

This paper or a similar version is not currently under review by a journal or conference, nor will it be submitted to such within the next three months. This paper is void of plagiarism or self-plagiarism as defined in Section 1 of ACM's Policy and Procedures on Plagiarism.

Moodplay: Interactive Mood-Based Music Discovery and Recommendation

Ivana Andjelkovic · Denis Parra · John O'Donovan

the date of receipt and acceptance should be inserted later

Abstract A large amount of research in recommender systems focuses on 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 paper introduces *MoodPlay*, a hybrid music-band recommender system which integrates content and mood-based filtering in a novel interactive interface, allowing the user to explore the items on a latent space visualization. We show how *MoodPlay* allows the user to explore a music collection by latent affective dimensions, and we explain how to integrate user input at recommendation time with predictions based on a pre-existing profile. We describe system architecture, design and interactive features followed by a use-case evaluation of the system. Results of a user study (N=279) are discussed, with four conditions being evaluated with varying degrees of visualization, interaction and control. Results of this study show that a low-dimensional visualization of mood information improves user acceptance and understanding of both the underlying data and the recommendations. Allowing users to interact with the visualization, for example to position their avatar as input to the recommender algorithm, produces an additional improvement in these metrics. However, user experience was reduced across all metrics in the more complex condition with trail-based interactions, indicating that cognitive overload was a factor. This was backed up through an analysis of feedback comments from users. The paper concludes with a discussion of this observed impact of visual and interactive features

I. Andjelkovic
University of California, Santa Barbara
Tel.: +123-45-678910
Fax: +123-45-678910
E-mail: ivana@mat.ucsb.edu

D. Parra
Pontificia Universidad Católica de Chile

J. O'Donovan
University of California, Santa Barbara

and associated cognitive limitations that aims to inform design of future interactive recommendation systems.

Keywords Algorithms · Design · Experimentation · Evaluation · Human Factors

1 Introduction

Recommendation systems have become an invaluable tool for finding desired content among millions of options. There are well established algorithms, such as Collaborative Filtering, Content-Based Filtering and Matrix Factorization, used across a variety of domains to recommend digital content or merchandise. However, because of its unique consumption characteristics, music falls into a domain suitable for improvement by alternative approaches to traditional recommendation problem. For instance, we can listen to the same track several times without decreasing satisfaction. Comparing to other domains (e.g. movies), the consumption of music is fast and more context dependent. In this paper, we focus on building an interactive recommender system that suggests music bands based on contextual information. We present interaction mechanisms that allow the user to guide the system based on affective state, which in turn adapts to changes of listening context. There are several music recommender systems that employ different types of context (daily activity [66], time of the day [4], music genre [38], etc.). However, no previous work has integrated affective context for music discovery into a visual and interactive recommendation system.

Experimental evidence shows a strong relation between emotion and music [34] and previous research in mood-aware and emotion-aware recommender systems produced improvements over their non-contextual alternatives [13, 61]. Furthermore, the importance of building interactive recommender interfaces that go beyond the static-ranked list paradigm to improve user satisfaction with a system has been studied in the past [12, 19, 6, 32, 65, 47, 42] and it is supported by results showing that small improvements in accuracy do not always correlate with better user satisfaction [41, 35]. Accordingly, our goal is to build a recommender system with an interactive interface (Figure 1) that supports users in discovering unknown, interesting items and making music choices over an affective music space. We frame our work following these research questions: How can metadata such as affective information be visually represented for a recommender system? How can interaction, explanation and control be supported over such a visualization? What are the effects of such interactive visualizations on the user experience with a recommender system, and what is the right amount of interaction for a music recommender? In our effort to answer these questions, we have produced the following key contributions:

🚧 **To-do:** Mention UMAP paper. Update the list of contributions and point out to differences between previous and current Mood-Play version and study.

🚧 **To-do:** Compared to our previous paper... *—jod*

taste profile, (2) static recommendations, highlighted in a latent affective space visualization, (3) dynamic recommendation lists generated via user interaction in the latent affective space visualization, using current user's preference and (4) dynamic, interaction driven, trail-based recommendations.

The rest of this article is structured as follow: In the first two sections we provide important definitions and we describe and relate important related work. Then we describe the overall architecture of our system followed by a section which provides details such as the interface and interactions, visual affective model and recommendation methods. Next, we describe the user study setup, followed by a discussion of the key results from the study:

- In general, the system was rated as highly novel and fun to use.
- As expected, perception of control is larger in conditions (2), (3), and (4).
- Visualization of affective information in a latent space significantly improved users' understanding of recommendations.
- Trail-based interaction (example shown in Figure 5) was considered too confusing.
- Visual conditions (2), (3) and (4) tend to improve system trust, after trust propensity is controlled for.

Finally, we share our ideas for future work and the resulting implication for design of future interactive recommender systems.

2 Definitions

Research presented in this article is concerned with mood based recommendation. However, throughout the article we use related terms, *emotion* and *affect*, to explain different concepts. Here we provide definitions for each of the terms and other relevant constructs.

Affect: Colloquial term that covers a broad range of feelings. It encompasses both emotions and moods [15,64].

Emotions: Intense, short lived feelings, speculated by most researchers to be directed at someone or something [14,67].

Moods: General, low intensity feeling states that often lack a contextual stimulus [25]. 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 [25,48]. 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 [25]. Therefore, emotion models such as Circumplex model of affect [52] are often used to represent moods as well.

(Affective) Latent Space: We use this term to refer to the 2-D space used to plot artists and moods in *Moodplay*'s visualization. Since this space is obtained by reducing the dimensionality of the initial artists-moods matrix, we call it interchangeably latent space or affective latent space.

GEMS: This acronym stands for Geneva Emotional Music Scales. It is music-specific emotion model proposed and validated by Zentner et al. [72,71].

3 Related Work

The following aspects are the most relevant to our research: affective-aware recommendations, recommendation of music bands, visual approaches to present recommendations beyond a rating list and affective-aware visualizations of music collections.

3.1 Affective Computing and Recommendations

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 [16], gait patterns [30], body posture [31], 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 [49, 70, 68]. Furthermore, currently largest data repository of face videos (2 million) owned by Affectiva¹ is efficiently used to train computers in detecting emotions from facial expressions in real time.

For the recommendation purposes, Masthoff *et al.* [39] integrated affective state in a group recommender system by modelling satisfaction as mood, while González *et al.* [17] incorporated the emotional context in a recommender system for a large e-commerce learning guide. More related to our work, Park *et al.* [44] developed probably the first context-aware music recommender that exploited mood inferred from context information. And more recently, Tkalcic *et al.* [62, 21] discussed the role of emotions in recommender systems and introduced a framework to identify the stages where emotion can be used for recommendation.

In the music recommendation domain, several works infer the users' mood for music recommendation based on movements, temperature and weather [11] or from the music content [51]. Furthermore, Griffiths *et al.* [20] measure 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 enables them to recommend music in the detected mood. Zwaag *et al.* [63] take target mood as an input from user and then select songs that direct the user towards the desired mood, while measuring skin conductance to verify the change. Skin temperature [27] and arm gestures [1] have also been used for inferring mood and querying music collections. Compared to these studies, in our up to date work we determine user's mood based on a set of provided artists and use it to suggest new artists in the similar mood. In addition, we propose rich user interface to help users explore mood space and choose music in desired mood. In the future, proposed 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.

¹ <http://www.affectiva.com>

3.2 Recommendation of Music Bands

Recommendations in the music domain is a well-established field within recommender systems, which have shown, among many others, approaches to recommend tracks [8, 37], albums [45], playlists [38, 3, 22], music targeted at specific venues [29], music targeted at daily activities [66], and artists and music bands [6, 24]. For the sake of space, since our proposed interface aims at recommending music bands, we focus on presenting related work on this sub-field. Hijikata *et al.* [24] used a Naive Bayes recommender to present recommendations of music bands, while Bostandjevic *et al.* [6] used a hybrid controllable recommender system with a visual interactive interface, TasteWeights. Compared to these previous approaches, we innovate by using bands' affective representation to compute similarity, by introducing a user-controllable recommendation interface and allowing users to explore the music band dataset interactively.

3.3 Visual Approaches to Recommendation

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.* [41] and Konstan *et al.* [35], who showed that small improvements in recommender accuracy do not necessarily improve users' satisfaction with a system. However, 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 [18], a sunburst visualization that allow users finding interest-based recommendation in last.fm, and Pharos [73], a social map visualization of latent communities. Other examples that, in addition to visualizations, include a richer user interaction are PeerChooser [43], SmallWorlds [19] –that focus in representing collaborative filtering– and TasteWeights [6] [32], an interactive system that represents a hybrid recommender of music bands. On a different domain, TalkExplorer [65] 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 [47], 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.* [50] and Chen *et al.* [9]. Finally, with a focus on making users aware of the filtering mechanisms on a social network, Nagulendra and Vassileva [42] 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. For a more detailed review of systems and techniques for interactive visual approaches for recommendation, we suggest to read Chen *et al.* survey [23].

To the best of our knowledge, ours is the first work that implements a novel recommender interface that maps the artists on an latent affective space and enables navigation though the space via an avatar. We also incorporate the notion of trails (connected sequential steps on the latent visualization) to allow the user more flexibility in an incremental process of obtaining recommendations.

3.4 Affective-aware Visualizations of Music Collections

Although affective-based music selection and recommendation are gaining popularity in both research and commercial settings, the development of visual aids for affective information is still scarce. Nearly all existing visualizations are built upon Russell's circumplex model of affect [52]. This model is today commonly used to represent emotions and moods as a mixture of two dimensions, valence and arousal, and position them in the coordinate system. Yang *et al.* [69] incorporated it into their music retrieval method, and commercial applications such as Habu² and Musicoverly³ use it as a platform for music selection based on mood.

However, many emotions cannot be uniquely characterized by valence and arousal values [10]. 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 [54]. 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, possibly induces more contemplative range of emotions [71]. To address this problem, we propose a novel visual representation of music specific affective dimensions, built upon the GEMS model derived from extensive psychological study by Zentner *et al.* [72].

4 System Overview

The MoodPlay system is accessible via web browser and consists of three sections: input, visualization and recommendation panel. Users construct profiles by entering names of artists via an interactive drop-down list (Figure 1a). Based on the mood information associated with profile artists, the system positions a user avatar in a precomputed latent mood space (Figure 1b) and recommends new artists (Figure 1c). In this section we provide an overview of the user interface and explain the method for constructing the mood space.

4.1 Interface Design

Visualization of the latent mood space along with the artists within it is central to solving the problem of navigation through the music collection and explanations of recommendations. The space consists of 266 moods - similar ones being positioned closer to each other than dissimilar ones. Furthermore, moods form a hierarchy with three primary categories at the top - vital, sublime and uneasy (see Table 1). This is portrayed on canvas by showing mood nodes in different categories in red, blue and green 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 our database are placed within the mood space based on moods associated with their

² <http://habumusic.com>

³ <http://musicoverly.com>

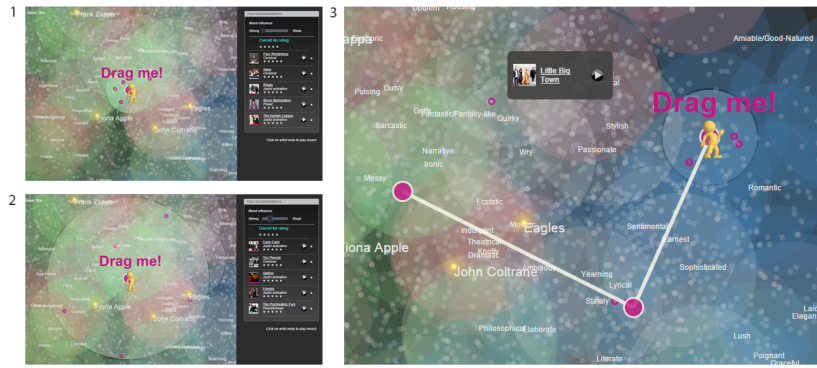


Fig. 2: Screen shots of interactive features in *MoodPlay*: (1) and (2) - the varying recommendation catchment area around user avatar, controlled by a hybridization slider, (3) - trail based interaction, along with a display of artist information box upon clicking on artist node in the visualization.

music. Users can stream their music in real-time and see additional artist information by clicking on the nodes.

Our interface design follows Shneidermann’s visual information seeking mantra “*Overview first, zoom and filter, then details on-demand*” [56]. The hierarchical view of large number of moods allows the user to explore starting from broad terms and filter down selection to specific subset of moods in the visualization. Users are able to zoom in and pan the visualization, and thus more closely inspect areas they are interested in. Furthermore, we implement a dynamic mood labeling algorithm in order to reduce cognitive load in a dense mood space. We rank the moods based on the frequency they are used to describe different artists, and show only limited number of the most popular moods at a given zoom level.

Ordered list of recommended artists is displayed in the right panel (Figure 1) and the corresponding artist nodes are highlighted in the mood space. In this way, we aim to provide transparency, trust, efficiency and satisfaction to the user, which are four out of the seven criteria identified by Tintarev and Masthoff [60] to design explanations in recommender systems (the other three are scrutability, effectiveness and persuasiveness). Items in the recommendation list are linked to audio streams on Rdio⁴ and to Last.fm⁵ profiles of artists. For each recommended artist we also display artist’s picture, color of the top mood category (red, blue or yellow) and name of the sub-category the artist belongs to, with the goal to help users gain some understanding of the music upon visual inspection. Furthermore, because recommended items change as a result of the user’s interaction with the system, we display up or down arrows next to artist name if its position changed, horizontal line if it remained the same or a star if the recommendation is new. Rating of recommended items is enabled

⁴ Rdio streaming service was discontinued in December 2015.

⁵ <http://www.last.fm/api/intro>

only for the purpose of user study, and is achieved by clicking on one of the five stars below artist names.

Adaptivity of music recommenders is particularly important due to the dynamic nature of listening context [58]. Keeping this in mind, we model the gradual change of a user's preference by enabling the movement of the avatar in latent affective space and maintaining the array of trail marks, weighted by distance from the current position (Figure 2.3). As the user navigates away from the initial position, we incorporate the mood information associated with each trail mark into the recommendation algorithm. Removing any of the trail marks is possible by simply clicking on it, and deleting the whole trail is achieved by clicking on the initial position of the avatar.

Finally, fine-tuning of recommendations is further supported by controlling the hybridization of recommendation process. Our 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). Therefore, we allow users to adjust the mood influence via a slider control which dynamically re-sizes a catchment area around the current avatar position (Figure 2.1 and 2.2). The weaker the mood influence, the more we rely on audio similarity to calculate recommendations, and vice-versa.

4.2 A Visual Model of Affect

A key challenge of this research was developing the mood representation and explaining it to the end users. We believe that the challenge of dynamically explaining a complex latent variable space to end users has applications beyond music recommendation. Numerous emotion models, both continuous and categorical have been proposed in the psychology field [26, 53, 55]. For the purpose of identifying potential clusters in our mood space, we explored whether our visual map fits into the hierarchical, and therefore categorical, music-specific emotion model proposed by Zentner *et al.* [72] - GEMS. This model consists of 3 main categories (vitality, uneasiness, sublimity), 9 sub-categories and 45 music relevant emotion words distributed across different sub-categories. Our hypothesis was that such hierarchy should emerge in the visual mood space built upon professionally curated artist-mood associations. It is important to note that psychology researchers focus on deriving emotion models rather than mood models, and for recommendation purposes music is generally tagged with mood descriptors. In this paper we use either of the terms depending on the field we address, and an overarching term, affect, in the context of our proposed system.

To perform our hierarchical classification of moods, we employed a WordNet⁶ similarity tool⁷ and calculated similarity scores between 289 Rovi and 45 GEMS mood words. Furthermore, since similarity between terms in WordNet is based on semantic relatedness and not strictly on synonymity, we evaluated mood classification by subjective observation. For example, the word *volatile* was found to be related more closely to *tender* than *tense* and was placed into sub-category *Tenderness*, rather

⁶ <https://wordnet.princeton.edu/>

⁷ <http://maraca.d.umn.edu/cgi-bin/similarity/similarity.cgi>

Category	Sub-category	No. of moods	Example moods
Sublimity	Tenderness	24	Delicate, romantic, sweet
	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
	Joyful activation	32	Animated, fun, playful, exciting
Unease	Tension	32	Nervous, harsh, rowdy, rebellious
	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
	Cerebral *	12	Detached, street-smart, ironic
	Mechanical *	7	Crunchy, complex, knotty

Table 1: Structure and description of *MoodPlay* mood hierarchy. Categories and sub-categories marked with * are the expansions from the original GEMS model.

than *Tension*. The following steps were taken to reduce the observed classification error rate: (1) we created new mood categories to accommodate moods that do not belong to any of the GEMS categories, (2) 23 of the least frequently used mood words to describe artists in Rovi were discarded, (3) remaining misclassified words were manually placed into categories that they are more likely to belong to. Table 1 shows the final list of categories and distribution of associated moods.

5 Technical Design and Implementation

MoodPlay uses diverse data collected from different sources, mostly through public Web APIs. Recommendations have to be computed very quickly, since they are immediately presented in the interface as a result of user interaction. Therefore, the system requires a special architectural design. As depicted in Figure 3, it has two main components: 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.

5.1 Dataset

MoodPlay relies on a static dataset of 4,927 artists, obtained in several iterations. First, 3,275 artists were randomly selected from a subset of the Million Songs Dataset⁸.

⁸ <http://labrosa.ee.columbia.edu/millionsong/pages/getting-dataset#subset>

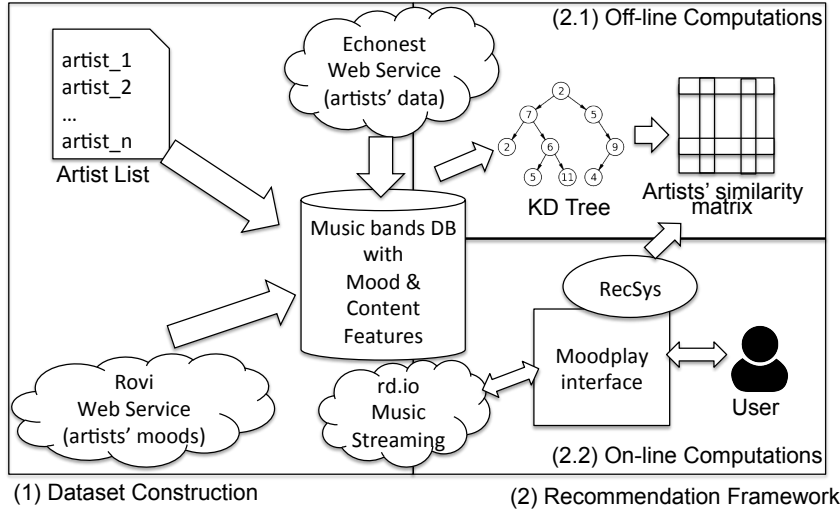


Fig. 3: MoodPlay system architecture indicating the modules for: (1) dataset construction and (2) recommendation framework.

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 popular artists from the public EchoNest⁹ database, as measured by proprietary metrics *familiarity* and *hotness*. We complemented the initial set in order to better facilitate an online user study with participants of different ages from different parts of the world. Artists for which we were not able to obtain mood or song data were discarded. Next, mood data for each artist was obtained via Rovi¹⁰ API and the top ten most popular songs for each artist along with corresponding audio analysis data were obtained from EchoNest. Different interpretations of the same song, having the same title in EchoNest database were discarded. Rdio¹¹ API was used for music streaming in MoodPlay. Finally, artist pictures and links to Last.fm profiles were obtained via Last.fm¹² API.

5.2 Recommendation Approaches

Our hybrid cascading recommender [7] operates in two stages as shown in Figure 4: (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 becomes the input to an audio content-based recommender, which re-ranks the artists

⁹ <http://developer.echonest.com>

¹⁰ <http://developer.rovicorp.com/io-docs>

¹¹ Rdio streaming service was discontinued in December 2015.

¹² <http://www.last.fm/api/intro>

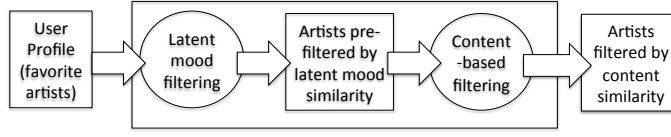


Fig. 4: Schematic representing our hybrid cascading recommender which pre-filters based on mood similarity and then post-filters based on content similarity.

and produces the final recommendation list. Such layered approach supports our goal to help user control and understand how recommendations are generated while navigating mood space. The following paragraphs describe the recommendation process in 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 affective space produced by correspondence analysis. To calculate audio-based similarity, we first identify the 10 most popular songs for each artist in our database via the EchoNest API and obtain audio analysis data for the total of 49,270 songs 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 $v_i \in \mathbb{R}^{515}$ [40] and build artist data into a KD-tree [5]. 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 has a good balance of accuracy, scale and efficiency [40]. 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 $\mathcal{O}(K \cdot \log n)$. To compute artist similarity, first, for each song we rank all other songs from the dataset from most to least similar. We then calculate average similarity rank of songs per artist [8], thus obtaining the artist similarity matrix (Algorithm 1).

Online recommendation. During a user session, MoodPlay recommends new artists similar to the artists the user enters into her profile. First, we determine the overall mood by calculating the centroid $C(u) = (c_x, c_y)$ of profile artist positions, where we then place the user avatar. The coordinates c_x and c_y are calculated as in Equation 1, where P is the set of artists in the user profile, and a_x is the x -axis coordinate of artist a in profile P .

$$c_x = \sum_{a \in P} \frac{a_x}{|P|} \quad c_y = \sum_{a \in P} \frac{a_y}{|P|} \quad (1)$$

Artists found within the adjustable radius around the centroid are all potential candidates for recommendation because they are considered to reflect the latent moods derived from the user’s input. Among the candidate artists, we select the ten most similar to the user profile based on pre-computed audio similarity data, rank them by distance from user position and display first five as recommended artists (Algorithm 2).

Algorithm 1 Algorithm for computation of audio similarity

Require:
 Set of artists: $A = \{a_1, a_2, \dots, a_n\}$
 Set of songs for all artists: $S = \{Sa_1 \cup Sa_2 \cup \dots \cup Sa_n\}$
Ensure: Audio similarity ranks: $ARanks = \{a_i \rightarrow \{a_j \rightarrow rank_{ij}\}\}$

```

1: function COMPUTEAUDIOSIMILARITYRANKS
2:   ARanks = {} ▷ dictionary of artist similarity ranks
3:   for each artist  $a_i$  in  $A$  do
4:     SRanks = {} ▷ dictionary of song similarity ranks
5:     for each song  $s_k$  in  $Sa_i$  do
6:       SRanks[ $s_k$ ] = COMUTESIMILARITYMAPOFSONGRANKS( $s_k, S$ )
7:     end for
8:     for each artist  $a_j$  in  $A$  do
9:       ARanks[ $a_i$ ][ $a_j$ ] = COMPUTEAVERAGESONGSIMILARITY( $SRanks, Sa_j$ )
10:    end for
11:  end for
12:  return ARanks
13: end function

14: function COMUTESIMILARITYMAPOFSONGRANKS( $s, S$ )
15:  Rank all songs from  $S$  based on audio similarity to song  $s$ 
16:  for each  $s_j$  in  $S$  do
17:    similarityMapOfSongRanks[ $s_j$ ] =  $rank_j$ 
18:  end for
19:  return similarityMapOfSongRanks
20: end function

21: function COMPUTEAVERAGESONGSIMILARITY( $SRanks, Sa$ )
22:  average = 0
23:  for each song  $s_i$  in  $Sa$  do
24:    for each song  $s_j$  in  $SRanks.keys$  do
25:      average +=  $SRanks[s_j][s_i]$ 
26:    end for
27:  end for
28:  average = average / ( $Sa.size + SRanks.size$ )
29:  return average
30: end function

```

Trail-based recommendation. Furthermore, we propose a novel, adaptive recommendation approach that accounts for the mood change, reflected by the repositioning of user's avatar in the affective space. We keep track of each new position and apply a decay function to the preference trail when recommending new artists. Recommendations from the last position in the trail are assigned the greatest weight, because we presume that the most recent mood area of interest is the most relevant to user. The weights further 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.

At each trail mark, we apply the recommendation algorithm described in the previous sub-sections, which produces an initial set of recommendation candidates. We then calculate adjusted distances d_a between trail marks and corresponding recommendation candidates in two steps. First, we scale distances between the trail mark and artists because radius can vary among trail marks. If the distances were not nor-

Algorithm 2 Basic algorithm for online music recommendation

Require:
 Artists in user profile: $P = \{a_1, \dots, a_n\}$
 User position: $u = \text{Centroid}(a_1, \dots, a_n)$, $a_i \in P$ or a position from user's trail $u \in T = \{u_1, \dots, u_n\}$
 Recommendation radius: r
 Audio similarity ranks: $ARanks = \{a_i \rightarrow \{a_j \rightarrow \text{rank}_{ij}\}\}$
 Number of recommendations: n_{rec}

Ensure:
 Recommended artists: $R = \{a_1, \dots, a_n\}$

```

1: function RECOMMENDMUSIC( $u$ )
2:    $M = []$  ▷ artists within mood radius
3:   for  $a_i$  in  $A - P$  do
4:     if  $\text{distance}(a_i, u) < r$  then  $M[i] = a_i$ 
5:     end if
6:   end for
7:    $H = \{\}$  ▷ dict. of artists & similarity with  $P$ 
8:   for  $a_i$  in  $M$  do
9:      $H[a_i] = \text{AVERAGESIMRANKING}(a_i, P)$ 
10:  end for
11:   $\text{sort}(H)$  ▷ sort artists by audio similarity
12:   $R = H[1..n_{rec}]$ 
13:  return  $R$ 
14: end function

15: function AVERAGESIMRANKING( $a, P$ )
16:    $\text{average} = 0$ 
17:   for each  $a_i$  in  $P$  do
18:      $\text{average} += ARanks[a][a_i]$ 
19:   end for
20:   return  $\text{average} \div P.size$ 
21: end function

```

malized, many relevant artists would be falsely considered irrelevant and would not appear in the final recommendation list. Next, we adjust the normalized distances for each trail mark based on the corresponding weights using the formula $d_a = d_n + \Delta \times (|T| - 1 - i)$, where d_n is a normalized distance, Δ is a decay constant, $|T|$ is a total number of trail marks and i is an iterator over the trail marks. After several tests, we found that weight constant Δ performs the best when calculated as: $\Delta = r_{min}/4$, where r_{min} is the minimal recommendation radius. The larger the value of Δ , the steeper the decay function is. Finally, the recommendation candidates are sorted based on adjusted distances, and top five are recommended to user.

6 Evaluation

Preliminary evaluation of an early version of MoodPlay has been described in [2]. Compared to the previously published study, here we present modified experiment design and more analysis of results. We performed a crowd-sourced study with entirely new set of participants. 378 users participated in the study and 279 remained after filtering out those that we did not deem as valid, i.e. those that incorrectly answered attention check questions or ended the study prematurely.

Algorithm 3 Hybrid recommendation with provenance trails.**Require:**

Trail of user positions: $T = \{u_1, u_2, \dots, u_n\}$, 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}

Ensure:

Recommended artists: $R = \{a_1, \dots, a_n\}$

```

1: function RECOMMENDMUSICBASEDONTRAIL
2:    $R = \{\}$  ▷ dict. of recommended artists
3:    $\Delta = r_{min}/4$ 
4:   for  $u_i$  in  $T$  do
5:     for  $a_j$  in RECOMMENDMUSIC( $u_i$ ) do
6:        $d_s = \text{SCALE}(\text{distance}(u_i, a_j), r, r_{min})$ 
7:        $d_a = d_s + \Delta \times (T.size - 1 - i)$ 
8:        $R[a_j] = d_a$ 
9:     end for
10:  end for
11:   $\text{sort}(R)$  ▷ sort artists in R by  $d_a$ 
12:  return  $R[1..n_{rec}]$ 
13: end function

14: function SCALE( $d, r, r_{min}$ )
15:    $d_c = \text{Convert } d \text{ from range } [0, r] \text{ to } [0, r_{min}]$ 
16:   return  $d_c$ 
17: end function

```

The focus of the evaluation was to understand the effects of mood-based interactions with a recommendation algorithm and to independently evaluate the influence of the MoodPlay visualization from an explanatory perspective. To improve the previous experiment design, in this study we gave users more freedom to naturally interact with the system and we tracked additional interaction metrics. Furthermore, in the previous study we focused the evaluation on user characteristics, interaction and experience, and placed less attention on ratings-based analysis. Here we report the results of both quantitative and qualitative analysis and address impact of mood based interactions on user experience.

6.1 Setup

As in the previous study, we set up four conditions having different features, as shown in Table 2. The conditions have increasing visual and interaction complexity. Conditions 1 and 2 are based on a preexisting user profile while condition 3 and 4 also allow for user input to the algorithm at recommendation time. Figure 2 shows the interaction mechanisms in the full system, as tested in condition 4.

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
Number of subjects	70	69	70	70

Table 2: Availability of different features per experimental condition. Last row in the table shows the number of valid subjects in each condition.

6.2 Participants

In total, 398 participants took the study, equally distributed across all 4 conditions. The distribution of users who completed valid sessions in 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.

6.3 Study Procedure

Participants accepted the study on Mechanical Turk and were redirected to a Qualtrics¹³ pre-study survey with demographic and propensity related questions. Following this, they were assigned to a random condition and performed the main task. Finally, participants gave qualitative feedback in a post-study survey, also administered through the Qualtrics platform.

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, shown on the left in Figure 1. 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 5 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. To ensure that users spent sufficient time in

¹³ <http://www.qualtrics.com>

the experiment, we displayed non-numerical timer and gave users the opportunity to proceed to the post-study after at least 1.5 minutes of interaction.

🚩 **To-do:** Can we quantitatively compare results from the first and second study?

7 Results

7.1 Interaction and Exploration

The interaction analysis shows important differences in user behavior among the different conditions, which are summarized in Figure 5.

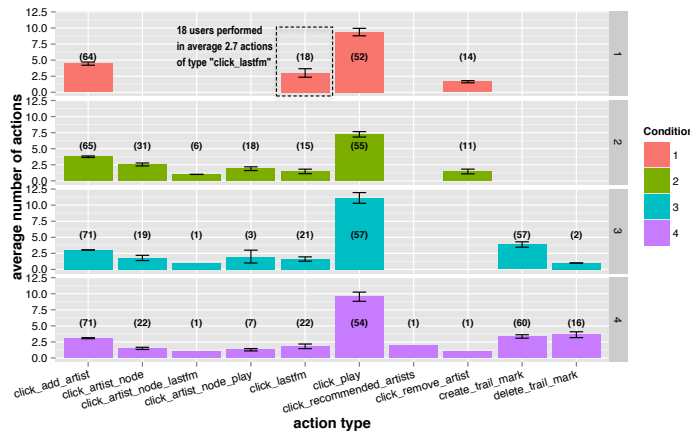


Fig. 5: Average amount of user actions in the four Moodplay conditions: (1) traditional ranked list, (2) visualization without avatar, (3) visualization including avatar, and (4) visualization with avatar and trails. Numbers in parentheses show the amount of users who performed the action.

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 in 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**, in addition to seeing the visualization, users were able to reposition an avatar in the mood space and thus update the recommendations. The interaction with the avatar (*create.trail.mark*) had a negative effect on the number of people who interacted with artist nodes in the visualization. However, more recommendation lists were generated, and users played more recommended artists. On average, users played more artists overall ($M=11.23$, $S.D.=6.28$) and they spent more time in seconds than in all other conditions ($M=446.63$, $S.D.=237.24$). Interestingly, none of the users attempted to update the recommendations by removing artists from their profile.

Finally, in **condition 4**, the system had the same features as in condition 3 with the addition of a provenance trail drawn between avatar positions. Compared to condition 3, users' attention seemed to be diverted to creating and deleting trail marks (action *delete.trail.mark*), and the average number of played artists was lower ($M=9.54$, $S.D.=5.24$). In terms of other 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*) conditions 3 and 4 were rather similar. 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, $S.D.=173.38$).

7.2 Diversity

One of the most interesting results of our study is that the right amount of interaction functionality in a visual interface can promote diversity among the consumed items. We measured this effect by comparing the number of unique artists rated and played per user in each condition. With respect to artists played, we considered "playing activity" in any part of the interface (visualization and recommendation panel) and in the recommendation list only, in order to make a fair comparison against condition 1. Plots in Figure 6 show these distributions. Significant differences were assessed with Wilcoxon signed-rank tests since data departs 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 ($M=7.63$, $S.E.=0.43$), $p < .02$.

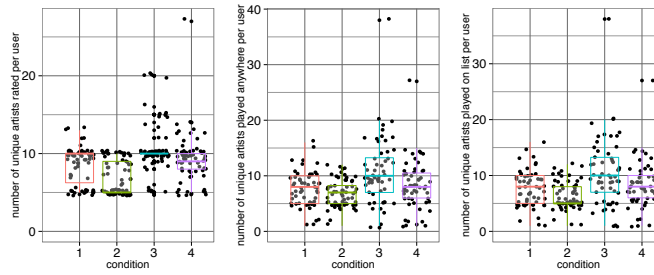


Fig. 6: Consumption of unique items per user: rating-based (left) , and played-based interactions in all interface (center) and on the recommendation list only (right).

⚠️ TO-DO: I skipped improving this section because I am not sure whether we will replace it with NDCG data –Ivana

7.3 Ratings and Ranking

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) 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 8 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 9 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.

In addition to differences in ratings, we analyzed differences in ranking among the different conditions. During the study, users had to rate an initial and a final list of recommendation, and they were free to rate more lists in between. Table [CITATION] shows the average nDCG of the first and last lists at each condition. To analyze the differences in nDCG ranking between conditions, we conducted multiple pairwise t-tests with Bonferroni correction. We found no differences in nDCGs at the initial lists, but by comparing the nDCGs at the end of the study, we found that condition 3 had a significantly larger nDCG ($M=0.58, SE=0.02$) than condition 2 ($M=0.47, SE=0.02$), $p=0.048$. which is consistent with the result on ratings. Since the recommendation

	NDCG and standard error per condition			
	1	2	3	4
First recommendation list	0.58 ± 0.02	0.54 ± 0.02	0.58 ± 0.02	0.55 ± 0.02
Last recommendation list	0.54 ± 0.03	0.47 ± 0.03	0.58 ± 0.02	0.53 ± 0.03

Table 3: Normalized Discounted Cumulative Gain (NDCG) and standard error for the first and last rated recommendation list, per condition.

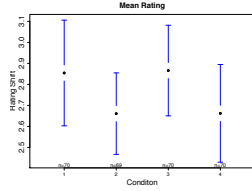


Fig. 7: Mean Final List Rating by Condition (No rating propensity)

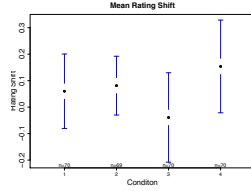


Fig. 8: Mean Rating Shift by Condition (Last - First list rated)

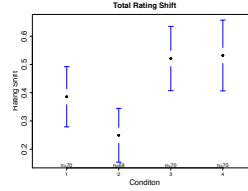


Fig. 9: Total Rating Shift by Condition (Last - First list rated)

algorithm was the same in all four conditions, only the visualization and interaction could explain the observed differences among conditions.

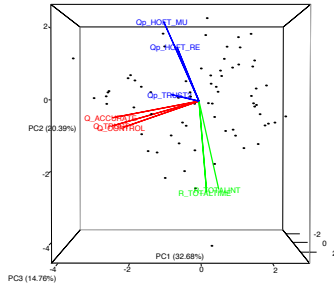
7.4 Connecting Quantitative and Qualitative Evaluation

In order to explore the relationships between quantitative and qualitative experimental results collected during the user study, we performed Principal Component Analysis (PCA) [36] –a technique for dimensionality reduction– over the variables that have shown significant effects in previous studies [32,33,46,6]. Figure 11 shows biplots drawn from the output of PCA conducted at each condition separately, and the variables presented as arrows within the plot are:

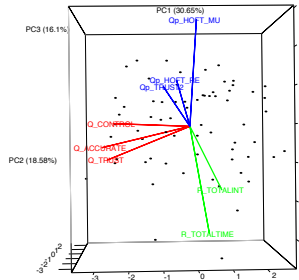
- Qp_HOFT_MU (pre-study question): *How often do you listen to music online?*
- Qp_HOFT_RE (pre-study question): *How often do you use recommender systems?*
- Qp_TRUST2 (pre-study question): *Are you a trusting person?*
- Q_ACCURATE (post-study question): *How accurate do you think the recommendations were?*
- Q_CONTROL (post-study question): *Did you feel in control of the interface?*
- Q_TRUST (post-study question): *How much do you trust the recommendations suggested during the study?*
- R_TOTALINT: Number of user interactions with the system (clicks, music plays, ratings, etc.)
- R_TOTALTIME: Duration of the user study.

For interpreting these plots we used the guidelines described in [36], which states:

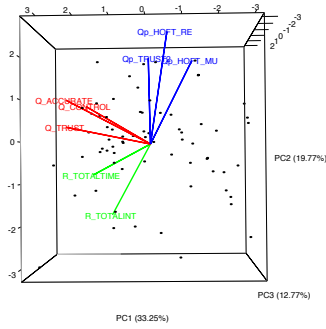
- (a) the length of a vector represents the variance of that variable (within the principal



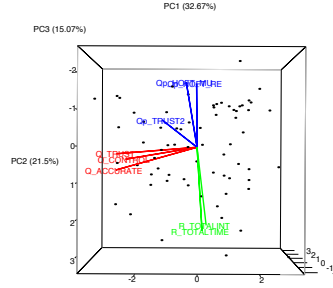
(a) Condition 1: no visualization, static list of recommendations.



(b) Condition 2: visualization without interactive update of recommendations.



(c) Condition 3: visualization with avatar and interactive recommendations.



(d) Condition 4: visualization with avatar, trails and interactive recommendations.

Fig. 11: 3D Biplots for Principal Component Analysis of the experiment variables: (i) pre-study survey (blue), (ii) post-study survey (red) and (iii) user interaction (green).

components used in the biplot), (b) the cosine of the angle between a vector and an axis indicates the importance of the contribution of the variable to the corresponding principal component, (c) the cosine of the angle between pairs of variables have a strong relation with how correlated they are, and (d) uncorrelated variables are at right angles to each other.

Based on these guidelines, we expect that in the conditions where the system had a stronger influence on their final evaluation, the vectors of time in interface and amount of user interaction should be positively correlated with the vectors representing user evaluation.

Condition 1. We observe that the red post-study variables (Q_CONTROL, Q_TRUST, Q_ACCURACY) are strongly correlated between each other. Also, they have little relation to the number of interactions (R_TOTALINT) and the duration of the study (R_TOTALTIME) because they are almost orthogonal, implying that the amount of time and interaction with the interface had small effect on explaining their user experience with the system. Instead, blue-colored post-study variables are loading in the same direction as users' pre-existing level of trust (Qp_TRUST2), although the

short length of this vector tells us that its total variability is not well explained by principal components PC1, PC2 and PC3. Finally, the familiarity of user with music (Qp_HOFT_MU) and how often they listen to music online (Qp_HOFT_RE) are strongly correlated, but they do not explain either the variability of post-study variables.

Condition 2. Similar to condition 1, the initial levels of trust (Qp_TRUST2) and familiarity with recommendation systems (Qp_HOFT_RE) load in the same direction as perceived control, trust and accuracy (Q_CONTROL, Q_TRUST, Q_ACCURACY) in PC1. Interestingly, Q_CONTROL, and Q_ACCURATE, and Q_TRUST load in the same direction with respect to PC1, control departs in PC2 from the other two variables. This implies that the user perception on accuracy and trust diverted from the perception on controllability, compared to condition 1, for instance. This observation allows to explain some previous negative results in condition 2, since users could explore the mood space visualization, but they could not update the list of recommendations by interacting with it.

Condition 3. The preference of users for this condition, shown in the previous analyses, could be explained holistically with this plot. We see that this is the only subplot where Q_TOTALTIME and Q_TOTALINT load in the same direction than Q_CONTROL, Q_ACCURATE and Q_TRUST in PC1, the PC which explains most of the variance. Notably, the acute angle between the performance variables in red and Q_TOTALTIME shows that the amount of time that users spend on the interface explains the performance variables more than in any other interface, but specially Q_TRUST. This is an important result since it validates our hypothesis that both visualization and interaction help to increase the trust that user have in the recommendation system. It can be also observed that the angle between Qp_TRUST2 (the initial level of user trust) is less correlated to the final level of user trust Q_TRUST than R_TOTALTIME.

Condition 4. This condition, just like in condition 1 and unlike condition 3, shows a disconnection between the amount of time and interaction with the final perception of the user in terms of Q_ACCURATE, Q_CONTROL and Q_TRUST over the system. The acute angle between Qp_TRUST2 and Q_TRUST shows that user's inherent trust determines more her final trust on the system than the time spent on it. The plots also show little relation between the user experience with recommender and the amount of user interaction, suggesting that the user was actually confused and not really taking advantage of advanced features on the interface. This can explain the drop in user satisfaction of this interface compared to condition 3.

7.5 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 12 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.

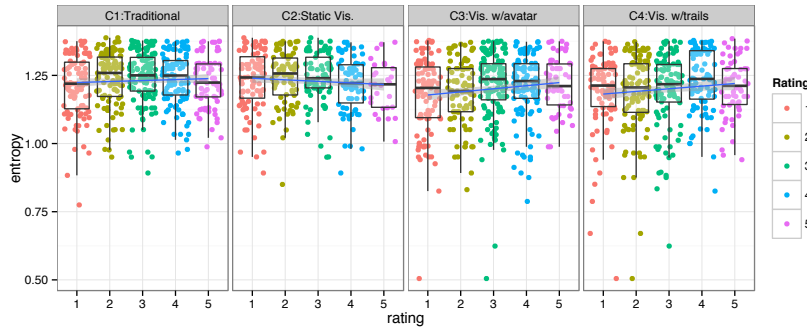


Fig. 12: Entropy-based interaction results.

7.6 Mood Preference

To-do: If the previous subsection get removed, change the first sentence here

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 $V = w_0, \dots, w_k$, $U = w_0, \dots, w_m$, $S = w_0, \dots, w_l$, and $O =$

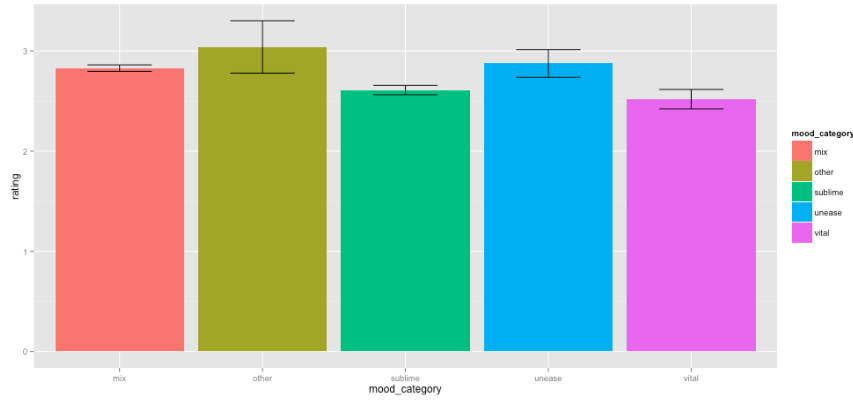


Fig. 13: Average ratings of artists belonging to different mood categories (*vital*, *uneasy*, *sublime* and *other*) and those that do not have a dominant category (*mix*).

w_0, \dots, 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:

$$v = \frac{\sum_{w \in V}}{\sum_{w \in A}}, u = \frac{\sum_{w \in U}}{\sum_{w \in A}}, s = \frac{\sum_{w \in S}}{\sum_{w \in A}}, o = \frac{\sum_{w \in O}}{\sum_{w \in A}} \quad (2)$$

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 13 shows average artist ratings for each one of the groups. Group *other* consists of only 25 artists, significantly less than the remaining groups, and therefore we don't use it in the analysis. We can see in the Figure 13 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 than 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 14). 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 *uneasy* ($M=3.06$, $S.E.=0.21$) 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* ($M=2.5$, $S.E.=0.08$), $p = 0.0005$.

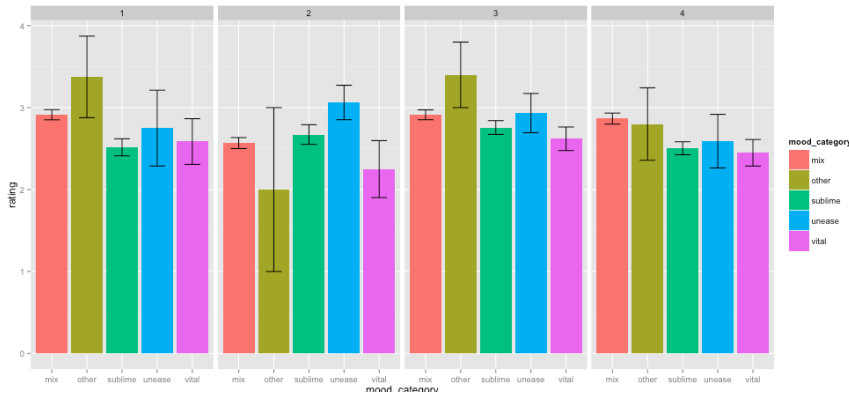


Fig. 14: Average ratings of artists belonging to different mood categories (*vital*, *uneasy*, *sublime* and *other*) and those that do not have a dominant category (*mix*), per experimental condition.

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. [15] [16] Ivana]Clarify previous sentence 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.

7.7 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 disagreement and agreement respectively. For simplicity, we report only answers to a set of the most relevant questions here (Table 4).

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.

⚠ **To-do:** Does it? The values are below 50. If 0 = disagree and 100 = agree, everything below 50 is disagreement.

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

 **To-do:** What about statistical test in this table?? –DP

Table 4: Summary of the most relevant variables in the post-study survey. Numbers indicate average user agreement (on a scale from 1-100) with mean \pm S.E.

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.

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.

7.8 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 5 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	<p>All good.</p> <p>It was really fun.</p> <p>I enjoyed using this!</p>	<p>Add more bands/artists to the search- for example, neither Silversun Pickups nor Smashing Pumpkins were found to add to my list.</p> <p>The recommendations didn't seem to match the artists I chose.</p> <p>Show more information on how the mood of a song/artist is determined.</p>
2	<p>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</p> <p>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.</p> <p>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.</p>	<p>I put in 3 rappers and it gave me like oldies and pop songs. Genre plays roles in certain moods.</p> <p>It runs a little slow, should improve optimization for older computers.</p> <p>I really didn't understand it.</p>
3	<p>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.</p> <p>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.</p> <p>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.</p>	<p>Make the interface simpler and more concise. Speed up loading times</p> <p>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.</p> <p>Larger music selection, possibly change the strong week slider, to broad or specific to the particular mood you are feeling.</p>
4	<p>its a cool design</p> <p>Neat program! If I could practice with it more I think I would really enjoy it.</p> <p>It was excellent! Thanks to the developers for developing wonderful tool.</p>	<p>Some of recommended artists didn't relate to my mood close enough.</p> <p>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.</p> <p>Make the interface faster and smoother. There was too much chopiness when I was using the visualization tool.</p>

Table 5: Selected positive and negative user feedback grouped by experimental condition.

8 Future Work

There is a fertile ground for expansions and branching of this research in several directions. The overarching idea is to build a system that recommends music according to user's musical taste, and guides the user from her current mood to the desired (target) mood. In this section we describe four main categories of future work.

Identifying user's mood and musical preference. As of now, users build their profile by manually selecting several artists that they like and we make recommendations based on the overall mood of the profile. We argue that it is acceptable to use mood data on artist level, rather than on song level, because multiple moods associated with each artist in our database describe its repertoire of songs. Nevertheless,

using individual songs as an input and recommending tracks accordingly could yield greater precision. Another important consideration is that user's profile in MoodPlay may correspond more to the musical taste than to the current mood. The taste and preference based on current mood can be treated as separate, but related parameters. Both can be determined by explicitly asking user to provide the information or implicitly, based on relevant data that has been collected automatically. In the following paragraphs, we focus on ideas for determining current preference as reflected by mood, which could model user's musical taste as well if tracked over longer period of time.

An approach used by some commercial recommendation systems is to let users type in a mood or select it from a predefined list. This is not always an efficient method and easy task for users given the large number of available mood tags. In particular, mood data in our system is very detailed and attempts to capture nuances that characterize different artists (e.g. rowdy, playful, graceful, elegant), whereas typical user uses smaller vocabulary and less specific words to describe mood of the music (e.g. happy, sad, energetic, calm). An alternative approach is to use mood hierarchy embedded in MoodPlay to pick a top mood category from a list, followed by a subcategory and finally select specific mood.

Implicitly determining user's mood in an automated fashion on a granular level is even more challenging. But an implicit approach can be effective if used with less specificity because it can entirely free the user from interaction. If greater granularity is desired, it can be improved by asking for minimal user's input. Extensive research in affective computing briefly outlined in section 3.1 informs multiple ways of improving MoodPlay to collect mood data. For example, user's mood and current preference could be determined from 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.

Identifying target mood. It is not always desired to play music in a user's current mood. Instead, listeners may be interested in hearing music that changes how they feel. For example, happy music can uplift listener in a sad mood, but some may enjoy bitter-sweet music when heart broken, while others would prefer springlike or playful songs. Target mood largely depends on a personal preference and current conditions, and therefore the recommender requires user's input or highly advanced sensing algorithm to determine it.

Depending on the listening context and preference, the recommender can either suggest music in the target mood or find and follow a path from current to target mood. Commercial recommendation systems already offer playlists for different moods and activities (e.g. mellow, music for work or gym), which are effective for short term, action based listening. However, to our knowledge, there are no recommenders that allow transitions from one state to another or adapt to changes in how user feels or changes in listening context.

Adaptive recommendation systems have been an active area of research in recent years. Looking beyond their applications in entertainment, adaptive music recommenders can be of particular value in music therapy. Up to date studies have shown positive effects of music on recovery of movement (e.g. in patients with stroke or Parkinson's disease) and speech [59]. Music therapy with the goal to modulate emo-

tions has been studied less extensively, but the benefits in pain and mood management have been documented [57,28]. Current version of MoodPlay has attracted interest from music therapists because its engaging interface can aid choosing music during therapy sessions for hospitalized children and elderly people with dementia. However, in a broad sense, adaptive recommendation systems can provide subtle but profound effect of music on a listener's well being, outside of therapeutic setting as well. By being able to continuously monitor feedback about a user's state and context, and adapt to changes, the therapeutic benefits of music can be improved.


✂ John, I initially wrote here "can become more ubiquitous" because I wanted to say that benefits can be experienced more outside of therapeutic setting. DId it not sound right?]

Path from one mood state to another. Our proposed trail algorithm is a crude way to create a path from one mood state to another and generate recommendations accordingly. Through the extensive evaluation of MoodPlay it was observed via numerous metrics that users preferred recommendations obtained by navigating music collection freely, over the recommendations given by a 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 engage more or less with the system. In cases when user chooses to build the trail, recommending items along the trail (in between the 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.

Recommendation algorithm and interface. 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.

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.

 **To-DO:** I believe John said he would add more content related to dimensionality reduction and recommendation –*Ivana*

Different dimensionality reduction techniques may yield different layout of moods, and therefore different recommendations. Although it may be tempting to compare the current mood layout to commonly used Russel’s mood model, our goal was to create visual hierarchy of moods that would encompass moods that are difficult to describe on valence-arousal scale, help users locate and explore moods via hierarchy and make a step forward in researching different mood scales and categories.

9 Conclusion

This paper presented and evaluated *MoodPlay* –a hybrid recommender system for musical artists which introduces a novel latent space visualization containing mood information for each artist. The system supports explanation and control of a recommender system through manipulation of an avatar within the visualization. Design and implementation of an online experiment (N=279) was presented to evaluate the *MoodPlay* system against a benchmark. Four conditions were evaluated with varying degrees of visualization, interaction and control. Results of this study show that a low-dimensional visualization of mood information significantly improves user acceptance and understanding of both the underlying data and the recommendations. Allowing users to interact with the visualization, for example to position their avatar as input to the recommender algorithm, produces an additional improvement in these metrics. However, user experience was lower across all metrics in the more complex condition with trail-based interactions, indicating that cognitive overload was a factor. This observation was backed up through an analysis of feedback comments from users. To conclude, the experiment highlighted that interaction and visualization with mood data is beneficial to recommender systems, however, caution must be taken in both UI and interaction design to avoid cognitive overload, for example by supporting system awareness of a user’s mood, context and limitations.

Acknowledgements This work was partially supported by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053; The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL, NSF, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on. The author Denis Parra was funded by Chilean research agency Conicyt, Fondecyt grant number 11150783.

References

1. D. Amelynck, M. Grachten, L. van Noorden, and M. Leman. Toward e-motion-based music retrieval a study of affective gesture recognition. *T. Affective Computing*, 3(2):250–259, 2012.
2. I. Andjelkovic, D. Parra, and J. O’Donovan. Moodplay: Interactive mood-based music discovery and recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, UMAP ’16, pages 275–279, New York, NY, USA, 2016. ACM.

3. C. Baccigalupo and E. Plaza. Case-based sequential ordering of songs for playlist recommendation. In *Advances in Case-Based Reasoning*, pages 286–300. Springer, 2006.
4. L. Baltrunas and X. Amatriain. Towards time-dependant recommendation based on implicit feedback. In *Workshop on context-aware recommender systems (CARS'09)*, 2009.
5. J. L. Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18(9):509–517, Sept. 1975.
6. S. Bostandjiev, J. O'Donovan, and T. Höllerer. Tasteweights: a visual interactive hybrid recommender system. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 35–42. ACM, 2012.
7. R. Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, Nov. 2002.
8. O. Celma and P. 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.
9. L. Chen and P. Pu. Interaction design guidelines on critiquing-based recommender systems. *User Modeling and User-Adapted Interaction*, 19(3):167–206, 2009.
10. G. L. Collier. Beyond valence and activity in the emotional connotations of music. *Psychology of Music*, 35(1):110–131, 2007.
11. S. Cunningham, S. Caulder, and V. Grout. Saturday night or fever? context-aware music playlists. *Proc. Audio Mostly*, 2008.
12. B. Faltings, P. Pu, M. Torrens, and P. Viappiani. Designing example-critiquing interaction. In *Proceedings of the 9th international conference on Intelligent user interfaces*, pages 22–29. ACM, 2004.
13. I. Fernández-Tobías, I. Cantador, and L. Plaza. An emotion dimensional model based on social tags: Crossing folksonomies and enhancing recommendations. In *E-Commerce and Web Technologies*, pages 88–100. Springer, 2013.
14. N. 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.
15. J. M. George. Individual differences and behavior in organizations. chapter Trait and State Affect, page 145. San Francisco: Jossey-Bass, 1996.
16. 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.
17. G. Gonzalez, J. L. De La Rosa, M. Montaner, and S. Delfin. Embedding emotional context in recommender systems. In *Data Engineering Workshop, 2007 IEEE 23rd International Conference on*, pages 845–852. IEEE, 2007.
18. L. Gou, F. You, J. Guo, L. Wu, and X. L. 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.
19. B. Gretarsson, J. O'Donovan, S. Bostandjiev, C. Hall, and T. Höllerer. Smallworlds: Visualizing social recommendations. In *Computer Graphics Forum*, volume 29, pages 833–842. Wiley Online Library, 2010.
20. D. Griffiths, S. Cunningham, and J. Weinl. A discussion of musical features for automatic music playlist generation using affective technologies., 2013.
21. B.-j. Han, S. Rho, S. Jun, and E. Hwang. Music emotion classification and context-based music recommendation. *Multimedia Tools and Applications*, 47(3):433–460, 2010.
22. N. Hariri, B. Mobasher, and R. Burke. Context-aware music recommendation based on latent topic sequential patterns. In *Proceedings of the Sixth ACM Conference on Recommender Systems, RecSys '12*, pages 131–138, New York, NY, USA, 2012. ACM.
23. C. He, D. Parra, and K. Verbert. Interactive recommender systems: a survey of the state of the art and future research challenges and opportunities. *Expert Systems with Applications*, 2016. in press.
24. Y. Hijikata, Y. Kai, and S. 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.
25. D. Hume. Emotions and moods. *Organizational behavior*, pages 258–297, 2012.
26. C. Izard. *Human Emotions*. Emotions, Personality, and Psychotherapy. Springer, 1977.
27. J. H. Janssen, E. L. van den Broek, and J. 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.
28. P. N. Juslin and P. Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

29. M. Kaminskas and F. Ricci. Location-adapted music recommendation using tags. In *User Modeling, Adaption and Personalization*, pages 183–194. Springer, 2011.
30. M. Karg, K. Kühnlenz, and M. Buss. Recognition of affect based on gait patterns. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 40(4):1050–1061, 2010.
31. A. Kleinsmith and N. Bianchi-Berthouze. Recognizing affective dimensions from body posture, 2007.
32. B. P. Knijnenburg, S. Bostandjiev, J. O'Donovan, and A. Kobsa. Inspectability and control in social recommenders. In *Proceedings of the sixth ACM conference on Recommender systems*, pages 43–50. ACM, 2012.
33. B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction*, 22(4-5):441–504, 2012.
34. S. Koelsch. A neuroscientific perspective on music therapy. *Annals of the New York Academy of Sciences*, 1169(1):374–384, 2009.
35. J. A. Konstan and J. Riedl. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction*, 22(1-2):101–123, 2012.
36. P. M. Kroonenberg. *Applied multiway data analysis*, volume 702. John Wiley & Sons, 2008.
37. B. Logan. Music recommendation from song sets. In *ISMIR*, 2004.
38. F. Mailliet, D. Eck, G. Desjardins, P. Lamere, et al. Steerable playlist generation by learning song similarity from radio station playlists. In *ISMIR*, pages 345–350, 2009.
39. J. Masthoff. The pursuit of satisfaction: affective state in group recommender systems. In *User Modeling 2005*, pages 297–306. Springer, 2005.
40. B. McFee and G. R. G. Lanckriet. Large-scale music similarity search with spatial trees. In A. Klapuri and C. Leider, editors, *ISMIR*, pages 55–60. University of Miami, 2011.
41. S. M. McNee, J. Riedl, and J. 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.
42. S. Nagulendra and J. 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.
43. J. O'Donovan, B. Smyth, B. Gretarsson, S. Bostandjiev, and T. Höllerer. Peerchooser: visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1085–1088. ACM, 2008.
44. H.-S. Park, J.-O. Yoo, and S.-B. 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.
45. D. Parra and X. 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.
46. D. Parra and P. Brusilovsky. User-controllable personalization. *Int. J. Hum.-Comput. Stud.*, 78(C):43–67, June 2015.
47. D. Parra, P. Brusilovsky, and C. 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.
48. R. J. D. Paul Ekman. *The Nature of Emotion: Fundamental Questions*. Oxford University Press, 1994.
49. V. A. Petrushin. Emotion recognition in speech signal: experimental study, development, and application, 2000.
50. P. Pu, B. Faltings, L. Chen, J. Zhang, and P. Viappiani. Usability guidelines for product recommenders based on example critiquing research. In F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, editors, *Recommender Systems Handbook*, pages 511–545. Springer US, 2011.
51. S. Rho, B.-j. Han, and E. 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.
52. J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980.
53. K. R. Scherer. *On the nature and function of emotion: A component process approach.*, chapter 14, pages 293–317. Lawrence Erlbaum, Hillsdale, NJ, 1984.
54. K. R. Scherer, T. Johnstone, and G. Klasmeier. Vocal expression of emotion. *Handbook of affective sciences*, pages 433–456, 2003.

55. U. Schimmack and A. Grob. Dimensional models of core affect: A quantitative comparison by means of structural equation modeling. *European Journal of Personality*, 14(4):325–345, 2000.
56. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.
57. S. L. Siedliecki and M. Good. Effect of music on power, pain, depression and disability. *Journal of Advanced Nursing*, 54(5):553–562, 2006.
58. S. Stober and A. Nürnberger. Adaptive music retrieval—a state of the art. *Multimedia Tools Appl.*, 65(3):467–494, Aug. 2013.
59. M. H. Thaut and G. C. McIntosh. Neurologic music therapy in stroke rehabilitation. *Current Physical Medicine and Rehabilitation Reports*, 2(2):106–113, 2014.
60. N. Tintarev and J. Masthoff. Designing and evaluating explanations for recommender systems. In *Recommender Systems Handbook*, pages 479–510. Springer, 2011.
61. M. Tkalčič, U. Burnik, and A. Košir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, 20(4):279–311, 2010.
62. M. Tkalčic, A. Kosir, and J. 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.
63. M. D. van der Zwaag, J. H. Janssen, and J. H. D. M. Westerink. Directing physiology and mood through music: Validation of an affective music player. *T. Affective Computing*, 4(1):57–68, 2013.
64. D. Västfjäll. Emotion induction through music: A review of the musical mood induction procedure. *Musicae Scientiae*, 5(1 suppl):173–211, 2002.
65. K. Verbert, D. Parra, P. Brusilovsky, and E. 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.
66. X. Wang, D. Rosenblum, and Y. Wang. Context-aware mobile music recommendation for daily activities. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 99–108. ACM, 2012.
67. H. M. Weiss and R. Cropanzano. Affective Events Theory: A theoretical discussion of the structure, causes and consequences of affective experiences at work. 1996.
68. S. Wu, T. H. Falk, and W.-Y. Chan. Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, 53(5):768–785, 2011.
69. Y.-H. Yang, Y.-C. Lin, H. T. Cheng, and H. H. Chen. Mr. emo: music retrieval in the emotion plane. In A. El-Saddik, S. Vuong, C. Griwodz, A. D. Bimbo, K. S. Candan, and A. Jaimes, editors, *ACM Multimedia*, pages 1003–1004. ACM, 2008.
70. F. Yu, E. Chang, Y. qing Xu, and H. yeung Shum. Emotion detection from speech to enrich multimedia content, 2001.
71. M. Zentner and T. EEROLA. Self-report measures and models. *Handbook of Music and Emotion: Theory, Research, Applications*, 2011.
72. M. Zentner, D. Grandjean, and K. R. Scherer. Emotions evoked by the sound of music: Characterization, classification, and measurement. *Emotion*, 8(4):494–521, 2008.
73. S. Zhao, M. X. Zhou, X. Zhang, Q. Yuan, W. Zheng, and R. 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.