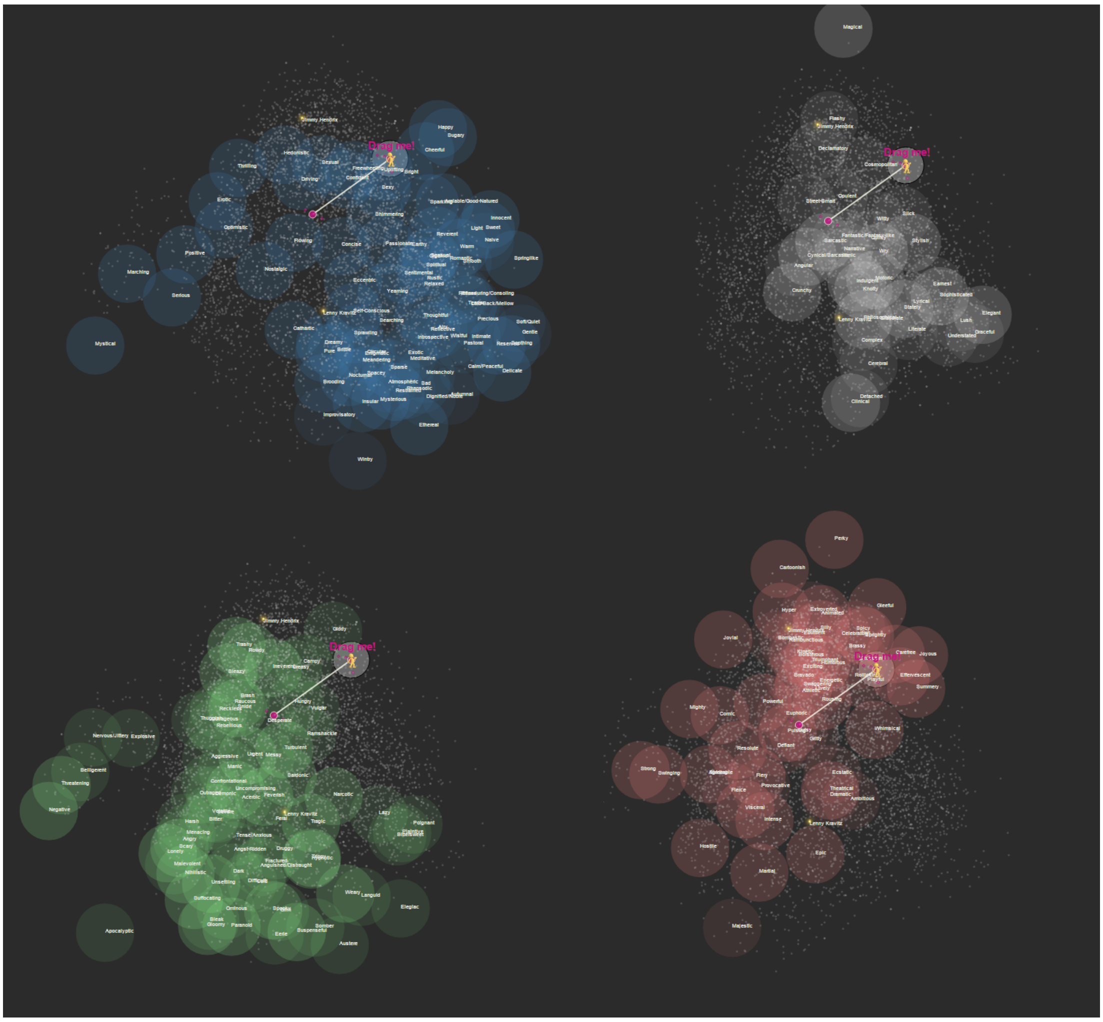


Figure 4.2: Four views of mood space filtered by top categories: sublime moods (top left), unclassified (top right), uneasy (bottom left), vital (bottom-right)

mood-artist space, filtered by top category. Furtermore, filtering by sub-categories is exemplified in Figure 4.3 for *Uneasy* moods.

4.2.2 Dynamic Labeling

Showing labels for all visible moods in the mood space on lower zoom levels (maximum 266) causes cognitive overload and affects user's ability to find desired

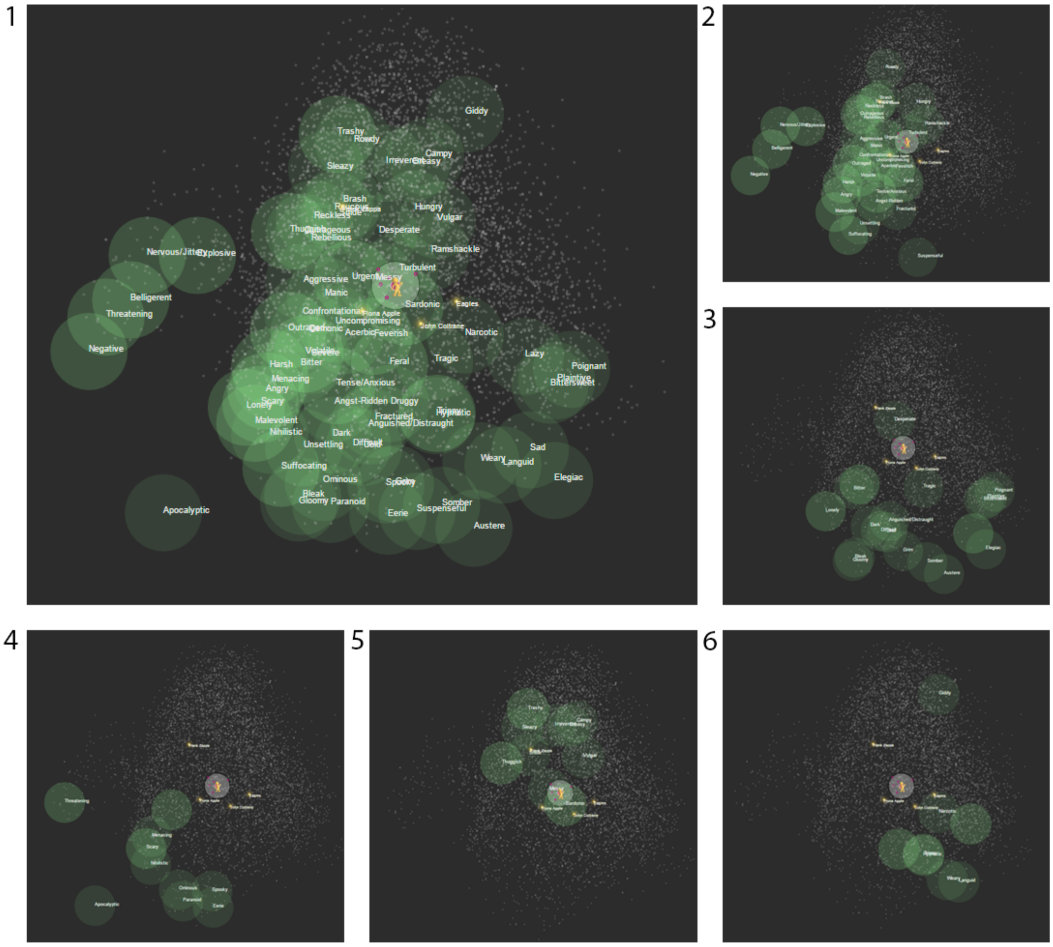


Figure 4.3: Four views of mood space filtered by top categories: sublime moods (top left), unclassified (top right), uneasy (bottom left), vital (bottom-right)

items. This problem is addressed by dynamically displaying labels for a subset of moods, deemed to be dominant in the visible area. Mood dominance is directly proportional to the number of artists tagged by it and the weights assigned. More specifically, dominance based ranked list was computed by summing up weights for individual moods across all artists, and then ordering them ascendingly. On each zoom level the system shows labels for up to 36 moods total, or 9 most prominent moods per category (number of labels is chosen by informal experimentation). This significantly unclutters the space at lower zoom levels when most of the moods space is visible. At the highest zoom level less than 36 moods are visible, and in that case all of the labels are displayed.

4.2.3 Artist Information

Being able to hear artists' music is a necessary component of a music recommendation system. Upon clicking on the artist node in the visualization, an info box is displayed, showing artist picture, link to Last.fm profile, and an audio play button (Figure 4.4). Artist list in the recommendation pane contains the same elements. Clicking the play button starts streaming of 30 seconds excerpts of songs from a randomly selected album.

Ordered list of recommended artists is displayed in the right-most panel (Figure 4.1-right) and the corresponding artist nodes are highlighted in the mood space. Items in the recommendation list are linked to audio streams and to Last.fm profiles of artists. For each recommended artist the system also displays artist's

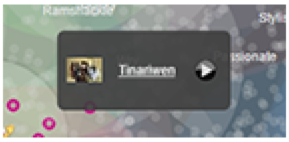


Figure 4.4: Artist information box, displayed upon clicking on an artist node in the visualization.

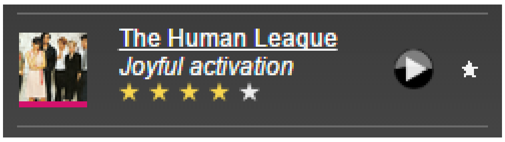


Figure 4.5: Item in the recommendation list, containing artist information.

picture, color of the top mood category (red, blue or yellow) and the name of the sub-category the artist belongs to, with the goal to help users gain some understanding of the music upon visual inspection (Figure 4.5). As users interact with the system, recommendation list is being updated. Up and down arrows next to the artist name inform the user that the item changed the position after the interaction, while star means the item is new. Rating of recommended items is enabled for the purpose of user study, and is achieved by clicking on one of the five stars below artist names.

4.2.4 Trail of User Mood Preferences

Adaptivity of music recommenders is particularly important due to the dynamic nature of listening context [79]. Keeping this in mind, the gradual change of user's preference is modeled by enabling the movement of avatar in latent mood space and maintaining the array of trail marks, weighted by distance from the current position (Figure 4.6). As user navigates away from the initial position, the mood information associated with each trail mark is incorporated into the recommendation algorithm. Trails can be modified or deleted entirely.

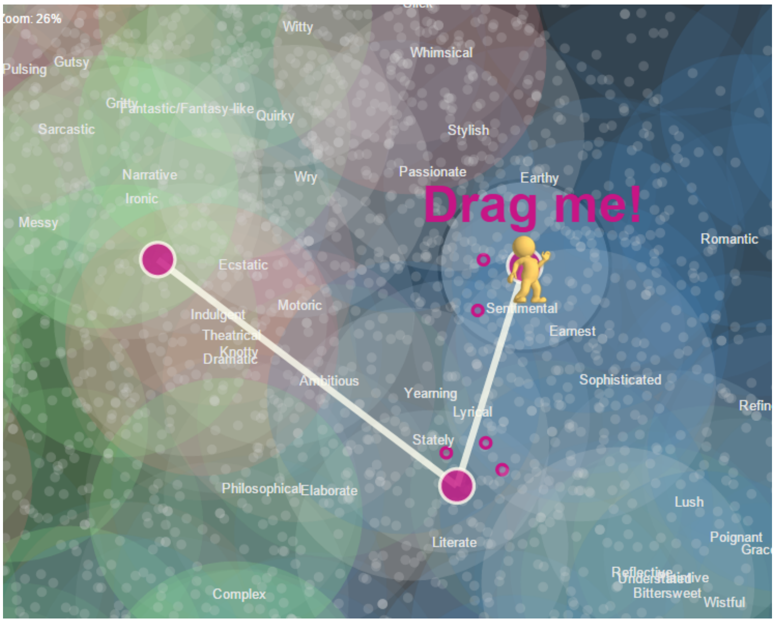
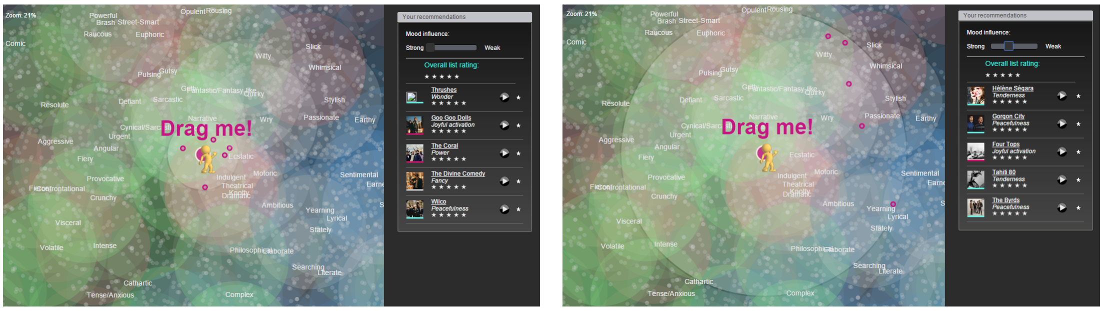


Figure 4.6: User trail containing three marks - one at the original user position calculated based on the profile, and two at new positions selected by user. Recommended artists are highlighted in pink.

4.2.5 Controlling Mood Influence

Finally, fine-tuning of recommendations is further supported by controlling the hybridization of recommendation process. The recommendation approach accounts for the fact that mood based similarity between artists does not necessarily match audio based similarity (e.g. techno and punk artists are both energetic, but they do not sound similar). Rather than using fixed, predefined settings, we allow user to adjust the mood influence in the recommendation algorithm via a slider control. The weaker the mood influence, the more we rely on audio similarity to calculate recommendations, and vice-versa. This is visualized though a novel radius-based interaction that dynamically re-sizes a catchment area around the current avatar position, as shown in Figure 4.7-left and 4.7-right respectively.


Your recommendations

Mood influence: Strong — Weak

Overall list rating: ★ ★ ★ ★ ★

Thrushes
Wonder
★ ★ ★ ★ ★

Goo Goo Dolls
Joyful activation
★ ★ ★ ★ ★

The Coral
Power
★ ★ ★ ★ ★

The Divine Comedy
Fancy
★ ★ ★ ★ ★

Wilco
Peacefulness
★ ★ ★ ★ ★

Your recommendations

Mood influence: Strong — Weak

Overall list rating: ★ ★ ★ ★ ★

Hélène Ségara
Tenderness
★ ★ ★ ★ ★

Gorgon City
Peacefulness
★ ★ ★ ★ ★

Four Tops
Joyful activation
★ ★ ★ ★ ★

Tahiti 80
Tenderness
★ ★ ★ ★ ★

The Byrds
Peacefulness
★ ★ ★ ★ ★

Figure 4.7: The effect of adjusting mood influence using a slider control. Left - narrow mood selection, Right - wide mood selection

4.3 System Implementation

Client side of the MoodPlay web application was developed using JavaScript and D3.js library for visualization and canvas interaction (e.g. dragging, zooming, filtering). User data and most of the artist data are stored in a Mongo database on the application server. The application also interacts with several external servers that host public music databases, in real time, in order to pull additional artist information per user request.

4.3.1 System Architecture

MoodPlay uses diverse metadata which is collected from different sources, mostly through public Web APIs, and therefore requires a special architectural design. In addition, recommendations have to be computed very quickly, since many types of user interactions refresh the recommendation list several times per session.

The system architecture has two main components as depicted in Figure 4.8: one for building the library of items with their metadata (*Dataset Construction*) and a second component that generates user recommendations (*Recommendation framework*). Following subsections describe the architecture design and implementation in detail.

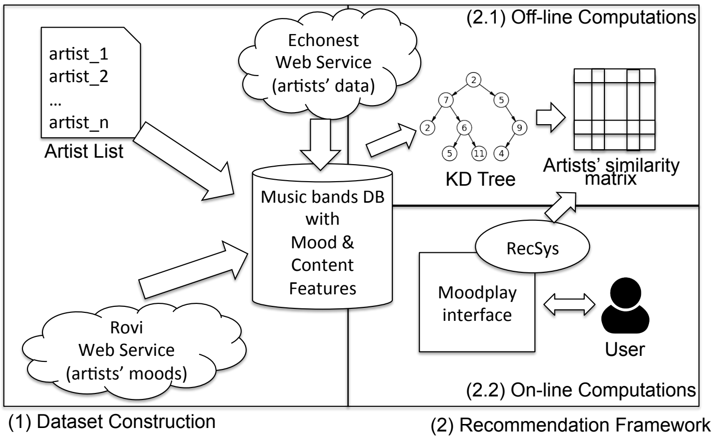


Figure 4.8: MoodPlay system architecture indicating the modules for: (1) dataset construction and (2) recommendation, which is divided in the component for (2.1) off-line computation and (2.2) online computations made at the moment the user interact with the system.

4.3.2 Dataset and Data Sources

MoodPlay relies on a static music dataset of 4927 artists, obtained in several iterations, which can be seen at the center of Figure 4.8. First, 3275 artists were randomly selected from a subset of the Million Songs Dataset [1]. Artists ranged from very popular to less known, and played music in a variety of genres and over different decades. The pool was then expanded by 2,000 most *familiar* and

[1] http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset#subset

hotttest artists from the public EchoNest[2] database. *Familiarity* and *hotttness* metrics are EchoNest numerical estimations of how known and popular given artist is in the world. The decision to complement the initial set of randomly chosen artists with the popular ones was made to better facilitate an online user study with participants of different ages from different parts of the world. Finally, artists for which we were not able to obtain mood or song data were discarded.

In addition to the list of artists, we needed to collect metadata since the computation of the visual space and recommendations is based on mood and audio data associated with artists. Mood data for each artist was obtained via Rovi[3] API. Artist positions in the mood space were pre-computed (Section3.2) and are loaded from database at application start time. Furthermore, top ten most popular songs for each artist and corresponding audio analysis data was obtained from EchoNest[4]. Different interpretations of the same song, having an exact same title in EchoNest database were discarded. Music streaming in MoodPlay is accomplished by sending a request to Rdio[5] API with the Rdio artist ID as a parameter. Finally, MoodPlay interface offers artist picture and link to an external profile, both obtained from Last.fm[6]. Description and sources of artist data are summarized in table 4.1.

[2] http://developer.echonest.com/docs/v4/artist.html
[3] http://developer.rovicorp.com/io-docs
[4] http://developer.echonest.com/docs/v4/song.html
[5] http://www.rdio.com/developers
[6] http://www.last.fm/api/intro

Field name	Description	Source
ID	Artist ID	-
Name	Artist name	-
Moods	Weighted moods associated with artist	Rovi
Position	Artist position in mood space	Computed based on mood data
Top category	Top mood category calculated based on assigned moods	Computed based on mood data
Sub-category	Mood sub-category calculated based on assigned moods	Computed based on mood data
Songs	Audio analysis of ten most popular songs	EchoNest
Picture	Artist's picture	Last.fm
External profile	Link to artist's profile on Last.fm	Last.fm
Streaming data	Rdio album ID used for streaming music	Rdio

Table 4.1: MoodPlay artist data

4.4 Recommendation Approaches

The main database of our system stores artists' information with audio content and mood features, and we use both categories of features to produce recommendations in two steps: off-line computation of artists' similarity based on audio data, and online computation of recommendations based on a cascading process of mood and audio content filtering. We take such hybrid approach for suggesting new artists in order to alleviate disadvantages of individual algorithms.

Online component returns the recommendations to the user while interacting with the interface. This component is a hybrid cascading recommender [14], diagrammed in Figure 4.9, which operates in two stages: (1) using the user profile as an input, our system produces a first candidate set of recommendations based on mood similarity, and (2) the output of the first recommender is the input to an audio content-based recommender, which re-ranks the artists and produces the final recommendation list. This layered approach, along with the related interface

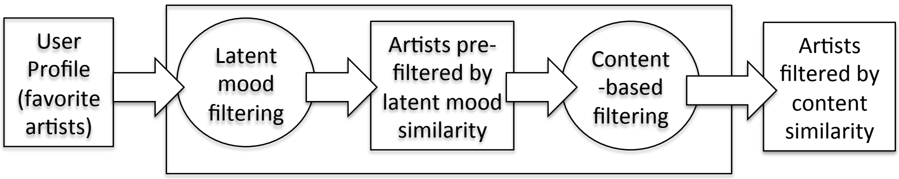

User Profile (favorite artists)
Latent mood filtering
Artists pre-filtered by latent mood similarity
Content -based filtering
Artists filtered by content similarity

Figure 4.9: Schematic representation of our hybrid, cascading recommender, which pre-filters artists based on mood similarity and then post-filters based on audio content.

components, supports our goal to help user understand how recommendations are generated while navigating mood space. Next, we describe the algorithms in more detail.

Offline computation of artist similarity. Artists' pairwise similarity, based on mood and audio content, is calculated offline and stored in two separate data structures. Mood based similarity between any two artists is a function of their Euclidean distance in the mood space (produced by Correspondence analysis, see section 3.2.2). To find audio based similarity, we obtain audio analysis data for representative songs of each artists and use a variance of nearest-neighbor retrieval algorithm to generate artist similarity map.

Top 10 most popular songs for each artist in our database were identified via Echonest API. Audio analysis data for resulting 49,270 songs was obtained from the same source. We used timbre, tempo, loudness and key confidence attributes, which amounted to approximately 10,000 numerical values per song. In order to make the similarity calculations efficient, we represent each song with a vector $v_i \in \mathbb{R}^{515}$ [55] and build artist data into a KD-tree [7]. Finally, an accelerated approach for nearest-neighbor retrieval that uses maximum-variance KD-tree data structure was used to compute similarity between songs, since it is has a good balance of accuracy, scale and efficiency [55]. In this way, time complexity of constructing a similarity matrix was reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n \log n)$, while the search for the K nearest neighbors of a given artist is reduced from $\mathcal{O}(K \cdot n)$ to $\mathcal{O}(K \cdot \log n)$.

Specifically, to find a ranked list of similar artists for a given artist, first, for each one of the artist songs (10 total) we rank all other songs from the dataset from most to least similar. Next, we get song authors and calculate average similarity rank for each one. Finally, we obtain a similarity matrix by repeating these steps for each one of 4,927 artists [17]. Details of the algorithm are in Algorithm 1.

Online recommendation. During a user session, MoodPlay recommends new artists similar to the artists the user enters into the system. First step in the process is to determine a user's overall mood based on the profile. This is achieved by locating given artists in the mood space and calculating the centroid of their positions, where we then position the user. Artists found in the surrounding area are all potential candidates for recommendation because they are considered to reflect moods derived from the user's input. The size of the surrounding area is set to a default value, and expandable by user. Among potential candidates, we select ten most similar to the user profile based on pre-computed audio similarity data, order them by distance from user position and display as recommended artists (Algorithm 2).

Algorithm 1 Algorithm for computation of audio similarity

```
Input:
    Set of artists: $A = \{a_1, a_2, ..., a_n\}$
    Set of songs for all artists: $S = \{Sa_1 \cup Sa_2 \cup ... \cup Sa_n\}$
Output: Audio similarity ranks: $ARanks = \{a_i \rightarrow \{a_j \rightarrow rank_{ij}\}\}$
function ComputeAudioSimilarityRanks
    ARanks = {}                                    ▷ dictionary of artist similarity ranks
    for each artist $a_i$ in $A$ do
        SRanks = {}                                ▷ dictionary of song similarity ranks
        for each song $s_k$ in $Sa_i$ do
            SRanks[$s_k$] = ComputeSimilarityMapOfSongRanks($s_k, S$)
        end for
        for each artist $a_j$ in $A$ do
            ARanks[$a_i$][$a_j$] = ComputeAverageSongSimilarity($SRanks$, $Sa_j$)
        end for
    end for
    return ARanks
end function

function ComputeSimilarityMapOfSongRanks($s, S$)
    Rank all songs from $S$ based on audio similarity to song $s$
    for each $s_j$ in $S$ do
        similarityMapOfSongRanks[$s_j$] = $rank_j$
    end for
    return $similarityMapOfSongRanks$
end function

function ComputeAverageSongSimilarity($SRanks$, $Sa$)
    average = 0
    for each song $s_i$ in $Sa$ do
        for each song $s_j$ in $SRanks.keys$ do
            average += $SRanks[s_j][s_i]$
        end for
    end for
    average = average / ($Sa.size$ + $SRanks.size$)
    return average
end function
```

Furthermore, we allow user to move in the mood space and thus obtain new recommendations that account for the mood change. We propose a novel, adaptive, recommendation approach, where we keep track of each new position and apply a decay function to the preference trail when recommending new artists. Assuming that the latest user position is the most relevant to user at a given moment, we assign the greatest weight to the recommendations from the latest trail mark. The weights decrease as a function of hop distance from the end of the trail. Pseudocode for the trail based recommendation algorithm is given in Algorithm 3, and here we outline the steps:

1. At each trail mark, calculate ten recommended artists using the hybrid method described above.

2. At each trail mark, scale distances between chosen artists and the trail mark as if the recommendation radius was minimal. We do this because radius can vary among trail marks. To produce the final recommendation list we sort all potential candidates based on the distance from their respective mark. Distances are scaled down to ensure the items are not too far from the respective trail mark.

3. For each artist, at trail mark i ($i = 0$: initial user position) the adjusted distance:

   $$D_a = D + \Delta \times (n - 1 - i) \tag{4.2}$$

   where D_a stands for adjusted distance, D is the original distance, n is the total number of trail marks and i is the trail mark number. We obtain empirically the weight constant by using the following formula:

   $$\Delta = r_{min}/4 \tag{4.3}$$

where r_{min} is the minimal recommendation radius. The higher the Δ value, the steeper the decay function is.

4. Combine all potential recommendation items from each trail mark (10 per trail mark) and sort them by adjusted distances. Pick top 5 artists from this sorted list, and recommend them to user.

Algorithm 2 Mood and Content-based Hybrid Recommendation

```
Input:
    Artists in user profile: $P = \{a_1, ..., a_n\}$
    User position: Profile based position $u = Centroid(a_1, ..., a_n), a_i \in P$ or a position from
    user's trail $u \in T = \{u_1, ..., u_n\}$
    Recommendation radius: r
    Audio similarity ranks: $ARanks = \{a_i \rightarrow \{a_j \rightarrow rank_{ij}\}\}$
    Number of recommendations: $n_{rec}$
Output:
    Recommended artists: $R = \{a_1, ..., a_n\}$
function RecommendMusic($u$)
    M = []                                          ▷ artists within mood radius
    for $a_i$ in $A - P$ do
        if distance($a_i, u$) < r then M[$i$] = $a_i$
        end if
    end for
    H = {}                                          ▷ dict. of artists & similarity with $P$
    for $a_i$ in $M$ do
        H[$a_i$] = AverageSimRanking($a_i$, $P$)
    end for
    sort(H)                                         ▷ sort artists by audio similarity
    R = H[1..$n_{rec}$]
    return $R$
end function

function AverageSimRanking($a, P$)
    average = 0
    for each $a_i$ in $P$ do
        average += $ARanks[a][a_i]$
    end for
    return average $\div P.size$
end function
```

Algorithm 3 Hybrid Recommendation with Provenance Trails.

```
Input:
    Trail of user positions: $T = \{u_1, u_2, ..., u_n\}$, where $u_1$ is profile based position and consecutive
    $u_i$ are positions that user navigated to
    Current recommendation radius: $r$
    Minimum recommendation radius: $r_{min}$
    Number of recommendations: $n_{rec}$
Output:
    Recommended artists: $R = \{a_1, ..., a_n\}$
function RecommendMusicBasedOnTrail
    R = {}                                              ▷ dict. of recommended artists
    $\Delta = r_{min}/4$
    for $u_i$ in $T$ do
        for $a_j$ in RecommendMusic($u_i$) do
            $d_s$ = Scale(distance($u_i, a_j$), $r$, $r_{min}$)
            $d_a = d_s + \Delta \times (T.size - 1 - i)$
            R[$a_j$] = $d_a$
        end for
    end for
    sort(R)                                             ▷ sort artists in R by $d_a$
    return R[1..$n_{rec}$]
end function

function Scale($d, r, r_{min}$)
    $d_c$ = Convert $d$ from range [0, $r$] to [0, $r_{min}$]
    return $d_c$
end function
```

Chapter 5

Evaluation

Evaluating recommender systems that contain interactive components is particularly challenging because of complex and potentially diverse interplay between the human participant and the automated algorithm. While the longitudinal study with real world users would be ideal, we believe that an crowdsourced study using Amazon Mechanical Turk[1] (MT) is a reasonable surrogate. There is a large body of research showing that a crowdsourcing platform like MT or CrowdFlower[2] can provide concrete and reliable results when evaluating higher complexity tools like MoodPlay. Many recent papers at top scientific conferences use MT to evaluate systems far beyond basic aspects [46, 11, 72, 43] and detailed comparison of experimental research in the lab setting versus using MT is presented in [9, 60, 13]. The main advantages of using MT to analyze different interaction patterns with the MoodPlay recommender are as follows:

[1] https://www.mturk.com

[2] http://www.crowdflower.com/

Effortless access to a diverse population sample. Participants live in different countries, belong to different age groups and have different education backgrounds.

Scalability. Depending on the complexity of a study and compensation, it is reasonable to expect over 50 people participating in an experiment in one day.

On the other hand, following are some of the limitations and the approaches taken to minimize negative effects:

Lack of control. Comparing to lab studies, during MT studies there is significantly less control over participants environments. However, we minimized the negative consequences by filtering out the participants that did not meet threshold interaction and attention requirements.

Low motivation. As reported in [60], MT participants are less motivated than those completing lab studies. Nevertheless, in our study participants spent 20 minutes on average and left very detailed, thoughtful text comments and feedback, showing that they were mentally engaged with the system.

5.1 Hypotheses

MoodPlay system was developed in two phases and evaluated through two user studies having a similar format. The specific setup details and results of both studies are reported in separate sections, but first we describe the hypothesis, common structure of the experiments and measures. Previous chapters presented the development of music specific, visual mood model and an interactive music recommender. The studies detailed in this chapter were designed to address research questions pertaining to the proposed visualization and recommendation system.

Research question: What are the effects of proposed interactive visualizations on the user experience with a recommender system?

Hypothesis: The expectation is that the interactive, controllable interface will improve several aspects of user experience. For example, the ability to navigate the mood space and to control the recommendation process should help users to find desired artists and increase recommendation accuracy. The number of interactive features introduces significant complexity and the acceptance among users is difficult to predict. Yet, the experiments measure the effect of complexity on users' experience. Overall, the expectation is that the proposed interface would make music listening and discovery more enjoyable than music recommendations in the form of ordered lists.

Research question: How does knowledge of and interaction with mood metadata influence recommendation accuracy and user experience?

Hypothesis: The explanation of moods related to profile artists and recommendations should increase users' high-level understanding of the underlying recommendation method. The interaction with mood metadata should further support this understanding and help user guide the recommendation system in a desired direction. It is expected that this will result in increased trust and recommendation accuracy. Finally, interaction with mood metadata should also increase the recommendation diversity, as users are able to explore the music collection through the mood space.

5.2 Study Setup

Participants accepted the study on Mechanical Turk and were redirected to a Qualtrics[3] pre-study survey with demographic and propensity related questions (Appendix B). Following this, they were assigned to one of the four random experimental conditions, as shown in Table 5.1, and performed the main task. Finally, participants gave qualitative feedback in a post-study survey, also administered through the Qualtrics platform.

The conditions have increasing visual and interaction complexity. Condition 1, that allows users only to enter profile items and see an ordered list of recommended items serves as a benchmark, against which other, more complex conditions are compared. Condition 2 is also based on a preexisting user profile, but displays the static visualization of mood space and highlights the recommendations within it. Conditions 3 and 4, have the same features as previous conditions, but also allow for user input to the algorithm at recommendation time through interaction with

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

	Condition			
Feature	1	2	3	4
Profile generation	x	x	x	x
Ordered list of recommendations	x	x	x	x
Display of latent mood space		x	x	x
Navigation in latent mood space			x	x
Hybridization control			x	x
Trail based recommendations				x

Table 5.1: Availability of different features per experimental condition. Last row in the table shows number of valid subjects in each condition.

the mood visualization. Figure 4.1 shows the full system, as tested in condition 4.

During the main task, participants were given step by step instructions in the form of interactive MoodPlay system tutorial. They were asked to enter at least three profile items (music bands) from a drop-down list. In all conditions, this profile was used to generate a list of 5 recommendations, that were shown on the right side of the screen. Ratings were collected for the recommendation list as a whole and for individual items in the list. Participants were then allowed to interact freely with the system and generate as many intermediate recommendation lists as they wished. Once satisfied, they again rated the full list of items prior to finishing the MoodPlay interaction task. Finally, participants gave qualitative feedback in a post-study survey, also administered through the Qualtrics platform.

The effect of interactive visualizations and knowledge of mood metadata on user experience was evaluated using objective and subjective measures. Following lists give an overview of measures, while the detailed lists of actions tracked and pre-survey and post-survey questions can be found in Appendix ??.

Objective measures are mostly derived from tracking user interactions with the interface:

- List of artists added and removed from the profile
- Ratings of recommended items and recommendation list as a whole (predictive accuracy of the recommendation algorithm)
- List of artists that user played during the session
- List of artists user clicked on to see their Last.fm profile
- Number and positions of added and removed trail marks
- Total mouse movement distances
- Time spent during the session

Subjective measures are derived from asking participants to answer questions about their experience with the system. The questions were designed to evaluate user's perceived:

- Recommendation accuracy
- Trust in the recommendations
- Diversity of recommendations

- Understanding of the recommendations
- Understanding of the interface
- Ease of use
- Overall satisfaction

Many of the listed measures are correlated, but the difference and relation between recommendation accuracy and user's trust in the recommendations in particular need further elaboration. Objectively, predictive recommendation accuracy was measured by collecting ratings of recommended items provided by users. On the other hand, we also measured perceived recommendation accuracy through a post survey question. The main difference between these two measures is in their granularity. Users provide ratings for individual items and recommendation lists as a whole during the evaluation session. While providing such ratings, users are most likely concerned with how well the recommended items satisfy their current need and expectations. In MoodPlay, the rating depends on factors such as perceived match in mood metadata, genre similarity to user's profile and possibly serendipity. Perceived accuracy measured at the end of the session reflects the overall impression of algorithmic accuracy, influenced by factors such as explanations and sense of control over recommendations during the session as a whole.

Trust in the recommendation system is heavily influenced by recommendation accuracy. It is built over time and depends on the user's understanding of the recommendation process, among other factors. For example, Buczak et al. [12] noticed that users thought their recommender was defective after suggesting unknown TV shows. Hence, they particularly focused on developing a ”trust building” mechanism. In addition to recommendation accuracy, MoodPlay has several features that are expected to increase users’ trust, most important of which are visualization and explanation of mood space and control over recommendations.

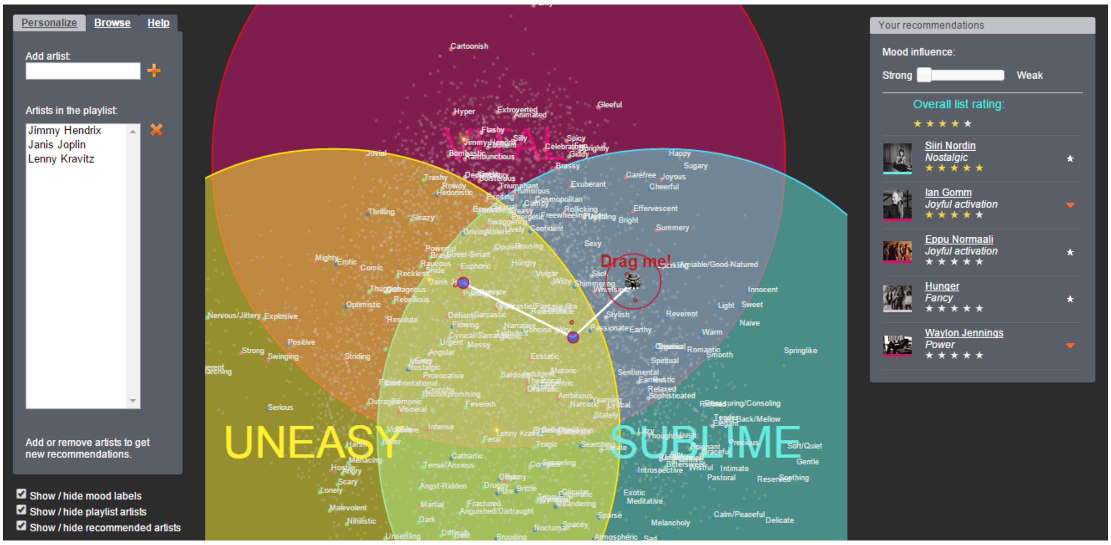
Chapter 6

Preliminary Study

The first formal evaluation of the system was conducted after several informal lab studies and one small MT pilot study. This chapter describes the study design and setup and presents results in three areas. First, we describe a user interaction and perception analysis, followed by a more holistic system evaluation using a structural equation model. Finally, an in-depth analysis of three important dimensions –trust, interaction degree and cognitive load.

6.1 Setup Details

Preliminary study was conducted on the first version of MoodPlay system (Figure 6.1). Comparing to the final version of the system described in Chapter 4, the first version differed in the following ways: (1) mood nodes were significantly smaller, and the distinction three top mood categories were highlighted



Personalize Browse Help

Add artist:

Artists in the playlist:

Jimmy Hendrix
Janis Joplin
Lenny Kravitz

Add or remove artists to get new recommendations.

☑ Show / hide mood labels
☑ Show / hide playlist artists
☑ Show / hide recommended artists

Your recommendations

Mood influence:

Strong Weak

Overall list rating:
★★★★★

Siiri Nordin
Nostalgic
★★★★★

Ian Gomm
Joyful activation
★★★★★

Eppu Normaali
Joyful activation
★★★★★

Hunger
Fancy
★★★★★

Waylon Jennings
Power
★★★★★

Figure 6.1: First version of MoodPlay interface

with Venn diagrams in different colors, (2) music streaming was not directly available. Instead, users had to visit artists' Last.fm profile and play their music and (3) artist information boxes were not available upon clicking on artist nodes in the visualization.

The study setup was as described in the section 5.2, with the additional specificity. Users were required to rate each recommendation list as a whole and at least first two items in it during the session. As we found out after the study, this had a limiting effect on users, as they were not able to interact freely with the system without rating recommendations after each action.

6.2 Participants

In total, 397 participants took the study. After filtering out users we didn't deem as valid because of incorrectly answering attention check questions in the survey or answering all the survey items with the same value we ended up with 240 validated users. The distribution of users from conditions 1 to 4 was: 68, 60, 51 and 61. Studies lasted an average of 25 minutes and participants were paid a fixed amount of \$1.30 per study. Participant age ranged from 18 to 65 with an average range of 25-30. 57% were male. 44% had a four year college degree, and 6% had High School or less. 66% were familiar with data visualization; 77% used a mouse for the interactive study and 19% had a trackpad. When asked about music tastes, 80% said they listen to music frequently. Reported use of streaming services such as Pandora was normally distributed. 40% of participants reported that they preferred popular music, while 6% reported that they had esoteric music taste. Participants were asked an indirect question to assess trust propensity and behavior. The results were approximately evenly distributed across low, medium and high trust bins. During the design stage of this experiment, approximately 10 informal lab-based studies were also conducted and participants were interviewed to gauge their experiences with the system.

6.3 Limitations

During the study setup, a computational error was made during the indexing of artists and their positioning in the mood space. This resulted in a number of the artists being assigned to incorrect mood meta-data. In particular the error affected 37% of the artists significantly. The consequence of this error was that the first step in the hybrid recommendation phase –prediction of artists with similar mood, contained some noise. However, the second step, which is based on content features, was unaffected by the error. Accordingly, we focus our evaluation on user characteristics, interaction and experience, and place less attention on ratings-based analyses. A follow-up experiment is underway with a corrected model to assess these aspects in detail.

6.4 User Interaction

For each participant, we examined the difference in ratings between the first and final recommendation lists. To recap, the initial list is generated based on the user profile only. Subsequent lists are influenced by user interactions in the UI, for conditions 2-4. Figure 6.2 shows a breakdown of this rating shift by condition. The first two columns show the interactive conditions, and there is a significant improvement in mean ratings compared to the static conditions. The difference is approximately half a star on the the 5 point Likert scale. A one-way ANOVA and post-hoc test revealed that this difference was significant at p=4.15E-11. This result shows that interaction in the mood space has a positive influence on ratings. We believe that this result can be more pronounced when data in a user profile is stale. In this particular study, user profile data was provided at the beginning of the recommendation session.

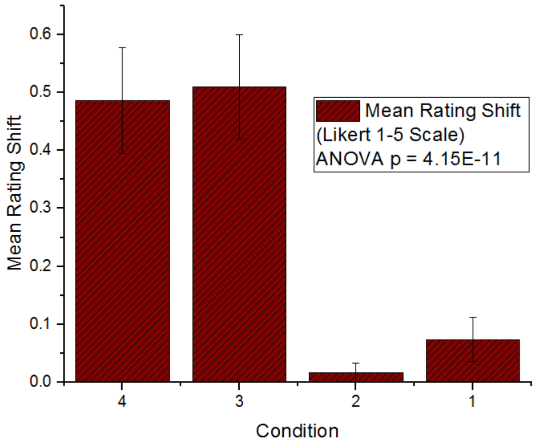


Figure 6.2: Mean rating shift between first list and last list rated, grouped by experimental condition. ANOVA shows differences are significant at $p < 0.05$

Conditions 1-4 in this experiment have increasing visual and interactive complexity. In order to understand the cost of the observed improvement in rating accuracy, an analysis of the time spent in the recommendation session was performed for each condition. Figure 6.3 shows the results of this analysis. In conditions 1 and 2, sessions lasted about 6 minutes on average, while in conditions 3 and 4, sessions averaged about 8 minutes.

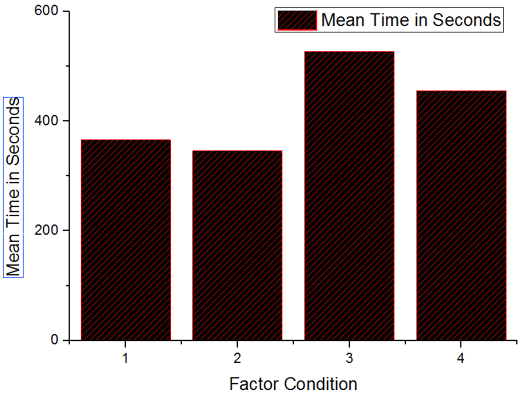


Figure 6.3: Total time spent in the recommendation session for each condition.

6.5 User Perception

Recently, researchers in recommender systems are recognising the importance of user experience in addition to traditional success metrics such as predictive accuracy and diversity. In this experiment, a large amount of qualitative data was collected in order to explore variations in user experience across the 4 experimental conditions. Figure 6.4 shows a small sample of these results for the full treatment condition. The figure shows results of mean agreement with a set of statments in the post study. Agreement was reported on a percentage scale. Examples of questions are as follows:

- **Q_DIVERSE** The recommendations were diverse
- **Q_ACCURATE** The system gave me accurate recommendations
- **Q_FUN** The system was fun to use
- **Q_TRUST** I trusted recommendations from the system

Analysis was also performed across the different conditions. In general, the two interactive conditions (described in 3 and 4 of Table 5.1) showed a small improvement over the other conditions along most dimensions. Of particular note is that both 3 and 4 showed a large (40%) increase in engagement with the recommender system, measured by total interactions over all control elements.

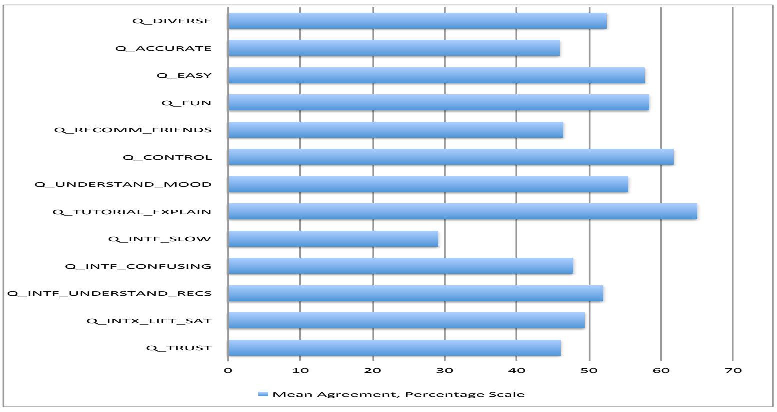


Figure 6.4: Sample of collected user experience metrics from post study for the full feature treatment (condition 4).

6.6 Structural Model

Since the MoodPlay system combines a recommendation algorithm, an interactive interface and subjective experiences of participants in the experiment, there are many variables that interact with each other. To study these interactions, several structural equation models [46] were tested over the personal characteristics of users from the pre-study; objective system aspects that were controlled in each condition; subjective aspects from the post study questionnaires and observed dependent variables from analysis of the system log data. Figure 6.5 shows the result of one such model with a reasonable fit to the data ($X^2(240) = 190, p < 0.05$). In this representation, edge thickness highlights the stronger effect sizes and val-

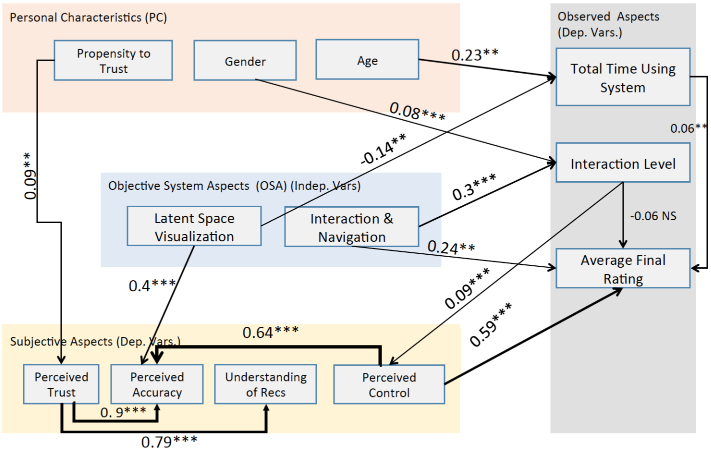


Figure 6.5: Structural equation model for variables in the experimental data, computed using Onyx. Significance levels are '***' p<.001, '**' p<.01, 'ns' p>0.05. All factors in the model have been scaled to have a standard deviation of 1. Arrows are directed and edge values represent β co-efficients of the effect.

ues can be positive or negative, indicating effect direction. Notably, trust (both propensity and perceptive trust) plays an important role in how users perceive and understand recommendations. Visualization of the latent space causes an improvement in perceived accuracy. Gender influences degree of interaction, while participant age was more likely to influence the total time spent in the system, with older people spending more time on their interactions.

6.7 Interaction and Trust

Previous studies on recommender systems have shown a relation between the inherent user's propensity to trust and its final perception of the system [63, 46, 47]. Considering those previous results and the relations we have found in the structural model, we dug deeper into how trusting propensity and the interface itself affect the final perceived trust of the user. During the experiment, subjects answered a question in the pre-study survey measuring trusting propensity, $trust\_propensity_u$, and a question in the post-study survey assessing their perception of trust while using the system, $post\_trust_u$. We calculated the *Trust gain* of a user as a ratio $Trust\_gain\_ratio = \frac{post\_trust_u}{trust\_propensity_u}$ and we compared this metric across different conditions, as shown in Figure 6.6. As an example, if a user had a trust gain ratio of 1.2, it implies that her perception of trust with the system increased by 20% with respect to her trusting propensity; while a value of 0.9 implies a decrease by 10% in trust. To summarize the results, Figure 6.6 has two plots, at the bottom four scatter plots showing for each user (each dot) the trust gain ratio (x axis) and the initial trusting propensity (y axis). We highlighted with a box in each condition the users (dots) with a gain larger than 1.5 and a trusting propensity smaller than 0.3. The only condition that shows very few users with gains in that area is condition 1 (5.8% of subjects), which had no latent mood space visualization, compared to conditions 2 (16.6%), 3(15.7%) and 4(18%). The other conditions had a considerable amount of users with trust gain ratios up to 100% ($Trust\_gain = 2.0$). In addition, the upper plot in Figure 6.6 shows the marginal distribution of $Trust\_gain\_ratio$ per condition, and the only one showing a clear peak below 1.0 is condition 1. Although more analysis is needed to make this finding more conclusive, these results are in line with the aforementioned studies on user controllability in recommender systems and indicate that the visualization had an impact on increasing user trust on the recommendations.

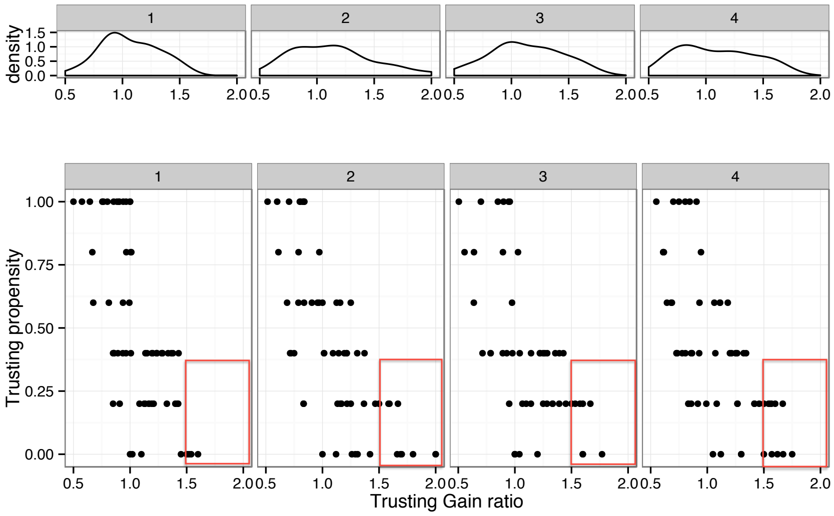


Figure 6.6: Trust gain ratio among different conditions. The red squares at the bottom-right corner of each plot highlight that only the conditions with latent mood space visualization (2, 3 and 4) obtained a considerable trust gain ratio in users with a low trusting propensity (< 0.3), implying that the visualization has an effect on the perceived trust over the recommendations. The upper plots show the marginal trust gain ratio distributions per condition.

6.8 How much Interaction?

Past studies have shown that although more interaction improves the user satisfaction with the system in general, this effect is mediated by personal characteristics such as trusting propensity, familiarity with the domain and language command [46, 47]. In addition, Hijikata et al. [35] and Parra et al. [63] have tried to go beyond in order to understand how much interaction and control is enough for the user before facing cognitive strain. Albers [1] states that learning new interactions requires additional work and remembering, and users prefer to optimize their cognitive resources instead of maximizing their work output. In our study, we observed an important effect relating amount of interactive features on the interfaces and user's perception of understanding them. Figure 6.7 shows the distribution of answers to the question in the post-study survey *The system helped me understand and compare moods of different artists.*

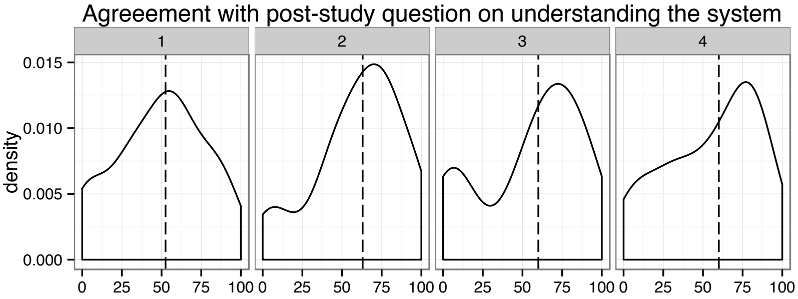


Figure 6.7: Distribution of raw agreement scores of study subjects with the question about perceived understandability on the system.

In the figure, the dashed line represents the median of the agreement with the statement, and it is clear that in conditions with the latent mood space visualization (2, 3, and 4) the agreement is larger than in the condition without it. Since the distributions depart from normality, we conducted a non-parametric Wilcoxon test rather than a t-test and we found that the agreement with the statement is significantly smaller in condition 1 than in condition 2, $W = 2389$, $p = 0.019$. This agreement is not significantly different among any of the other conditions. Another question in the post-study asked subjects about how confusing was the interface. Condition 1 was perceived, as expected, significantly less confusing than the other three interfaces. Now, condition 2 and 3 were not perceived significantly different in this item (Wilcoxon test, $W = 1472.5$, $p = 0.85$), whereas condition 4 was actually perceived as significantly more confusing more condition 2 (Wilcoxon test, $W = 1417$, $p = 0.04$), which only adds trails with respect to the interactive features of condition 3.

Notable, these results might indicate that for understandability of the system as a whole, a holistic interactive visualization, such as our proposed latent mood space, might promote user understanding of the system by allowing exploration even without facilitating specific aid for personalized exploration as in conditions 3 and 4. Now, adding a user avatar to personalize the user exploration adds some cognitive strain but it is still well managed by users. However, the complexity of adding trails significantly increases the user perception that the interface becomes too difficult to utilize effectively as an information filtering system.

Chapter 7

Comprehensive Study

7.1 Setup Details

MoodPlay system has been developed in two stages. The evaluation of the first version of the system is presented in the previous chapter and the second study described here was conducted on the final version, described in Chapter 4. Comparing to the first version, we improved graphical design and several interface features based the user feedback. Mood space visualization has been updated to show smoother transition between mood categories, we added music streaming and enabled access to artist information by clicking on artist nodes. Several modifications were made to improve the experiment design as well – we gave users more freedom to naturally interact with the system and we tracked additional interaction metrics. Finally, in the previous study we focused the evaluation on user characteristics, interaction and experience, and placed less attention on ratings-based analysis. Here we describe the modified experiment, conducted with an entirely new set of participants, report the results of both quantitative and qualitative analysis and address impact of mood based interactions on user experience.

The following sections describe the results of our user experiment (N=398, before filtering, N=279 after). We first describe participants, and follow with a discussion of results on interaction, ratings, mood and user experience metrics.

7.2 Participants

In total, 398 participants took the study, equally distributed across all 4 conditions. Here we report the user participant statistics for 279 users who completed valid sessions. The distribution of users from conditions 1 to 4 was: 70, 69, 70 and 70. Studies lasted an average of 20 minutes and participants were paid an amount of $1.00 per study. Age ranges of participants were reported from 18 to over 65, with an average range of 25-30. 52% were female. 13% did not finish college, 40% had a four year college degree and 47% had a graduate degree. 74% were familiar with data visualization; 66% used a mouse for the interactive study and 34% had a trackpad. When asked about music tastes, 89% said they listen to music frequently. Reported use of streaming services such as Pandora was normally distributed. 71% of participants reported that they preferred a mix of popular and esoteric music. Participants were asked an indirect question to assess trust propensity and behavior. The results were approximately evenly distributed across low, medium and high trust bins. During the design stage of this experiment, approximately 10 informal lab-based studies were also conducted and participants were interviewed to gauge their experiences with the system.

7.3 Interaction and Exploration

The interaction analysis shows important differences in user behavior among the different conditions, which are summarized in Figure 7.1.

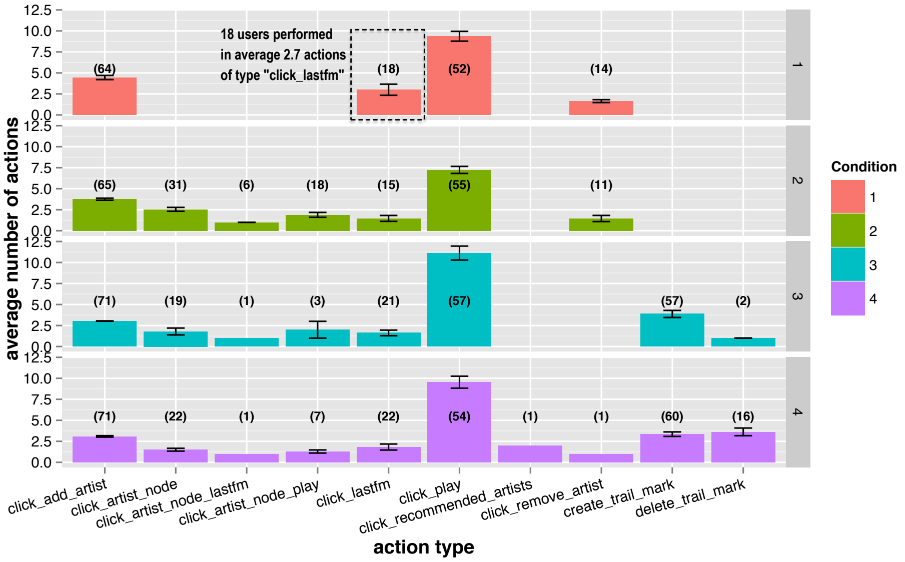


Figure 7.1: Average amount of user actions in the four Moodplay conditions: (1) traditional ranked list, (2) visualization without avatar, (3) visualization including avatar, and (4) visualization with avatar and trails. Numbers in parentheses show the amount of users performing the action.

In **condition 1** users could see only the widget to add and remove profile items (actions *click_add_artist* and *click_remove_artist*), and the ranked list of recommendations. Users could also play the music of the recommended bands (*click_play*) and follow links to artists' Last.fm profiles (*click_lastfm*). This limited amount of interactive features made users focus on the aforementioned actions to control the recommendations and explore their quality.

Condition 2 showed the visualization which allowed users to explore the artists positioned on the mood space, but they could not update their recommendations by interacting with an avatar. This probably explains why we see similar number of actions to add (*click_add_artist*) and remove artists (*click_remove_artist*) from the preference list. However, we see how users decreased the average amount of bands whose music was played from the recommendation list (action *click_play*) and who fetched additional information in the bands' Last.fm pages (*click_lastfm*). This does not mean a lack of user engagement with the system, but rather they alternatively performed these explorations by interacting with the canvas, by clicking on the artist nodes (action *click_artist_node*), playing bands' music directly from the nodes in the mood space (action *click_artist_node_play*) and fetching for additional bands' information in Last.fm pages (action *click_artist_node_lastfm*). In this condition, user sessions were shorter in seconds (M=385.78, S.D.=189.69) compared to those in condition 1 (M=417.7, S.D.=261.27).

In **condition 3**, users where able to see the mood space and they additionally had an avatar which could be dragged, and after each movement the list of recommendations was updated. These interface features reduced the amount of exploration, in terms of users and actions, over the mood space, but increased the users' activity in general since now users played more artists in average (M=11.23, S.D.=6.28) and they spent on average more time in second than all other conditions (M=446.63, S.D.=237.24). Interestingly, users did not remove artists from their user profile, since the simple movement of the avatar over the canvas was enough to get recommendations updated.

Finally, in **condition 4**, the system had the same features as condition 3 with the addition of a trail drawn between the positions the avatar. This produced a decrease in amount of bands played per user with respect to condition 3 (M=9.54, S.D.=5.24), but in terms of other actions was rather similar than condition 3 such as interactions over the canvas (actions *click_artist_node*, *click_artist_node_play*, and *click_artist_node_lastfm*) and fetching for additional artist information on Last.fm (action *click_lastfm*). In this condition there was a new action to allow users remove the trail (action *delete_trail_mark*) which probably made users divert their attention to deleting their previous avatar location to update the recommendations. Users spent on average more time in condition 4 than in conditions 1 and 2, but less than in condition 3 (M=415.73 seconds, S.D.=173.38).

7.3.1 Diversity

One of the most interesting results of our study is that a precise combination of visualization and interactions can effectively promote diversity among items consumed. We measured this effect by analyzing the amount of unique artists rated and played per user in each condition. With respect to artists played, we compared

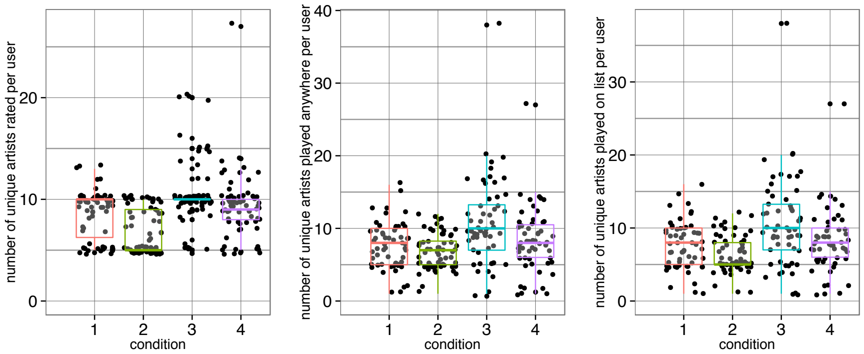
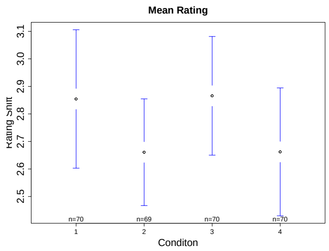


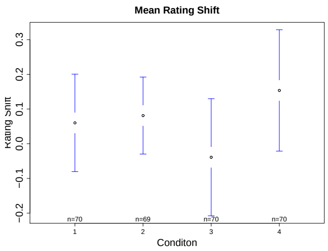
Figure 7.2: Consumption of unique items per user: rating-based (left) , and played-based interactions in all interface (center) and on the recommendation list only (right).

"playing activity" in any widget of the interface (visual space and recommendations) and also in the recommendation list only, to make a fair comparison against condition 1. Plots in Figure 7.2 show these distributions. Significant differences were assessed with Wilcoxon signed-rank tests since data departs significantly from normality. The most important result is that condition 3 significantly outperforms all the other conditions in the three aforementioned metrics: rated items (M=10.59 , S.E.=0.41), p <.001, artists played anywhere (M=10.71, S.E.=0.81), p = .002, and artists played on the recommendation panel only (M=10.61, S.E.=0.82), p <.003. Also notable, condition 1 shows significantly more diversity than condition 2 in terms of unique artists rated 1 (M=8.56, S.E.=0.28), p <.001, and played in the recommendation list (M=7.63, S.E.=0.43), p <.02.

7.3.2 Ratings

Now that we have discussed the various interactions that users made with the canvas, we examine the impact of that interaction and exploration on actual ratings provided. An initial examination of the mean rating per condition (Figure 7.3) shows that mean rating of the final recommendation list is lowest in condition 2 and 4, but an anova shows that this is not a significant difference. To account for rating propensity differences between users, a rating was taken for an initial list of items at the beginning of each session, and then again for a list of recommended items at the end of the session. Figure 7.4 shows the mean improvement in rating observed across each condition from first list rated to last list rated. Ratings were taken on a 5 point Likert scale. Here, condition 4 shows the largest improvement in rating, but our data does not show that this is significant. However, looking at the total shift in rating, regardless of direction, Figure 7.5 shows us clearly that the more interactive conditions (3 and 4) produce a significant ($p<0.05$) shift in ratings compared against the less interactive conditions (1 and 2). This result indicates that simply explaining a mood space visually has minimal impact on resulting item ratings, while interacting with an avatar, either with or without trails, creates more variability in ratings. We are interested in further exploring this effect to understand what patterns of interaction, if any, correlate with the observed positive and negative changes in observed ratings.





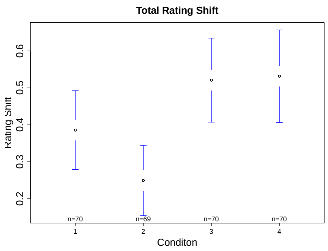


Figure 7.3: Mean Final List Rating by Condition (No rating propensity)

Figure 7.4: Mean Rating Shift by Condition (Last - First list rated

Figure 7.5: Total Rating Shift by Condition (Last - First list rated)

7.4 Mood-based Analysis

MoodPlay combines facets from data in the wild, automated algorithms, UI design, perception and interaction. Since there are many interacting variables, it is difficult to evaluate every possible causal relation for observed effects.

7.4.1 Mood Entropy

In order to explore the effects of mood data and interaction on observed user ratings, we introduce the concept of *mood entropy.* For example, if there are n different moods available, an artist has highest mood entropy if their music is evenly distributed across all three top mood categories (sublime, vital and uneasy). We believe this is a useful metric for MoodPlay since the interactive mood space allows a user to navigate towards the areas where mood categories overlap or towards the areas in distinct categories.

Figure 7.6 shows the results of our analysis of user ratings and entropy for each of the four experimental conditions. Each data point is an individual musical artist. The x-axis shows rating bins for each condition and the y-axis shows the entropy score. A low value on the y-axis means that an artist's music tends to focus on one mood category, while a high score shows a more even distribution across the categories. Each group of box-plots represent the entropy of items that received the given rating in each condition. We can observe from the right side plots for conditions 3 and 4, that items that received ratings of 4 and 5 tend to have higher entropy –that is, they are less associated with any one particular category. Furthermore, if we look only at the lower entropy items, shown below 1.00 on the y-axis, there is a clear increase in the number of artists receiving 4 and 5 star ratings in conditions 3 and 4. This tells us that interaction in the mood space also helps users find relevant artists whose music is focused on one particular sentiment, as identified by our main mood categories.

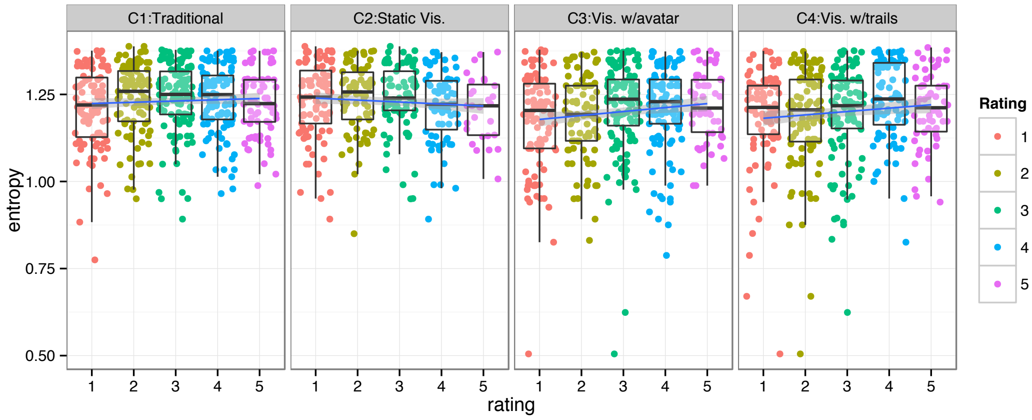


Figure 7.6: Entropy-based interaction results.

7.4.2 Mood Preference

As described in the previous subsection, we observed that more interactive conditions 3 and 4 helped users find artists representative of top mood categories – *sublime*, *vital* and *uneasy*. To further explore relation between different mood categories and rating accuracy, we compare ratings of artists across 5 groups: *sublime*, *vital*, *unease*, *other* and *mix*. All groups except *mix* contain artists representative of the corresponding categories, whereas *mix* group contains artists that do not have a representative category. Next, we describe the process for forming the groups.

An artist is characterized by weighted moods, each belonging to one mood category. We form sets $V = w_0, ..., w_k$, $U = w_0, ..., w_m$, $S = w_0, ..., w_l$, and $O = w_0, ..., w_n$ where w_i are weights of *vital*, *uneasy*, *sublime* and *other* moods respectively. We define set A as a union of sets V, U, S and O. Finally, we calculate the ratios v, u, s and o of each category that characterize the artist:

$$v = \frac{\sum_{w \in V}}{\sum_{w \in A}}, u = \frac{\sum_{w \in U}}{\sum_{w \in A}}, s = \frac{\sum_{w \in S}}{\sum_{w \in A}}, o = \frac{\sum_{w \in O}}{\sum_{w \in A}} \tag{7.1}$$

We then place all the moods that have one prevalent category, with the ratio greater than 0.5, into the corresponding group. All artists that do not have a prevalent category, or in other words none of the categories has the ratio greater than 0.5, are placed into group *mix*. Figure 7.7 shows average artist ratings for each one of the groups. Group *other* consists of only 25 artists, significantly less

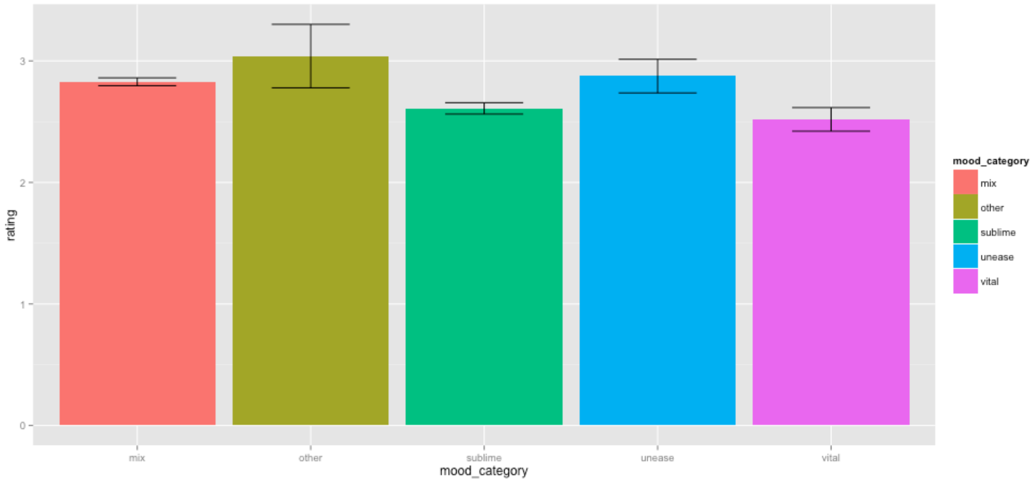


Figure 7.7: Average ratings of artists belonging to different mood categories (*vital*, *uneasy*, *sublime* and *other*) and those that do not have a dominant category (*mix*).

that than the remaining groups, and therefore we don't use it in the analysis. We can see in the Figure 7.7 that groups *mix* and *uneasy* have the highest rating and there is no significant difference between the two. However, we find that the mean rating of artists in group *mix* (M=2.83 , S.E.=0.03) is significantly higher that mean ratings in groups *sublime* (M=2.61, S.E=0.05), p = 0.0001 and *vital* (M=2.52, S.E=0.1), p = 0.003.

To examine the effect of interactive MoodPlay features, we compare mean ratings of artists in each group per experimental condition (Figure 7.8). In condition 1, the only significant difference shown by t-test is that *mix* (M=2.91, S.E=0.06) is higher than *sublime* (M=2.52, S.E=0.1), p = 0.001, and in condition 2 *unease* (M=3.06, S.E.=0.21) is higher than *mix* (M=2.57, S.E.=0.07), p = 0.03. However, in conditions 3 and 4 we see the rating pattern observed earlier in the aggregated ratings from all conditions. In condition 3, t-test doesn't show significant differences, but from the plot we see the tendency for *mix* to be higher than *sublime* and *vital.* But, in condition 4, *mix* (M=2.87, S.E.=0.07) is higher than *vital* (M=2.45, S.E.=0.16), p = 0.02 and *sublime* (M=2.5, S.E.=0.08), p = 0.0005.

Considering that users were updating recommendations only by adding or removing artists into their profile in conditions 1 and 2, rating differences among different mood groups may be due to preferences of users who participated in those conditions. In conditions 3 and 4, users were able to control recommendations by moving the avatar or adjusting the hybrid recommendation algorithm, and on average they gave higher ratings to the artists with mixed moods. Although more research is needed to make conclusive results, this indicates that people on average prefer recommendations that carry distributed mood content, over those that have a dominant mood category.

7.5 Qualitative Analysis

Recently, researchers in recommender systems are recognising the importance of user experience in addition to traditional success metrics such as predictive accuracy and diversity. In this experiment, a large amount of qualitative data was collected in order to explore variations in user experience across the 4 experimental conditions. The participants were asked to rate around 20 statements (numbers per conditions differ slightly) with values from 1 to 100 which indicate

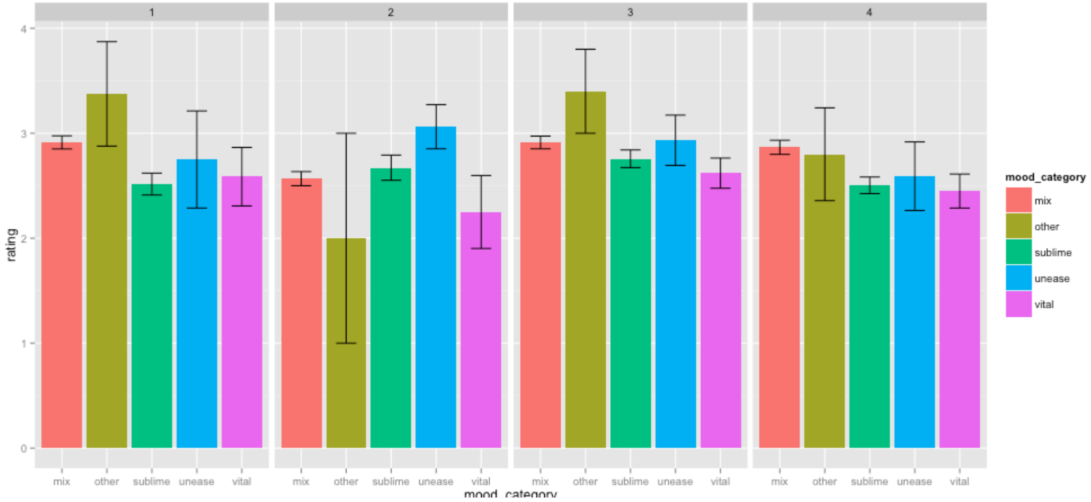


Figure 7.8: Average ratings of artists belonging to different mood categories (*vital*, *uneasy*, *sublime* and *other*) and those that do not have a dominant category (*mix*), per experimental condition.

disagreement and agreement respectively. For simplicity, we report only answers to a set of the most relevant questions here (Table 7.1) and list all questions of this type in the Appendix B.

Perceived trust was measured from two different angles: trust that the system produces good recommendations and the effect of interface on the trust in the recommendations. In all four conditions users believe that interface improves their inherent trust in the system. This is most prominent in condition 1, which does not display the mood space, but still shows mood categories and sub-categories for artists in the recommendation list. System was perceived as the most trustworthy overall in condition 3, whereas interface increased the trust the least in condition 4.

	Mean agreement and standard error per condition			
Statement	1	2	3	4
I trusted recommendations from the system	37.1 ± 3.6	44.6 ± 3.5	48.8 ± 3.5	38.4 ± 3.6
Interaction with the interface increased my trust in the recommendations	43.4 ± 3.6	47.1 ± 3.8	49.4 ± 3.8	39.2 ± 3.7
The recommendations were diverse	60.9 ± 3.3	65.3 ± 3.3	68.6 ± 3.3	59.9 ± 3.3
The interface helped me understand and compare moods of different artists	49.4 ± 3.5	55.7 ± 3.5	55.7 ± 3.3	46.3 ± 3.4
The interface helped me understand how recommendations were generated	42.8 ± 3.6	54.4 ± 3.9	58.6 ± 3.8	50.3 ± 3.7
The interface was confusing	23.1 ± 3.1	45.3 ± 4.1	46 ± 3.9	52.6 ± 3.9
Overall, the recommendations were accurate	36.2 ± 3.6	40.7 ± 3.6	49.8 ± 3.7	38.7 ± 3.5
The system was easy to use	73.9 ± 3.6	58.3 ± 3.8	63.8 ± 3.6	53.2 ± 4
By the end of the session I was satisfied with the recommendations	42.2 ± 4.1	44.2 ± 4	49.3 ± 3.9	38 ± 3.6

Table 7.1: Summary of the most relevant variables in the post-study survey. Numbers indicate average user agreement (on a scale from 1-100) with mean ± S.E.

We also examined how interface affects understanding of moods and the recommendation process. As expected, perceived ease of use drops-off with the increasing interface complexity and confusion rises. Interface in the condition 3 has the best balance of explanation and clarity, followed by condition 2. Furthermore, we did not see a clear differences in average ratings per condition, but the perception of accuracy in condition 3 is significantly higher than in conditions 1, 2 and 4. Similarly, participants perceived recommendations in condition 3 as the most diverse, and those in condition 4 as least diverse. All of these results indicate that the visual layout of moods and artists, with the addition of ability to re-position the avatar in the mood space and control the hybrid recommendation algorithm, gradually improve user experience. However, introduction of trails in condition 4 has a negative effect, most likely because of the cognitive overload. In addition, we suspect that trails may be perceived as limiting for exploration and may be causing a conflict between user's expectation to receive recommendations only from the most recent mood area rather than all previous trail points.

7.5.1 Participant Feedback

In each condition, participants were asked in the post survey to leave feedback on their experience and give suggestions for improving the system. Table 7.2 lists representative comments, grouped by condition and sentiment. On the positive side, many users had fun using *MoodPlay* and enjoyed discovering new artists in different moods in conditions 3 and 4. Drawbacks observed across all four conditions are small artist database and mixing genres in the recommendation lists. In addition, visualization rendering was sluggish for some users. These problems can be addressed in the future by considering genre in the recommendation algorithm and by optimizing visual solution for even larger artist database.

Cond.	Positive comments	Negative comments
1	All good.	Add more bands/artists to the search- for example, neither Silversun Pickups nor Smashing Pumpkins were found to add to my list.
	It was really fun.	The recommendations didn't seem to match the artists I chose.
	I enjoyed using this!	Show more information on how the mood of a song/artist is determined.
2	I think this could be a great tool. Good luck with the progress I am anxious to give it a try when it is finished	I put in 3 rappers and it gave me like oldies and pop songs. Genre plays roles in certain moods.
	I really liked this, it is a new concept that I've never seen. It helped introduce me to artists in different genres that I had never heard before and were very good.	It runs a little slow, should improve optimization for older computers.
	The mood cloud is awesome, and I didn't know there could be so many different music moods, that was great, but not being able to explore the artists within each specific mood circle causes some frustration. Making the cloud more dynamic to dragging and clicking would enhance the tool.	I really didn't understand it.
3	Really good player, i would change nothing it actually made me listen to a couple of artists i did not know about and liked their music.	Make the interface simpler and more concise. Speed up loading times
	An interesting concept. I use Pandora a lot, and my stations are usually based off of my mood that day. This tool would be useful for randomization of choices of music.	It was slow and laggy and some of the recommendations didn't have a play button. I'd like the option to buy a track if I heard one I really liked, or to save a playlist if I really enjoyed it.
	This is really cool, I do not listen to much music and I think this would help me find some new artists or even be used as a therapy tool.	Larger music selection, possibly change the strong week slider, to broad or specific to the particular mood you are feeling.
4	its a cool design	Some of recommended artists didn't relate to my mood close enough.
	Neat program! If I could practice with it more I think I would really enjoy it.	There is a lot of text on the page and it's a little overwhelming. Instead of starting off with so many "moods," maybe just have 20 initially listed.
	It was excellent! Thanks to the developers for developing wonderful tool.	Make the interface faster and smoother. There was too much choppiness when I was using the visualization tool.

Table 7.2: Selected positive and negative user feedback grouped by experimental condition.

Chapter 8

Summary

This dissertation is motivated by a need to fill in the gaps in music recommendation research by addressing topics in user interface and interaction design, user modeling and automated algorithms. It proposes novel visualization of mood-artist associations, that serves as a basis for novel interaction techniques, aimed to improve user experience with a recommender system. These novelties are built in and demonstrated in MoodPlay – an interactive, hybrid music recommender. Previous chapters explain in detail where MoodPlay stands in relation to previous research, what interface design decisions were made to build it and hybrid recommendation algorithms that were employed. Two extensive user studies were also conducted that show how proposed features improve user acceptance and understanding of artist moods and recommendations, and what level of interaction is suspected to cause cognitive overload. This chapter discusses answers to research questions listed at the beginning of the document, dissertation contributions and future work.

8.1 Discussion

The development of MoodPlay system, and the quantitative and qualitative analysis of experiment results illuminate answers to dissertation research questions.

Question 1: *How can metadata such as mood information be visually represented for a music recommendation system?*

Mood information has been visually represented in several preceding works, with the goal to enable user selection of artists in desired moods. Typically, user would choose a mood point in the visual space and the system would play music associated with the selected mood. To our knowledge, all up to date visualizations of moods for this purpose are based on circumplex model of affect, that represents moods along valence and arousal dimensions. Chapter 3 argues that there exists a need to use a music specific mood model for the purpose of music recommendation, and proposes an approach to fulfilling it. Specifically, a dimensionality reduction method was applied to high-dimensional data containing mood-artist associations. It was then shown that a mood model, previously developed in music psychology research, emerges in the obtained two-dimensional, latent space.

To use this space during recommendation process, and help users understand it better, several design aspects were addressed when incorporating it into an interactive system. Choice of colors, item sizes and transparency, dynamic labeling of mood nodes and node filtering based on mood categories all aim towards explaining the mood space and supporting the recommendation.

Question 2: *How can interaction, explanation and control be supported over such a visualization?*

One possible answer to this question was found through the development of MoodPlay system and described in detail in Chapter 4. The interaction with the system ranges from zooming and panning the visualization to explore the moods and artists, to controlling the hybridization of recommendation algorithm. Both user's profile items and recommended artists are highlighted in the visualization, which helps user understand how those two sets are related based on moods. The explanation and exploration are further supported by providing links to external artist profiles, allowing music streaming and displaying mood categories for recommended items. User avatar is positioned within the mood space as a centroid of user profiles items. The ability to move the avatar and form a trail of mood markers serves as a mechanism for modeling the change in user preference. This is one way to control the recommendations, as they are computed whenever the position is changed. Second way to influence the recommendation algorithm is by setting the ratio between importance of mood versus audio based filtering. This is achieved by controlling simple slider and visually explained to user by resizing the catchment area around the avatar. As the area is increased, the recommendation results depend more on the audio similarity to the profile items and less on the mood metadata.

Question 3: *What are the effects of such interactive visualizations on the user experience with a recommender system? For example, how much interaction is too much?*

User study results clearly showed that the proposed interface design and a certain combination of interactive features improve objective and perceived recommendation accuracy, as well as self-reported user satisfaction. We have shown that introduction of hybridization control for recommendation algorithm and the ability to move user avatar, yielded positive effects across a variety of examined metrics. An important finding was that such features increase user trust and recommendation accuracy, even when the artist are misplaced in the mood space or in other words, mood component of the recommendation algorithm is not fully functional. However, tracking of user mood states in the form of proposed trail introduced undesirable effects. First, we suspect this increased system complexity above comfortable threshold and caused cognitive overload. Second, users who are unfamiliar with the system and participate in short listening sessions, may be more inclined to rapidly investigate the mood space than familiar users in a more natural setting. The trail may have been perceived as a limitation during relatively short experiment sessions. Nevertheless, modeling of changing mood preference is a fruitful research endeavor and our future work can address trails that follow smoother mood transitions, are optional and used during longer listening sessions.

Question 4: *How does knowledge of and interaction with mood metadata influence recommendation accuracy and user experience?*

The interactive mood space encouraged exploration of music data, as evidenced by results in Figure 7.6. We observed that user profile items were overall a mix of three main mood categories, but the opportunity to navigate in the mood space led users to explore more artists with more homogeneous moods. Interestingly, not all discovered artists were received positively, especially in the trail based condition which imposed limitations to exploration, along with a cognitive overload. We also observed that when given the opportunity to explore music collection by mood metadata and control the recommendation algorithm, users preferred artists characterized by a mix of mood categories over those that have a dominant category.

8.2 Contributions

The main contributions of this work lie in the areas of user interface design, interaction techniques and recommendation algorithms.

1. *Novel visual interface for recommendation*

This is the first visual interface that maps music moods and artists in the space alternative to circumplex valence – arousal model. Numerical analysis in combination with holistic approach show that a music specific mood model, derived from extensive studies in music psychology, emerges in the visualization of real world music data.

2. *Hybrid mood-aware recommendation algorithm*

   Although mood tags were used in recommendation algorithms before, and hybrid approaches are common, the contribution here is in the manner of using mood metadata and audio content to suggest new music. These two components are combined in a cascading recommender, in such a way that the method is visually explained to the user at a conceptual level. First, the mood-aware recommendation is computed using the two dimensional representation of high-dimensional mood-artist associations. As a result, it is simple for a user to see where the suggested items lie in the mood space. The second step is filtering of mood-aware recommendations based on audio similarity to the items in the user profile. This filtering can be adjusted by user and the effects are visually explained in the mood space.

3. *Enhanced interaction techniques*

   Main contribution in this area are two following techniques: (1) user control over hybridization of recommendation algorithm (as explained in the previous paragraph), realized in the form of a slider control and (2) trail algorithm that enables the system to keep track of users marker points in the mood space as she navigates it.

The outlined novelties were evaluated in two user studies and the results show important connections between interaction with new features and user's experience with the system. The findings presented in this document inform the design of future interactive recommenders in various domains. Proposed visualization of hierarchical mood model can serve as a tool for further investigation of music perception, and moods expressed and invoked by music. The interactive system itself, here prototyped as MoodPlay, has applications not only in entertainment but also in music therapy, as indicated by several users. Lessons learned during up to date research can lead to the improvements of the system and adaptation for different situations. On the other hand, the interaction techniques such as hybridization control and preference trail can be applied to any recommendation system that maps items in a navigable space.

8.3 Future Work

There is fertile ground for expansions and branching of this dissertation work in several directions. Besides recommending specific songs rather than artists based on mood, MoodPlay system can be improved in the following ways.

Preference trail algorithm. Through the extensive evaluation of the system it was observed via numerous metrics that users preferred recommendations obtained by navigating music collection freely, over the recommendations received by means of available trail based algorithm. This does not mean that modeling the changing preference is not desirable, but rather that the method for doing so needs improvement. For example, users could be given a choice whether to use the system in an exploratory mode and freely navigate, or in preference modeling mode and build the trail. Depending on user's activity, available time and listening context, she could choose to be more or less engaged in the interaction with the system. In cases when user chooses to build the trail, recommending items along the trail, in between trail marks, could provide more gradual change in the recommendations and possibly give a more enjoyable listening experience during long sessions. Such recommendation method would require evaluation in a more natural setting, over a longer period of time.

Comparison of proposed mood space to Circumplex model of affect. To overcome some shortages of using Russel's Circumplex model of affect to map artists in a mood space, this dissertation proposes usage of music specific mood model. We justified the validity of proposed mood space by showing that mood model formed as a result of extensive research in psychology emerges in a low-dimensional space formed by real-world artist – mood associations. However, further research is needed to investigate to what degree the proposed model overcomes the shortcomings of Circumplex model. For example, one metric for comparison is the number and similarity degree of dissimilar moods appearing close to each other in space. It is also important to note that using different set of artists to build the space would produce different results. But the expectations is that the spaces built on sufficiently large sets would have very similar mood layout. Another suspected advantage of the proposed model is that users can find desired moods faster and easier by using the hierarchical structure. Additional experiment to test this hypothesis would be beneficial.

Recommendation algorithm. The average rating values in user studies were approximately around 3 (on a 5 point scale), independent of experimental conditions and other examined factors. In addition, many users said in the post-survey that the system suggested artists that either didn't match their mood or played music in different genre. First, building the mood space using larger artist database should improve the mood based component of the recommendation algorithm. Next, MoodPlay system accounts for audio similarity when recommending music, but audio content analysis doesn't always accurately distinguish between music genres. Therefore, recommendation algorithm can be improved by incorporating genre information. In addition, the system uses audio similarity method that previously yielded satisfactory results but further investigation and comparison of algorithms could yield better results.

Scalability. MoodPlay system was developed on a database of around 5000 artists. In comparison, online streaming services offer access to tens of millions of artists. In order to maximally scale the system, extensive work is needed in several areas. Even though there are efficient ways for dimensionality reduction of millions of data points, visualization design has to be adapted to accommodate such a large number. One simple way to achieve this is to show only limited number of artists on different zoom levels, according to some criteria such as popularity or user's preference. A challenge in such a filtering method is to determine what artists the user is interested in seeing, and to show popular artist but also encourage discovery by introducing less known artists. Finally, more sophisticated algorithms are needed to for computing recommendations on the large scale.

Automated profile building. As of now, users build their profile by manually selecting several artists that they like. Rich research in affective computing briefly outlined in Section 2.2 informs multiple ways of improving MoodPlay to collect user's taste and mood data. For example, user's mood and current preference could be determined based on contextual data such as: social media statuses, time of the day, weather, activity automatically inferred from GPS location or proximity of friends in the network, facial expression captured by mobile device or bodily functions measured by wearable devices.

Appendix A

Correspondence Analysis

Correspondence analysis (CA) is a multivariate statistical method introduced by Hirschfeld (Hirschfeld, 1935) and developed by Jean-Paul Benzecri in the early 1970s. It is conceptually similar to Principal component analysis (PCA), but applies to categorical rather than continuous data. It is most often used to analyze contingency tables – tables that display frequency distribution of variables (e.g. geographical areas and smoking habits or symptoms and treatments in medical research). CA helps identify relations between the variables by graphically representing them in a compact form. This technique and its variations are used to solve problems in a variety of fields: engineering, ecology, humanities, marketing etc. In this dissertation project, CA is used to find relations between 3275 artists and 289 moods. Each artist is characterized by a number of moods, weighted according to their relevance. For the purpose of demonstrating usage of CA, table A.1 shows sample data, where 6 artists are described by 4 different moods.

Artist / Mood	Ambitious	Bombastic	Gentle	Refined
Erykah Badu	6	3	1	0
Jovanotti	2	7	0	0
Woodkid	5	0	3	4
Nneka	4	2	4	1
Kopps	3	8	1	0

Table A.1: Sample contingency table with artists and associated moods as variables

The contingency table with artists and moods as variables, is represented as matrix X and transformed into row and column factor scores R and C. Details of mathematical transformations involved in CA can be found in Abdi, 2010. Here, it is sufficient to say that the data for visualizing relations between the variables is obtained from R and C, where the values correspond to x and y coordinates for artists and moods respectively.

$$M = \begin{bmatrix} 6 & 3 & 1 & 0 \\ 2 & 7 & 0 & 0 \\ 3 & 7 & 0 & 0 \\ 5 & 0 & 3 & 4 \\ 0 & 0 & 6 & 6 \\ 0 & 1 & 5 & 7 \end{bmatrix} \longrightarrow R = \begin{bmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \\ r_{31} & r_{32} \\ r_{41} & r_{42} \\ r_{51} & r_{52} \\ r_{61} & r_{62} \end{bmatrix}, C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \end{bmatrix}$$

Appendix B

Survey Questions

B.1 Pre-survey

Before starting the experiment main task, participants in the study answered questions about demographics, previous experience with recommendation systems and several other relevant topics.

1. What is your age range?
   (a) 18-25
   (b) 25-35
   (c) 35-50
   (d) 50-65
   (e) Over 65
2. What is your gender?
   (a) Male
   (b) Female
3. What is your education level?
   (a) High School or Less
   (b) Some College
   (c) College Graduate
   (d) Master's Degree

(e) Doctoral Degree

4. What is 4 + 8? ____________

5. How familiar are you with data visualization, network visualization, or graphing tools? This could be Matlab, Microsoft Excel, statistics packages like R, or things like parallel coordinate plots, heatmaps, treemaps, or scatter plots.

   (a) Not Familiar
   (b) Somewhat Familiar
   (c) Very Familiar

6. What type of input device are you using on your computer today?

   (a) Mouse
   (b) Trackpad
   (c) Other

7. How often do you use a computer workstation? This could be a notebook computer over 15" or a desktop computer, either for work or personal use.

   (a) Never
   (b) Rarely
   (c) Sometimes
   (d) Often
   (e) All of the Time

8. How often do you play strategy games or role playing games, e.g. Starcraft or Grand Theft Auto?

   (a) Never
   (b) Rarely
   (c) Sometimes
   (d) Often
   (e) All of the Time

9. How often do you listen to music?

   (a) Never
   (b) Rarely
   (c) Sometimes
   (d) Often
   (e) All of the Time

10. How often do use recommender systems (systems that predict items for you, such as Netflix or Amazon)

   (a) Never
   (b) Rarely

(c) Sometimes
(d) Often
(e) All of the Time

About how many songs have you listened to through streaming applications such as Spotify, Pandora or Last.fm?

(a) Don't Use
(b) 50 or less
(c) 51-500
(d) 501-1000
(e) over 1000

11. Are you a native English speaker?

(a) Yes
(b) No

12. Which animal is heavier on average, an elephant or a mouse?

13. Please select the option that best describes your music listening experience

(a) Very easily satisfied
(b) Moderately easy to satisy
(c) Difficult to satisfy
(d) Very difficult to satisfy

14. How would you describe your tastes in music?

(a) Mostly like popular music
(b) Like a mix of popular and esoteric music
(c) Mostly like esoteric music

15. How much does your current mood usually influence your decisions

(a) Not at all
(b) Rarely
(c) Moderately
(d) Frequently
(e) Very Frequently

16. Please rate your agreement with the following. (0="Not at all", 100="Fully Agree"). Compared to my peers I...

(a) 0 - 100 I listen to a lot of music.
(b) 0 - 100 am an expert on music
(c) 0 - 100 am a music lover

17. Consider the following hypothetical scenario...You have $50. You can keep this money and do with it whatever you wish or you can send some or all

of it to another person in another room (whom you will never see or meet). They are also given $50 and the same instructions. Any money sent will be tripled on the way to the other person. Thus, if you send them $10, they will receive $30; if they send you $30, you will receive $90, and so on. You can send them any amount that you wish. You can send them nothing if you wish. This decision is completely up to you.How much of your $50 would you send?

   (a) $0
   (b) $10
   (c) $20
   (d) $30
   (e) $40
   (f) $50

18. How much do you agree with the following statement: "I am a trusting person."

   (a) Strongly Disagree
   (b) Disagree
   (c) Neither Agree nor Disagree
   (d) Agree
   (e) Strongly Agree

19. Which of the following best describes your current mood?

   (a) Sublime (e.g.: joyful, warm and tender moods)
   (b) Vital (e.g.: stimulating moods such as 'lively', 'energetic' or 'fierce')
   (c) Uneasy (e.g.: negative moods such as 'sad', 'tense' or 'fearful')

20. What type of audio will you use for this HIT?

   (a) Speakers
   (b) Headphones
   (c) None

21. Browser Meta Browser Version Operating System Screen Resolution Flash Version Java Support User Agent

B.2 Post-survey

After completing the main experiment task, participants in the user study answered a range of questions regarding their experience with the system. Following Table B.1 contains the list of post-survey questions and average response values in the second user study.

	Average response value per condition			
Question	1	2	3	4
The interface helped me understand and compare moods of different artists	49.4	55.7	55.7	46.3
I trusted recommendations from the system	37.1	44.6	48.8	38.4
Interaction with the interface increased my satisfaction with recommendations	43.4	50.5	52.4	42.0
Adding items to my profile increased my trust in the recommendations	43.4			
Interaction with the interface increased my trust in the recommendations		47.1	49.4	39.2
The interface helped me understand how recommendations were generated	42.8	54.4	58.6	50.3
The interface was confusing	23.1	45.3	46.0	52.6
The interface was slow	22.2	32.3	31.9	37.8
The tutorial explained the system reasonably well.	72.6	62.9	65.8	57.5
By the end of the session I was satisfied with the recommendations.	42.2	44.2	49.3	38.0
Overall, the recommendations were accurate	36.2	40.7	49.8	38.7
The recommendations were diverse	60.9	65.3	68.6	59.9
The interface helped me to control the recommendations	42.7	53.8	63.8	52.7
The system was easy to use	73.9	58.3	63.8	53.2
The system was fun to use	61.8	59.5	65.8	56.3
I would recommend this system to my friends	48.7	48.4	55.5	43.9
I found useful the ability to explore moods and songs in the canvas		59.1		
I found the ability to move my avatar around useful			62.5	
The interface helped me express my current mood			51.1	
I found the trail-based navigation useful				44.7
I understood what the slider on the top right (strong or weak mood influence) controlled			54.6	60.6

Table B.1: Questions given to users in the post-study and answered via slider control with values ranging from 1 (*Strongly disagree*) to 100 (*Strongly agree*). Last four columns contain average response values for each question, across four experimental conditions. Missing average response value means that the question was irrelevant and thus not presented in the corresponding condition.

Bibliography

[1] Michael J. Albers. Cognitive strain as a factor in effective document design. In *Proceedings of the 15th Annual International Conference on Computer Documentation*, SIGDOC '97, pages 1–6, New York, NY, USA, 1997. ACM.

[2] Denis Amelynck, Maarten Grachten, Leon van Noorden, and Marc Leman. Toward e-motion-based music retrieval a study of affective gesture recognition. *T. Affective Computing*, 3(2):250–259, 2012.

[3] Jean J. Aucouturier and Francois Pachet. Improving timbre similarity : How high's the sky? *Journal of Negative Results in Speech and Audio Sciences*, 1(1), 2004.

[4] Claudio Baccigalupo and Enric Plaza. Case-based sequential ordering of songs for playlist recommendation. In *Advances in Case-Based Reasoning*, pages 286–300. Springer, 2006.

[5] Linas Baltrunas and Xavier Amatriain. Towards time-dependant recommendation based on implicit feedback. In *Workshop on context-aware recommender systems (CARS'09).*

[6] Dominikus Baur, Frederik Seiffert, Michael Sedlmair, and Sebastian Boring. The streams of our lives: Visualizing listening histories in context. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1119–1128, 2010.

[7] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, September 1975.

[8] Adam Berenzweig, Beth Logan, Daniel P. W. Ellis, and Brian P. W. Whitman. A large-scale evaluation of acoustic and subjective music-similarity measures. *Comput. Music J.*, 28(2):63–76, June 2004.

[9] Adam J. Berinsky, Gregory A. Huber, Gabriel S. Lenz, and Edited by R. Michael Alvarez. Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis*, 2012.

[10] D. Bogdanov and P. Herrera. How much metadata do we need in music recommendation? a subjective evaluation using preference sets. In *International Society for Music Information Retrieval Conference (ISMIR)*, Miami, Florida, USA, 24/10/2011 2011.

[11] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 35–42. ACM, 2012.

[12] Anna L Buczak, John Zimmerman, and Kaushal Kurapati. Personalization: Improving ease-of-use, trust and accuracy of a tv show recommender. In *in Proceedings of the TV'02 workshop on Personalization in TV, Malaga*. Citeseer, 2002.

[13] Michael Buhrmester, Tracy Kwang, and Samuel D. Gosling. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, 6(1):3–5, 2011.

[14] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[15] Michael A. Casey, Remco Veltkamp, Masataka Goto, Marc Leman, Christophe Rhodes, and Malcolm Slaney. Content-Based music information retrieval: Current directions and future challenges. volume 96, pages 668–696, April 2008.

[16] Oscar Celma. *Music Recommendation and Discovery: The Long Tail, Long Fail, and Long Play in the Digital Music Space*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[17] Òscar Celma and Perfecto Herrera. A new approach to evaluating novel recommendations. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, RecSys '08, pages 179–186, New York, NY, USA, 2008. ACM.

[18] Oscar Celma, Perfecto Herrera, and Xavier Serra. Bridging the music semantic gap. volume 187, Budva, Montenegro, Nov-0Jun-Feb00Jun 2006. CEUR, CEUR.

[19] Li Chen and Pearl Pu. Interaction design guidelines on critiquing-based recommender systems. *User Modeling and User-Adapted Interaction*, 19(3):167–206, 2009.

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

[21] Stuart Cunningham, Stephen Caulder, and Vic Grout. Saturday night or fever? context-aware music playlists. *Proc. Audio Mostly*, 2008.

[22] Ricardo Dias, Joana Pinto, and Manuel J. Fonseca. Interactive visualization for music rediscovery and serendipity. In *BCS-HCI 2014 Proceedings of the 28th International BCS Human Computer Interaction Conference, Southport, UK, 9-12 September 2014*, 2014.

[23] Boi Faltings, Pearl Pu, Marc Torrens, and Paolo Viappiani. Designing example-critiquing interaction. In *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 22–29. ACM, 2004.

[24] Nico H. Frijda. Moods, emotion episodes and emotions. In M. Lewis and J. M. Haviland, editors, *Handbook of Emotions*, pages 381–403. New York: Guilford Press, 1993.

[25] Emden R. Gansner, Yifan Hu, Stephen G. Kobourov, and Chris Volinsky. Putting recommendations on the map – visualizing clusters and relations. *CoRR*, abs/0906.5286, 2009.

[26] Jennifer M. George. Individual differences and behavior in organizations. chapter Trait and State Affect, page 145. San Francisco: Jossey-Bass, 1996.

[27] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. Toward a minimal representation of affective gestures. *Affective Computing, IEEE Transactions on*, 2(2):106–118, April 2011.

[28] Gustavo Gonzalez, Josep Lluís De La Rosa, Miquel Montaner, and Sonia Delfin. Embedding emotional context in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 845–852. IEEE, 2007.

[29] Liang Gou, Fang You, Jun Guo, Luqi Wu, and Xiaolong Luke Zhang. Sfviz: interest-based friends exploration and recommendation in social networks. In *Proceedings of the 2011 Visual Information Communication-International Symposium*, page 15. ACM, 2011.

[30] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. Smallworlds: Visualizing social recommendations. In *Computer Graphics Forum*, volume 29, pages 833–842. Wiley Online Library, 2010.

[31] Darryl Griffiths, Stuart Cunningham, and Jonathan Weinel. A discussion of musical features for automatic music playlist generation using affective technologies. In Katarina Delsing and Mats Liljedahl, editors, *Audio Mostly Conference*, pages 13:1–13:4. ACM, 2013.

[32] Masahiro Hamasaki and Masataka Goto. Songrium: A music browsing assistance service based on visualization of massive open collaboration within music content creation community. In *Proceedings of the 9th International*

Symposium on Open Collaboration, WikiSym '13, pages 4:1–4:10, New York, NY, USA, 2013. ACM.

[33] Byeong-jun Han, Seungmin Rho, Sanghoon Jun, and Eenjun Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, 2010.

[34] Negar Hariri, Bamshad Mobasher, and Robin Burke. Context-aware music recommendation based on latenttopic sequential patterns. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 131–138, New York, NY, USA, 2012. ACM.

[35] Yoshinori Hijikata, Yuki Kai, and Shogo Nishida. The relation between user intervention and user satisfaction for information recommendation. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 2002–2007. ACM, 2012.

[36] D. Howell. *Statistical Methods for Psychology.* Cengage Learning, 2009.

[37] Xiao Hu. Music and mood: Where theory and reality meet. In *Proceedings of the iConference*, 2010.

[38] David Hume. Emotions and moods. In *Organizational Behavior*, pages 258–297. 2012.

[39] Joris H. Janssen, Egon L. van den Broek, and Joyce H. D. M. Westerink. Tune in to your emotions: a robust personalized affective music player. *User Model. User-Adapt. Interact.*, 22(3):255–279, 2012.

[40] Patrik N Juslin and Petri Laukka. Communication of emotions in vocal expression and music performance: Different channels, same code? *Psychological bulletin*, 129(5):770, 2003.

[41] Marius Kaminskas and Francesco Ricci. Location-adapted music recommendation using tags. In *User Modeling, Adaption and Personalization*, pages 183–194. Springer, 2011.

[42] Michelle Karg, Kolja Kühnlenz, and Martin Buss. Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 40(4):1050–1061, 2010.

[43] Aniket Kittur, Ed H. Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[44] Andrea Kleinsmith and Nadia Bianchi-Berthouze. Recognizing affective dimensions from body posture. In *Proceedings of the 2Nd International Conference on Affective Computing and Intelligent Interaction*, ACII '07, pages 48–58, Berlin, Heidelberg, 2007. Springer-Verlag.

[45] Peter Knees, Markus Schedl, Tim Pohle, and Gerhard Widmer. An innovative three-dimensional user interface for exploring music collections enriched. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 17–24, New York, NY, USA, 2006. ACM.

[46] Bart P Knijnenburg, Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 43–50. ACM, 2012.

[47] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 141–148. ACM, 2011.

[48] Stefan Koelsch. A neuroscientific perspective on music therapy. *Annals of the New York Academy of Sciences*, 1169(1):374–384, 2009.

[49] Joseph A Konstan and John Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.

[50] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '12, pages 129–138, New York, NY, USA, 2012. ACM.

[51] Anita Shen Lillie. Musicbox : Navigating the space of your music. Master's thesis, Massachusetts Institute of Technology, 2008.

[52] Beth Logan. Music recommendation from song sets. In *ISMIR*, 2004.

[53] François Maillet, Douglas Eck, Guillaume Desjardins, Paul Lamere, et al. Steerable playlist generation by learning song similarity from radio station playlists. In *ISMIR*, pages 345–350, 2009.

[54] Judith Masthoff. The pursuit of satisfaction: affective state in group recommender systems. In *User Modeling 2005*, pages 297–306. Springer, 2005.

[55] Brian McFee and Gert R. G. Lanckriet. Large-scale music similarity search with spatial trees. In Anssi Klapuri and Colby Leider, editors, *ISMIR*, pages 55–60. University of Miami, 2011.

[56] Sean M McNee, John Riedl, and Joseph A Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1097–1101. ACM, 2006.

[57] Sayooran Nagulendra and Julita Vassileva. Understanding and controlling the filter bubble through interactive visualization: A user study. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, HT '14, pages 107–115, New York, NY, USA, 2014. ACM.

[58] Keith Oatley, Dacher Keltner, and Jennifer M Jenkins. *Understanding emotions .* Blackwell publishing, 2006.

[59] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. Peerchooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1085–1088. ACM, 2008.

[60] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, pages 411–419, 2010.

[61] Han-Saem Park, Ji-Oh Yoo, and Sung-Bae Cho. A context-aware music recommendation system using fuzzy bayesian networks with utility theory. In *Fuzzy systems and knowledge discovery*, pages 970–979. Springer, 2006.

[62] Denis Parra and Xavier Amatriain. Walk the talk: Analyzing the relation between implicit and explicit feedback for preference elicitation. In *Proceedings of the 19th International Conference on User Modeling, Adaption, and Personalization*, UMAP'11, pages 255–268, Berlin, Heidelberg, 2011. Springer-Verlag.

[63] Denis Parra and Peter Brusilovsky. User-controllable personalization. *International Journal of Human-Computer Studies*, 78(C):43–67, June 2015.

[64] Denis Parra, Peter Brusilovsky, and Christoph Trattner. See what you want to see: Visual user-driven approach for hybrid recommendation. In *Proceedings of the 19th International Conference on Intelligent User Interfaces*, IUI '14, pages 235–240, New York, NY, USA, 2014. ACM.

[65] Richard J. Davidson Paul Ekman. *The Nature of Emotion: Fundamental Questions.* Oxford University Press, 1994.

[66] Valery A. Petrushin. Emotion recognition in speech signal: experimental study, development, and application. In *In: Proc. ICSLP 2000*, pages 222–225, 2000.

[67] Rosalind W Picard. *Affective computing.* MIT press, 2000.

[68] Pearl Pu, Boi Faltings, Li Chen, Jiyong Zhang, and Paolo Viappiani. Usability guidelines for product recommenders based on example critiquing research. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 511–545. Springer US, 2011.

[69] Seungmin Rho, Byeong-jun Han, and Eenjun Hwang. Svr-based music mood classification and context-based music recommendation. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 713–716. ACM, 2009.

[70] Curtis Roads. *Composing Electronic Music: A New Aesthetic.* Oxford University Press, first edition, 2015.

[71] J.A. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.

[72] James Schaffer, Prasanna Giridhar, Debra Jones, Tobias Höllerer, Tarek F. Abdelzaher, and John O'Donovan. Getting the message?: A study of explanation interfaces for microblog data analysis. In *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI 2015, Atlanta, GA, USA, March 29 - April 01, 2015*, pages 345–356, 2015.

[73] Markus Schedl, Sebastian Stober, Emilia Gómez, Nicola Orio, and Cynthia C.S. Liem. User-aware music retrieval and recommendation. In Meinard Müller, Masataka Goto, and Markus Schedl, editors, *Multimodal Music Processing*, volume 3 of *Dagstuhl Follow-Ups*, pages 135–156. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 2012.

[74] Klaus R Scherer, Tom Johnstone, and Gundrun Klasmeyer. Vocal expression of emotion. *Handbook of affective sciences*, pages 433–456, 2003.

[75] Bo Shao, Tao Li, and Mitsunori Ogihara. Quantify music artist similarity based on style and mood. In *Proceedings of the 10th ACM Workshop on Web Information and Data Management*, WIDM '08, pages 119–124, New York, NY, USA, 2008. ACM.

[76] Malcolm Slaney. Precision-recall is wrong for multimedia. *IEEE Multimedia*, 18(3):4–7, 2011.

[77] A. Soriano, F. Paulovich, L.G. Nonato, and M.C.F. Oliveira. Visualization of music collections based on structural content similarity. In *Graphics, Patterns and Images (SIBGRAPI), 2014 27th SIBGRAPI Conference on*, pages 25–32, Aug 2014.

[78] Harald Steck, Roelof van Zwol, and Chris Johnson. Interactive recommender systems: Tutorial. In *Proceedings of the 9th ACM Conference on Recommender Systems*, RecSys '15, pages 359–360, New York, NY, USA, 2015. ACM.

[79] Sebastian Stober and Andreas Nürnberger. Adaptive music retrieval—a state of the art. *Multimedia Tools Appl.*, 65(3):467–494, August 2013.

[80] Barbara G. Tabachnick and Linda S. Fidell. *Using Multivariate Statistics (5th Edition)*. Allyn & Bacon, Inc., Needham Heights, MA, USA, 2006.

[81] Marko Tkalcic, A Kosir, and Jurij Tasic. Affective recommender systems: the role of emotions in recommender systems. In *Proc. The RecSys 2011 Workshop on Human Decision Making in Recommender Systems*, pages 9–13. Citeseer, 2011.

[82] Marjolein D. van der Zwaag, Joris H. Janssen, and Joyce H. D. M. Westerink. Directing physiology and mood through music: Validation of an affective music player. *T. Affective Computing*, 4(1):57–68, 2013.

[83] Daniel Västfjäll. Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, 5(1 suppl):173–211, 2002.

[84] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 international conference on Intelligent user interfaces*, IUI '13, pages 351–362, New York, NY, USA, 2013. ACM.

[85] Xinxi Wang, David Rosenblum, and Ye Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 99–108. ACM, 2012.

[86] David Watson, Lee A Clark, and Auke Tellegen. Development and validation of brief measures of positive and negative affect: the panas scales. *Journal of personality and social psychology*, 54(6):1063, 1988.

[87] Howard M. Weiss and Russell Cropanzano. Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. 1996.

[88] Siqing Wu, Tiago H. Falk, and Wai-Yip Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768–785, 2011.

[89] Yi-Hsuan Yang, Yu-Ching Lin, Heng Tze Cheng, and Homer H. Chen. Mr. emo: music retrieval in the emotion plane. In Abdulmotaleb El-Saddik, Son Vuong, Carsten Griwodz, Alberto Del Bimbo, K. Selçuk Candan, and Alejandro Jaimes, editors, *ACM Multimedia*, pages 1003–1004. ACM, 2008.

[90] Feng Yu, Eric Chang, Ying qing Xu, and Heung yeung Shum. Emotion detection from speech to enrich multimedia content. In *Second IEEE Pacific-Rim Conference on Multimedia*, pages 550–557, 2001.

[91] Marcel Zentner and Tuomas Eerola. Self-report measures and models. *Handbook of Music and Emotion: Theory, Research, Applications*, 2011.

[92] Marcel Zentner, Didier Grandjean, and Klaus R Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4):494–521, 2008.

[93] Shiwan Zhao, Michelle X Zhou, Xiatian Zhang, Quan Yuan, Wentao Zheng, and Rongyao Fu. Who is doing what and when: Social map-based recommendation for content-centric social web sites. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(1):5, 2011.