Kim et al. *EPJ Data Science* (2020) 9:13
https://doi.org/10.1140/epjds/s13688-020-0221-9

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE

Open Access

PepMusic: motivational qualities of songs for daily activities

Yongsung Kim[1*], Luca Maria Aiello[2] and Daniele Quercia[2,3]

*Correspondence:
yk@u.northwestern.edu
[1] Northwestern University, Evanston, USA
Full list of author information is available at the end of the article

Abstract

Music can motivate many daily activities as it can regulate mood, increase productivity and sports performance, and raise spirits. However, we know little about how to recommend songs that are motivational for people given their contexts and activities. As a first step towards dealing with this issue, we adopt a theory-driven approach and operationalize the Brunel Music Rating Inventory (BMRI) to identify motivational qualities of music from the audio signal. When we look at frequently listened songs for 14 common daily activities through the lens of motivational music qualities, we find that they are clustered into three high-level latent activity groups: calm, vibrant, and intense. We show that our BMRI features can accurately classify songs in the three classes, thus enabling tools to select and recommend activity-specific songs from existing music libraries without any input required from user. We present the results of a preliminary user evaluation of our song recommender (called PepMusic) and discuss the implications for recommending songs for daily activities.

Keywords: Music; Music recommendation; Motivational qualities; Daily activities

1 Introduction

Music captures attention, raises spirits, triggers and regulates emotions, and increases work output [1, 2]. To arouse the desired feelings, the type of music should match the type of activity [3]. For example, the music people commonly listen to when seeking motivation for a workout is usually different from the music one needs to delve into relaxation. Accordingly, people curate their activity-specific playlists either by putting together songs they deem appropriate, which might be time consuming or bothersome, or by drawing from existing popular playlists that have been suitably composed for the desired activity, which may lack personalization.

In an attempt to meet these user needs, previous work has looked into automatically recommending songs suited for a specific activity [4–7]. Many of them are dependent on a variety of signals [8] including music genre [9], popularity of the song [10], or demographic information of the user [11], which limits their generality. While others use audio signals, but they still focus on single activity, for example, recommending songs for running sessions [12]. This motivates the need for ways to recommend songs that are motivational for various activities.

A key challenge is thus to understand and identify which musical properties are motivational for which activity. Leveraging existing listening histories and their associated preference ratings could be a starting point [13]; however, the songs people like to listen to might not be motivational in the context of certain activities. One may survey and crowdsource user preferences by asking people to rate whether or not the song will be motivational for a given activity, but that would require large resources to acquire a large-scale dataset.

To address this challenge, we introduce the idea of propagating activity labels for songs by using latent "motivational" characteristics that can be identified from audio features. As opposed to previous approaches, we do not find similar songs that people listened to in the past for a given activity, but we do rely on latent "motivational" characteristics to recommend songs that may "motivate" people for the activity.

To achieve this goal, our main contributions depart from previous work in three main aspects:

- Our selection of audio features is directly informed by the music psychology literature. For the first time, we operationalize the Brunel Music Rating Inventory (BMRI) [14], an instrument to assess the motivational qualities of music in exercise and sport, which we extend to other activities (Sect. 4).
- We consider a set of common daily activities coming from established literature, and we map frequently listened songs during these activities to the motivational sound feature space and use clustering to identify prototypical motivational range for the activities. We find that these 14 common daily activities naturally fall into three main clusters representing three music archetypes: *calm*, *vibrant*, and *intense* (Sect. 5).
- We train a "motivation-based" classifier to map song into those three archetypes (Sect. 6). We found that, with our best performing classifier, it achieved 88.9% accuracy for *calm* group, 86.7% for *vibrant* group, and 86.5% for *intense* group.

The rest of the paper is organized as follows. We first review related work to motivate the need for extracting audio-signal based on motivational qualities and identifying prototypical motivational ranges based on these audio-signal (Sect. 2). We then introduce our data collection process that led to 1k+ songs and their metadata from Spotify and YouTube for 14 common daily activities identified in previous literature [15, 16] (Sect. 3). We present methods for extracting audio signal based on BMRI (Sect. 4), clustering songs to identify motivational range (Sect. 5), and training motivation-based classifiers (Sect. 6). We present a preliminary user evaluation to assess our "motivation-based" classifiers by asking users to rate whether or not they found the songs were good for each activity group (Sect. 7). We conclude our paper with the discussion of theoretical implications of gaining a better empirical understanding of the relationship between the motivational properties of music and daily activities, as well as practical implications of using such recommendations throughout a user's daily life (Sect. 8).

2 Related work

Our research builds upon the literature in a variety of fields from music psychology and sports psychology, to music recommender systems, to music information retrieval.

Our daily activities can benefit much by listening to music. Research from music psychology and sports psychology shows that music can regulate moods and emotions [17–19], increase productivity, increase the intensity or endurance of exercise [20, 21], encourage rhythmic movement, and evoke memories and raise spirits [2]. Motivated by this line of theoretical work, we base our classifier on features informed by psychometric measures reflecting motivational properties of music (Sect. 4).

Researchers have sought to improve music recommendation systems by incorporating different factors such as user context, user properties, and music content [8]. User context factors include location and time [22–24], physiological state [5], and emotion [25, 26]. User properties include demographics [11], listening histories [10], and users' music play sequence [27]. Music content factors include genre and artists [9], popularity [10], and music audio features [12, 13].

There exists music recommender systems that recommend songs to motivate specific activities, such as driving [6], running [4, 5], working [7], and traveling [24]. Baltrunas et al. study ways to incorporate factors such as driving style, mood, road type, weather, and traffic conditions to recommend songs for driving [6]. Systems like PersonalSoundTrack [4] and TripleBeat [5] use runner's pace [4] and physiological state to recommend songs to motivate runners. FocusMusicRecommender estimates user's concentration level and recommends songs to help users focus on work [7]. We are motivated by this prior work and we envision future music recommender systems that can recommend different sets of songs to motivate users' current activities—especially considering that user activities will soon be more readily detectable thanks to the advance in context-aware computing and sensing capabilities.

A few approaches recommend songs for common activities by using audio features [12, 13]. Core difference between prior work and our work is that we operationalize psychometric measures to extract music features that are related to motivational qualities. When we map the frequently listened songs for 14 common daily activities to motivational music feature space, we find that these 14 activities can be grouped into three latent activity groups: calm, vibrant, and intense (Sect. 4). The number of groupings resembles that of Yadati et al. [13] in which they also identified three high-level activity groups (relaxing, studying, exercising).

3 Data collection

We first need to choose a list of daily activities and pair them with songs that people listen to while engaged in those activities. Previous work in music information retrieval either defined an arbitrary set of activities [12], or mined user-generated content from platforms like Youtube to cluster activities that are frequently mentioned [13]. To ground our selection in established literature, we also relied on previous work that identified comprehensive taxonomies of daily activities that are *generally* conducted indoors or outdoors (with no specific relation to music) [15, 16]. We found that the intersection of these two activity sources results in eleven main daily activities: intimate relations, socializing, pray and worship, relaxing, eating, preparing food, exercising, shopping, working, commuting, and napping.

To gather songs that are frequently listened to while engaging in these activities, we resorted to Spotify. Spotify is an appropriate database for our purpose because it is a widespread service (180 million monthly active users all over the world), and it publicly exposes playlists curated by a variety of users, along with rich metadata. We chose a simple set of keywords for each activity. If an activity was a verb, the keywords were the verb and the verb+"ing" (e.g., if an activity was driving, we searched for both "drive" and "driving"). If an activity was noun, we only queried the activity in its own noun form (e.g., if an activity was "office", we only searched for "office"). We first submitted each keyword of an activity as a query to the Spotify search API[a], and collected the top 100 among the returned playlists. This provided a wide coverage of both popular and rarer songs that people listen to when engaged in a certain activity. Because the retrieval policy of Spotify search was not transparent, the set of returned playlists was likely ranked according to a mix of factors, including not only the relevance but also the popularity or prestige of the playlist owner. To only include playlists that were relevant to the activity query, we filtered out those that contained the corresponding search term in neither their names nor in their description. We then retrieved the metadata of the songs contained in the remaining playlists and retained only the songs that occurred in at least two playlists. This allowed us to filter out songs that may reflect strong personal tastes and may not be necessarily associated with a specific activity.

We wanted to retain activities that have at least 100 unique songs for each activity for clustering and classification purposes. Based on preliminary observations for a few distinctive activities (such as running and sleeping), we found that 50 songs already provided a strong signal to distinguish activities; for robustness, we doubled that number. With this last filtering step, shopping, praying, cooking, napping, and socializing were excluded. Therefore, we replaced napping with sleeping, and socializing with partying and drinking. We also found that activity terms such as eating and working were associated with playlists with other purposes; for example, the most common playlist names and descriptions associated with eating were about eating disorder, and those associated with working concerned working out. Hence, we replaced eating with breakfast, lunch, and dinner, and replaced working with studying and office. Finally, we also expanded commuting with commuting and driving, and expanded exercising with exercising and running. With this procedure, we ended up with 14 common daily activities, namely, *relaxing and sleeping, exercising, running, office, partying, drinking, sex, commuting, driving, breakfast, lunch, dinner, and studying.*

As the Spotify API did not return audio files, we searched and downloaded each song on YouTube with a query composed by a song title and an artist, separated by a whitespace. We chose a song where the difference between the duration of the YouTube audio file and the song duration from the Spotify metadata was the smallest. This was an important step

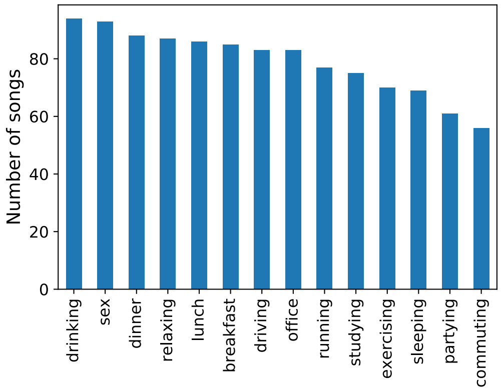


Figure 1 The number of songs per activity in our dataset. The dataset includes a total of 1107 songs for 14 common daily activities

because even if it had the same title and artist the audio file downloaded could be quite different depending on the duration (e.g., a music video with a long narrative at the beginning of the song or a video from a live concert in which the artist talks to the audience). We manually inspected 300 songs at random to ensure that these songs were matching the title correctly. Among these 300 songs, 91.3% of them (274 out of 300) matched the title correctly. All mismatched songs belonged either to the relaxing or the sleeping category. The main cause of these mismatches (16 out of 26 cases) was the song not being available on YouTube due to copyright restrictions or to the low popularity of the artist or album (e.g., artist names or album names like "Study Music", "Einstein Study Music Academy"). In the remaining 10 cases, 2 were "soft" mismatches (a slightly different version of the same song was selected), and the remaining 8 were actual mismatches, which accounted for about only 3% of the cases. We randomly sampled 100 songs from each activity, and we managed to collect 1107 songs for the 14 common daily activities overall (Fig. 1).

4 Operationalizing Brunel Music Rating Inventory

The Brunel Music Rating Inventory (BMRI) is a psychometric measure to assess the motivational qualities of music in the exercise and sport domain. Factors that determine motivational qualities of music are rhythm response (i.e., rhythmical elements of music), musicality (i.e., pitch-related elements of music), cultural impact, and association [14]. Elements for the *rhythm response* factor include: rhythm, stimulative qualities of music (loudness and tempo [14]), and danceability. Elements for the *musicality* factor include: harmony (how the notes are combined), and melody (the tune). *Cultural impact* refers to the effect of music on an individual's cultural experiences, whereas *association* refers to "extra-musical thoughts, feelings and images that the music may evoke [14]."

From the audio signal, we can extract music elements related to the first two factors out of the four. More specifically, we extracted music elements related to rhythm, tempo, harmony, melody, stimulative qualities of music (loudness and tempo [2]), and danceability. For each element, we use well-established third-party libraries or state-of-the-art music information retrieval techniques for accurate descriptors (Table 1).

For *rhythm*, we use Rhythm Patterns and Rhythm Histogram as descriptors. Rhythm Patterns describe amplitude modulations for a range of modulation frequencies (e.g., fluctuations or rhythm) on frequency bands that are within the human audible range. The algorithm computes a power spectrum that reflects human loudness sensation on 24 "critical bands"; see [28] for more details about how the algorithm transforms the spectral data

Table 1 Descriptors for each music elements that are being extracted as music feature, and its dimension

Factor	Music element	Descriptors
Rhythm Response	Rhythm	Rhythm Histogram [28] Rhythm Patterns [28]
	Stimulative	Loudness [30] Tempo (Beats Per Minute) [30]
	Danceability	Danceability [30, 31]
Musicality	Melody	Pitch Bihistogram [29]
	Harmony	Most Frequent Chord [30] Most Frequent Chord Scale [30] Key [30] Key Scale [30]

of the music signal into the specific human loudness sensation. Then, it transforms the power spectrum into amplitude modulations on the individual critical bands. Because the notion of rhythm ends above 15 Hz on human hearing, it computes amplitude modulations for the modulation frequencies ranging from 0 to 10 Hz (i.e., 60 bins) on the individual critical bands. The algorithm thus outputs a feature vector that has 24*60 dimensions. In contrast to Rhythm Patterns, Rhythm Histogram describes a general rhythm. Rhythm Histogram sums up the magnitudes of all critical bands per modulation frequency ranging from 0 to 10 Hz (i.e., 60 bins) to form a histogram of "rhythmic energy".

For *stimulative qualities of music* [2], we use tempo and loudness. We use beats per minute (BPM) and EBU R128 loudness as descriptors [30] for tempo and loudness, respectively.

The algorithm for *danceability* is derived from [31] and implemented in the Essentia audio analysis library [30]. The core idea behind the algorithm [31] is to use Detrended Fluctuation Analysis, which has the ability to indicate long-range correlations in non-stationary time series, to measure how the presence of strong and regular beats influence the DFA exponent α. For example, music with sudden, intense jumps result in a lower level of α than music with a smoother varying series of intensity values; this means that music with pronounced, regular beats has lower α values than music with a more "floating, steady nature" [31]. The algorithm outputs values range from 0 to 3 (higher value means the song is more danceable) [30].

For *melody*, we use Pitch Bihistogram as a descriptor. "Pitch bihistogram describes how often pairs of pitch classes occur within a window d of time." In the implementation, to form a chromagram with 60 discrete bins, the algorithm wraps the pitch content to a single octave. The window length is set to $d = 0.5$ in the [0,1] range and the feature values are normalized [29].

For *harmony*, we use a key and a chord as descriptors. The key and the scale of the key are computed given a pitch class profile (HPCP). For the chord, we compute the most frequent chord of the progression and the scale of the most frequent chord of the progression. In cases where multiple chords are equally frequent, the chord is hierarchically chosen from the circle of fifths. Valid chords are C, Em, G, Bm, D, F#m, A, C#m, E, G#m, B, D#m, F#, A#m, C#, Fm, G#, Cm, D#, Gm, A#, Dm, F, Am [30]. The scales of keys and chords are either "major" or "minor."

While the Spotify API also provided some of these features such as danceability, at the time of writing, many of those APIs were in beta testing and did not provide details on *how* features were computed. Instead, we opted for open-source algorithms that provided extensive documentation, were published in peer-reviewed papers, and have become widely used in the music information retrieval community.

5 Clustering activities

Since we don't have user ratings or labels to link activities to motivational music, we first use historical preferences of songs for different activities and map these songs to the BMRI sound feature space to identify motivational sound range for the activities.

We expect songs that are labeled with the same activity to be close to each other in the feature space. We also hypothesize that, when two activities need the same type of motivational stimula, their respective songs are clustered together in the latent feature space.

Given a set of n songs, represented by their d-dimensional feature vectors $(\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$, where $d = 1508$), we use k-means clustering to partition them into k clusters $(C_1, \ldots, C_k)$ such that within-cluster sum of squares is minimized:

$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{\mathbf{s} \in C_i} \|\mathbf{s} - \mu_i\|^2,$$

where μ_i is the centroid of cluster C_i. We used the euclidean distance for K-means clustering and not a weighted distance function because we assumed all dimensions are equally important for our first study.

To identify optimal k for the clustering, we used both elbow criterion that looks at the "elbow" in a plot that shows the sum of squared errors (Fig. 2) and silhouette score (Fig. 3). As we can see from Fig. 2, one may choose either 3 or 4 as k, while from Fig. 3, one will choose 2 as k. Therefore, we choose median/mean of these possible cluster sizes as k, which leads to $k = 3$.

Each cluster C_i contains songs that might be labeled with a variety of activities. For each activity a, we aim to find its most representative cluster. To do that, for each cluster, we compute the ratio of the number of its songs labeled with a (denoted as $song^{a,c}$) over the

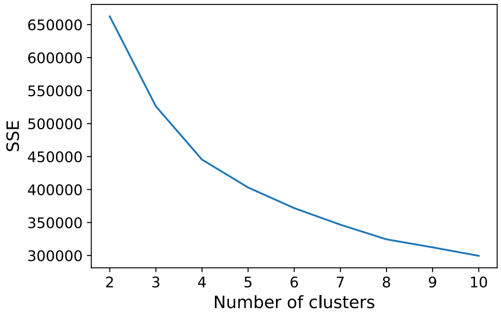


Figure 2 Sums of squared errors (SSE) for number of clusters k in k-means clustering. Based on "elbow" criterion, $k = 3$ or $k = 4$ will be a choice of k in this figure

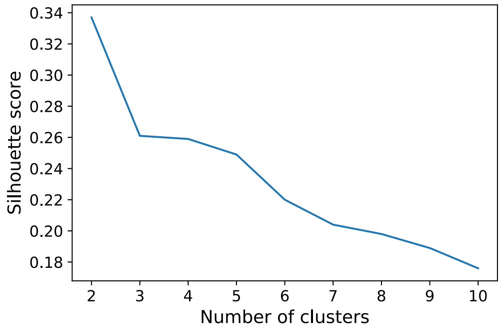


Figure 3 Silhouette score for number of clusters k in k-means clustering. Based on "elbow" criterion, $k = 2$ will be a choice of k in this figure

total number of songs labeled with a:

$$\frac{\sum_{c \in C} song^{a,c}}{\sum song^{a}}$$

and logically assign a to the cluster with the largest fraction.

5.1 Clustering results

When we looked at frequently listened songs for 14 common daily activities through the lens of motivational music qualities, they were clustered into 3 high-level groups that we call *calm* (containing 'relaxing' and 'sleeping'), *vibrant* ('commuting', 'driving', 'breakfast', 'lunch', 'dinner', and 'studying'), and *intense* ('exercising', 'running', 'office', 'partying', 'drinking', and 'sex'). The grouping is summarized in Table 2. Example songs in the calm group included compilations of nature sounds, instrumental and classical music. In the intense group, we found rock, electronic, and pop songs (e.g., Bonjovi's "It's my life"). The vibrant group included vivacious songs but less danceable compared to the intense group (e.g., The Beatles' "Hey Jude").

To get a visual cue about how the feature space differentiated songs belonging to different activities and activity groups, we ran a Principal Component Analysis (PCA) on the feature vectors of all the songs and plotted each song against the two largest PCA components (Fig. 4). In such 2-dimensional PCA space, calm songs were mostly located in the leftmost area; intense songs were on the right; and vibrant songs were in between them, partially mixed with the vibrant cluster.

To characterize the three clusters, we compared their distributions on the different BMRI dimensions. Loudness and danceability were lowest in the calm group and highest in the intense group (Fig. 5). The results align with expectations: fast-paced activities (e.g., 'exercising', 'running') are best accompanied by songs that are more danceable compared to quieter solitary or social activities that require more focus (e.g., 'studying' or 'dining') or during time for relax. Figure 5 (left) shows the tempo (beats per minute) for the three groups. Surprisingly, the BPM was the highest for the calm group ($\mu = 121.91$, $\sigma = 24.62$, $min = 61.56$, $max = 184.57$), followed by intense ($\mu = 117.19$, $\sigma = 21.72$, $min = 68.18$, $max = 184.57$) and vibrant ($\mu = 116.31$, $\sigma = 27.58$, $min = 67.21$, $max = 184.57$). We performed non-parametric ANOVA test, Kruskal-Wallis, for BPM, loudness, and danceability because our data could not assume the normality. A Kruskal Wallis test revealed a significant effect of group on BPM ($\chi^2 = 7.55$, $p = 0.02$), loudness ($\chi^2 = 176.66$, $p < 0.001$) and danceability ($\chi^2 = 87.52$, $p < 0.001$). Since the results of BPM do not align with expectations, we manually inspected outliers in the calm group whose BPM was greater than

Table 2 14 common daily activities can be clustered into three musical archetypes, namely *calm*, *vibrant*, and *intense*, according to motivational qualities of the associated songs

Calm	Vibrant	Intense
Relaxing	Commuting	Exercising
Sleeping	Driving Breakfast Lunch Dinner Studying	Running Office Partying Drinking Sex

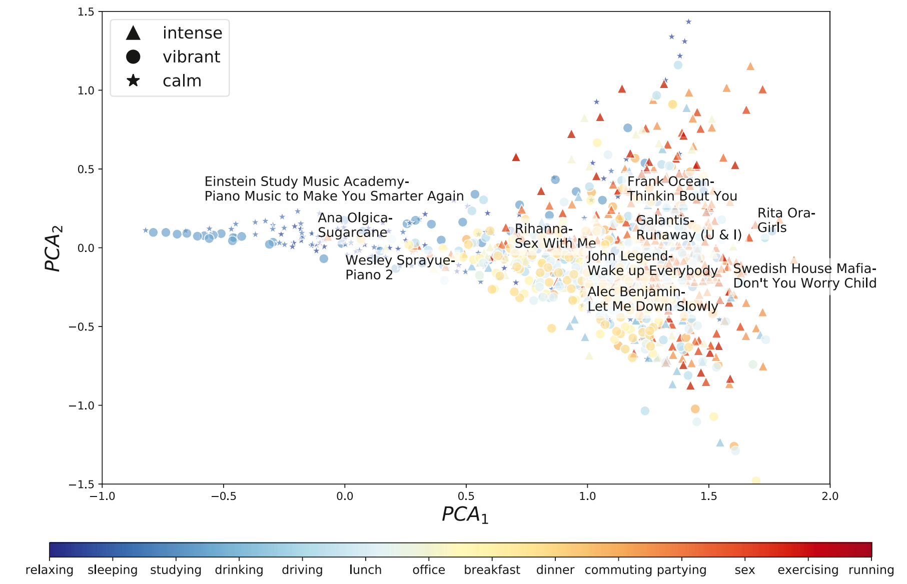


Figure 4 Projection of the songs' feature vectors on a 2-dimensional PCA space. Individual songs are drawn with markers representing the activity group, and color-coded with their activity. Example songs close to the centroid of points for each activity are reported. Calm songs, such as those for sleeping and relaxing, are mostly located in the leftmost side of the PCA space; intense songs, such as those for exercising, sex, and partying are on the right; and vibrant songs, such as those for breakfast, lunch, dinner, and driving, are in between the calm and intense songs

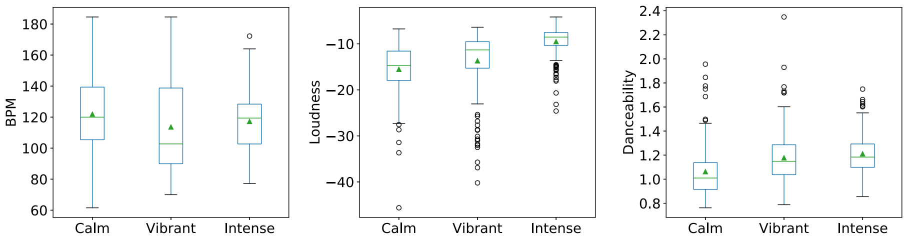


Figure 5 Tempo (left), Loudness (middle), and danceability (right) of the songs across the three music archetypes, namely, *calm*, *vibrant*, and *intense*. Both loudness and danceability are the highest for *intense* group and are the lowest for *calm* group, and the *vibrant* is in-between the two. For tempo, measured as beats per minutes, was the highest for the *calm* group, followed by the *intense* and *vibrant* groups. That is because the songs in the *calm* group are instrumental, such as classical music and music for mediation, and their beats are harder to automatically detect. Means are shown as triangles and medians as solid lines. The box whiskers indicate range including outliers

the mean plus one standard deviation. Twenty-four out of 28 songs were instrumental including classical songs and meditation songs. When the songs do not use steady metronomic time, it becomes harder to automatically detect the beats since the time is not kept by percussion. That is why the addition of features other than tempo is important.

We also looked at the most common musical keys of the songs for the three groups (Fig. 6). To best interpret the results, we linked the musical keys to feelings that they commonly provoke in people, as reported in the musical theory literature [32, 33]. The most common musical key in the *calm group* was F Major, followed by G major, and A Major. F Major is associated with calm, complaisance and repose, G Major with rustic, moder-

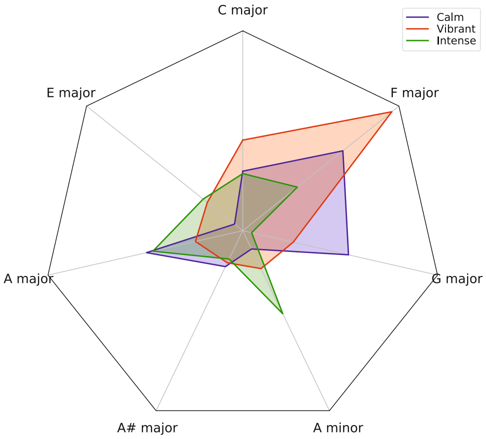


Figure 6 Most common music keys for the songs across the three music archetypes. Top-3 musical keys for the *calm* group are F Major, G Major, and A Major; top-3 musical keys for the *vibrant* group are F Major, C Major, and G Major; top-3 musical keys for the *intense* group are A Major, A Minor, and F Major. This figure is normalized per group, and the distance represents the percentage of each key in the group

ately idyllic and lyrical, and A Major with contentment over its situation, and youthful cheerfulness. Based on such characteristics, it is not surprising that people would listen to songs with these keys to seek relaxation. In the *vibrant* group, the three most common musical keys were F Major, C Major, and G Major. The presence of C Major was the most distinctive aspect compared to the calm group. C major is a cheerful key and is often described as gaiety, mirth, victorious, and innocent [33] and it also conveys joy [34]. These characteristics fit quite well for monotonous activities when people can use a stimulus to raise their spirits, or in social activities such as dining, when music can bring joy and vibrancy to the table. Last, the most common musical keys for the *intense* group were A Major, A Minor, and F Major, and that meets expectation: people need contentment and cheerfulness while working out or partying. Although these general characteristics of musical keys give us a lens through which to interpret our results, we caution readers that such descriptions tend to be too specific for a key's character, and there is a criticism over the belief in the uniqueness of character for each key, which was unanimous from late 17th till early 19th with music theorists [33]. Also, such characterization was based on classical music in early 17th–19th, which may not align perfectly with contemporary music.

6 Classification

Clustering results show that the songs that accompany the 14 most common daily activities can be grouped into only three groups when it comes to the motivational properties of their sound. To build a recommender that picks the best song for an activity, we need to learn, for any given song, to which of these three groups it belongs to. To establish that, we run a classification task that aims at classifying a song into its correct group.

This is a three-class classification task that we approach with a combination of three 'one vs. rest' binary classifiers: given a song, we calculate the confidence that it falls into cluster c_i or not, $\forall i \in [1,3]$ and we select the cluster with higher confidence. We accomplish that using a Random Forest classifier trained on: *i)* each individual feature, *ii)* all features, and *iii)* all features except Rhythm Histogram, Rhythm Patterns, and Pitch Bihistogram. We used 10-fold cross validation with 70-30 train-test split. To balance the training, in each

Table 3 Results of classification for three groups, in a binary 'one vs. rest' classification setup

	Accuracy	Precision	Recall	F1
Calm				
Rhythm Histogram (RH)	0.889	0.876	0.915	0.894
Rhythm Patterns (RP)	0.879	0.869	0.896	0.881
Stimulative Loudness	0.779	0.734	0.879	0.799
Stimulative Tempo	0.726	0.684	0.840	0.754
Danceability	0.736	0.730	0.753	0.739
Melody	0.828	0.792	0.891	0.838
Harmony	0.603	0.590	0.689	0.634
All except RH and RP	0.834	0.801	0.891	0.843
All	0.886	0.863	0.921	0.89
Baseline	0.496	0.496	0.474	0.485
Vibrant				
Rhythm Histogram (RH)	0.867	0.874	0.866	0.867
Rhythm Patterns (RP)	0.842	0.815	0.886	0.848
Stimulative Loudness	0.798	0.769	0.853	0.808
Stimulative Tempo	0.752	0.749	0.762	0.755
Danceability	0.728	0.709	0.778	0.741
Melody	0.81	0.787	0.851	0.818
Harmony	0.627	0.617	0.668	0.641
All except RH and RP	0.844	0.821	0.880	0.850
All	0.854	0.824	0.901	0.860
Baseline	0.495	0.496	0.494	0.494
Intense				
Rhythm Histogram (RH)	0.841	0.811	0.890	0.849
Rhythm Patterns (RP)	0.863	0.826	0.920	0.870
Stimulative Loudness	0.828	0.813	0.852	0.832
Stimulative Tempo	0.741	0.719	0.792	0.753
Danceability	0.719	0.676	0.845	0.751
Melody	0.817	0.791	0.864	0.826
Harmony	0.627	0.631	0.617	0.623
All except RH and RP	0.860	0.844	0.884	0.863
All	0.865	0.855	0.880	0.867
Baseline	0.501	0.503	0.505	0.503

fold, we randomly sampled the same number of positive and negative instances. The classifiers were optimized to maximize the accuracy, and we used a stratified random classifier that generated predictions by respecting the training set's class distribution as the baseline.

The classification performance is shown in Table 3. Overall, the features with the highest mean classification accuracy are two rhythm-related features: Rhythm Histogram (RH), and Rhythm Patterns (RP). Rhythm Histogram alone achieves 88.9% of accuracy for the calm group, 84.1% for the intense group, and 86.7% for the vibrant one. Rhythm Patterns alone achieves similar results between 84% and 88%. The third most predictive feature is a melody feature with Pitch Bihistogram achieving accuracy of 83% for the calm group, 81% for vibrant one, and 81.7% for intense one. Stimulative music features (i.e., loudness and tempo) was the fourth most predictive (78% to 83%).

For the misclassified songs per classifier, we also investigated the total number of songs in the testset for each classifier, the average number of misclassified songs, and the average percentage of misclassified songs per activity group (Table 4). For the calm classifier, there were 39.67% of vibrant songs were misclassified as calm, and 24.71% of intense songs were misclassified as calm. For the vibrant classifier, there were 14.11% of calm songs were misclassified as vibrant, and 51.44% intense songs were misclassified as vibrant. Lastly, for the intense classifier, 5.22% of calm songs were misclassified as intense, and 54.15% of vibrant

Table 4 This table shows the total number of songs in the test-set for each classifier, the average number of misclassified songs, and the average percentage of misclassified songs per activity group, which resulted from 10-fold cross validation for each classifier

Classifier	Correct	Incorrect	Accuracy
Calm	82.3	11.7	87.55% (82.3/94)
Vibrant	241.5	42.5	85.04% (241.5/284)
Intense	247.7	38.3	86.61% (247.7/286)

songs were misclassified as intense. This result shows that it is more common for calm and intense classifiers incorrectly classify vibrant songs, possibly due to the fact that the vibrant songs are somewhat in between the motivational ranges, and some of the songs are more or less suitable for other activity groups as well.

Rhythm Histogram, Rhythm Patterns, and Pitch Bihistogram are very informative yet quite expensive to compute, as they take about 30 seconds to a minute to extract for each song on a computer with 2.5 GHz Intel Core i7 and 16 GB memory. However, a classifier that combines all the features with the exception of Rhythm Histogram, Rhythm Patterns, and Pitch Bihistogram still yields accuracies in the 83%–86% ballpark, which is comparable to the top results. To build our recommender and to test it in the wild we used this reduced, yet efficient model. In the next section, we described our preliminary user evaluation with these reduced classifiers.

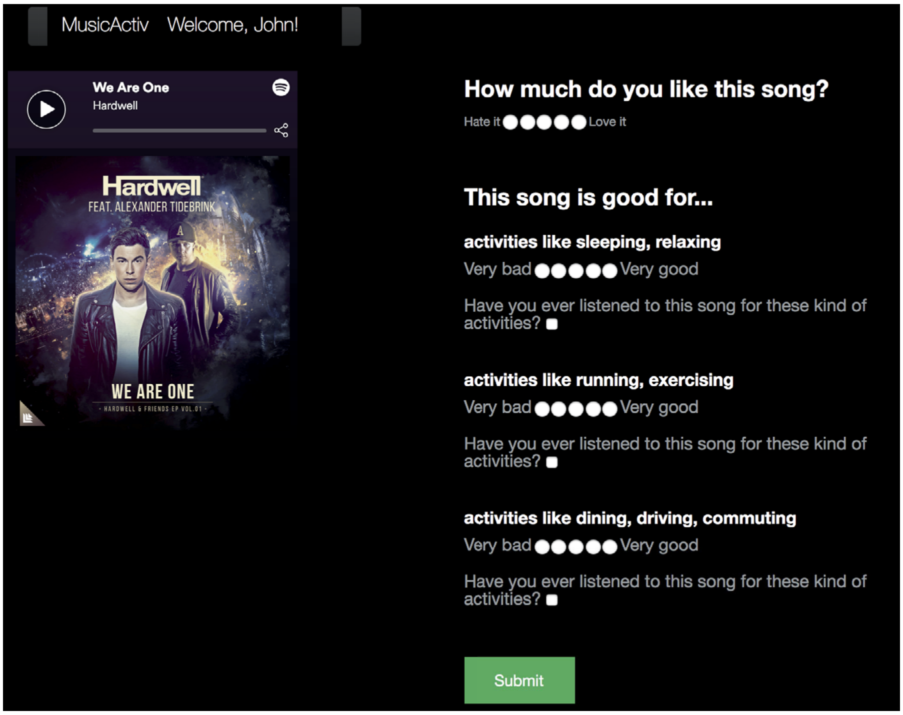
7 Preliminary user evaluation

We have shown that it is possible to accurately predict which activity type a song would be relevant to. Here we take a step further by classifying each of the user's song into 3 categories based on the songs the user has listened in the past, and by then asking the user to rate the quality of those recommendations.

7.1 Procedure and apparatus

We recruited participants through social media, mailing list, and word of mouth. Participants volunteered for their time and, upon accessing the website we set up for the experiment, they were asked to login with their Spotify account. In a pre-survey section, participants were asked to provide two pieces of information: basic demographic data (e.g., age and gender) and the frequency of listening to music while engaging in each of our three music archetypes. For the sake of clarity and exhaustiveness, we omit the cluster labels we arbitrarily picked and, instead, for each archetype, we listed the activities they include. We hypothesize that people who listen more often to music while performing a given activity can more reliably estimate the appropriateness of a song for that activity.

While the participants filled out the pre-survey, we gathered and processed their Spotify data. Specifically, we retrieved the last 50 songs played on each of the last 20 days. We chose 50 songs (which is approx. 4 hours) as a Nielsen report of 2017 showed that people spent 4.5 hours per day listening to music [35].[b] We then randomly sampled 10 songs from this set, extracted their audio features, and used our pre-trained classifier to determine the likelihood of the song belonging to each of the three music archetypes.[c] We randomly selected a total of 6 songs such that each song has high confidence score (> 0.5) for at least one of the three music archetypes.



How much do you like this song?

Hate it ●●●●● Love it

This song is good for...

activities like sleeping, relaxing

Very bad ●●●●● Very good

Have you ever listened to this song for these kind of activities? ■

activities like running, exercising

Very bad ●●●●● Very good

Have you ever listened to this song for these kind of activities? ■

activities like dining, driving, commuting

Very bad ●●●●● Very good

Have you ever listened to this song for these kind of activities? ■

Figure 7 An interface for participants to evaluate the recommended songs. A user can listen to a 30-second preview (left), rate how much they like the song (top right), rate how good the song is for three activity groups, and whether or not they listened to this song for each activity group in the past

Once the participants filled out the pre-study survey, they were directed to a sequence of six pages (one per selected song) that were identical in structure (Fig. 7). They could listen to a 30-second snippet of the song to answer a short questionnaire, which asked users to: *i)* rate how much they like the song on a scale from 1 ('hate it') to 5 ('love it'); *ii)* and separately, assess how good the song is for the activities included in the three macro-groups, from 1 ('very bad') to 5 ('very good'); and *iii)* specify if they have ever listened to the song while engaging in those activities.

7.2 Results

7.2.1 Participants

The 25 participants we recruited rated 150 songs on a 5-point scale. Among them, 18 were male (72%) and 7 were female (28%). Ages were distributed as follows: 18–24 years old (24%), 25–34 years old (60%), 35–44 years old (8%), 45–54 years old (8%).

Figure 8 shows the frequency of listening to music while engaging in the three music archetypes. For calm category (sleeping and relaxing), 40% of participants listened to music more than 3 times a week, 12% of participants listened once a week, 12% of participants listened once a month, 12% of participants listened two to three times a month, 12% of participants listened once every couple of months, 8% of participants never listened, 4% of participants listened twice a week. For intense category such as exercising and running, 28% of participants listened to music more than 3 times a week, 20% of them never listened to music, 16% of them listened once a week, 16% of them listened two to three times a month, 12% of them listened twice a week, 4% of them listened once a month, and

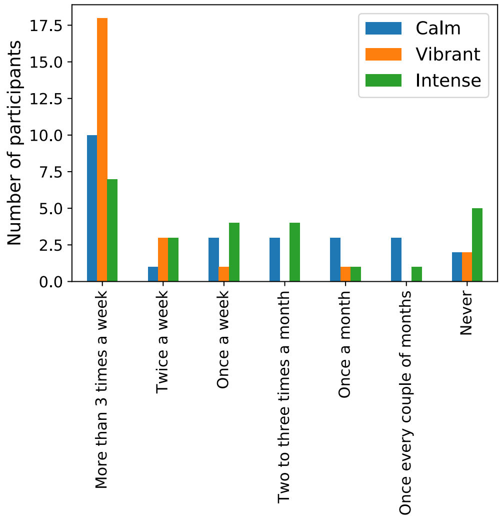


Figure 8 Participants' frequency of listening to music for the three music archetypes. Twenty five participants participated in the study and rated how often they listen to music while engaged in the three music archetypes, namely calm (Blue), vibrant (Orange), and intense (Green)

another 4% of them listened once every couple of months. For vibrant category such as dining, commuting and driving, 72% of participants listened more than 3 times a week, 12% of them listened twice a week, 8% of them never listened, 4% of them listened once a month, and another 4% of them listened once a week. Based on these results, we can conclude that activities in the vibrant category are more likely to be paired with music listening.

7.2.2 Evaluation of recommended songs

To focus on people who were more likely to have vivid recollection of listening to music while engaged in an activity, we only considered the responses of those who listened to music for a given activity at least twice a week. This resulted in 65 songs rated for the calm category, 115 songs rated for the vibrant category, 49 songs rated for the intense category.

To evaluate how the confidence score of our classifier affects user responses, we looked at the three confidence score ranges (low: [0, 0.33), mid: [0.34, 0.66), high: [0.67, 1.00]) and their corresponding user responses (Fig. 9). For the calm category, the mean for the user responses was 2.5 ($\sigma = 1.61$) when the confidence score was in the low range, 2.92 ($\sigma = 1.44$) when the confidence score was in the mid range, and 3.5 ($\sigma = 1.22$) when the confidence score was in the high range. We conducted pair-wise comparisons, and the Mann-Whitney U test showed that there's a marginally significant difference between low and mid ($U = 316, p = 0.112$), between low and high ($U = 37, p = 0.079$), and no significant difference between mid and high ($U = 91, p = 0.187$). For the intense category, the mean for the user responses was 3 ($\sigma = 1.41$) when the confidence score was in the mid range, and 3.86 ($\sigma = 1.24$) when the confidence score was in the high range. The Mann-Whitney U test showed that there's a significant difference between mid and high ($U = 129, p = 0.017$). We did not include the low range in the significance testing because there was only one datapoint from a single user who rated 2. The results show that when our motivation-based classifiers were more confident about classification for the music archetype, the users also rated that the recommendation was good for the given music archetype.

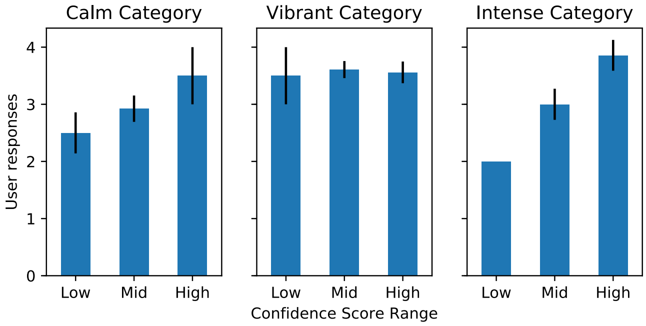


Figure 9 User ratings for songs per three confidence score ranges for each activity group. Participants rated how good the recommended songs were for a given activity group. Answers were given on a 5-point Likert scale. The confidence score range is divided into low: [0,0.33), mid: [0.34, 0.66), and high: [0.67, 1.00]. The error bars indicate the standard error of the mean. Statistical analysis detected significant differences across three confidence score ranges for the calm category and the intense category, but did not detect any significant differences for the vibrant category

However, for the vibrant category, we did not see any significant difference of confidence score range on the user responses. The average user responses was 3.5 (σ = 0.71) when the confidence score was low, 3.61 (σ = 1.32) when the confidence score was in the mid range, and 3.56 (σ = 1.11) when the confidence score was in the high range. The absence of a significant difference in the last case might be due to the nature of the activities in the vibrant group, which are suitable to a wider variety of music types compared to activities with a specific focus such as relaxing or exercising. Our classifier is trained to find the 'prototypical' songs for, say, commuting or having breakfast, but because these activities tend to have more nuanced characteristics, even songs that are not a perfect suit for those activities might be perceived as equally fit.

8 Discussion and conclusion

8.1 Implications

Music has a strong motivational potential on people. Still, this potential is only partly understood in relation to the variety of activities that people engage in daily. With this paper, we contribute to advance this understanding by translating BMRI, an inventory from music psychology that lists music features related to motivational properties, into a module that extracts those properties directly from the audio signal. From the practical perspective, this tool—which we make publicly available to the community[d]—will provide practitioners with the means of studying these properties at scale. From the theoretical standpoint, the application of the tool to annotate songs from Spotify allowed us to discover that music does not need to be activity-specific to increase motivation in that activity. Rather, there are three types of emergent music archetypes that include multiple daily activities each. However, to fully characterize music that people listen for different activities, we urge readers to take all our clustering dimensions into account when interpreting the results since the expression or perception of music is dependent on, and a combination of, musical keys, progression of chords, tempo, dynamics, and rhythmic pattern.

8.2 Limitations and future work

In this work, as our main purpose is to understand and identify motivational songs for daily activities, we only used 14 common daily activities. An interesting extension would be to collect songs for other, rare activities or events beyond daily activities to explore which type of activities can be grouped together and what might be the characteristics of newly emerged activity groups.

Our preliminary user evaluation only focused on the songs that the users have listened in the past to avoid confounds related to personal music taste. We also recognize the importance of serendipity and novelty in recommendations [36, 37]. In the future, we would like to introduce serendipitous recommendations [37] in a way that a recommended playlist has a mixture of songs that already meet user's personal music taste and other new songs that they have never listened to but could still be suitable for the user's current activity. For example, based on a user's listening history, we may first acquire music content information such as preferred genres or artists, then based on such information we may find songs that are suitable for the user's current activity and that the user has never listened before. By doing this, we can find songs that are both motivational for a given activity and that meet users' personal tastes. On top of that, we may also improve our motivation-based classifiers by including other descriptors for the musical elements (e.g., including onset patterns and scale transforms as descriptors for rhythm, and 2D Fourier transform magnitudes and intervalgram as descriptors for melody [29]).

While our preliminary user evaluation mainly focused on how good the recommended songs were given an activity group, which served as first milestone towards evaluating motivation-based classifiers, future work should evaluate such recommender systems in a more realistic setting. For example, we may set up a study where we provide a user's a playlist that is generated by a motivation-based recommender system, and ask the user to be engaged in an activity that the playlist was intended for, such as asking the user to go for a run or study for 30 minutes. Having more realistic settings like this will provide us with more accurate measures for the performance of such motivation-based recommender systems.

Another limitation of our preliminary user study is the relatively small number of datapoints, which limited our ability to run a significant correlation analysis between the user responses and our classifiers' confidence scores. Future work should conduct a large-scale user study to provide strong evidence with a correlation analysis to prove the effectiveness of our classifiers.

As existing music psychology literature has investigated the correlation between personality traits and music preferences [38], in the future, we may also explore how personality traits are linked with people's music listening habits with respect to different activities. As a first step towards this effort, we explored the correlation between the prevalence of 5 personality traits (the respondents of our user evaluation took a ten-item personality inventory [39]) and the frequency of listening to music while performing activities in the three music archetypes (Fig. 10). Respondents who were emotionally stable (low in Neuroticism) tended to listen to music while engaged in activities in the calm group (e.g., relaxing and sleeping), while those high in Extraversion tended to listen to music while engaged in activities in the intense group (e.g., partying and drinking). We speculate that emotionally stable people may tend to use music to regulate their emotions [38], while extroverts tend to engage more in social gatherings and parties. These early findings suggest that information on personality traits might be helpful at the beginning of the recommendation stage to offer tailored music recommendations, or to even nudge users into listen-

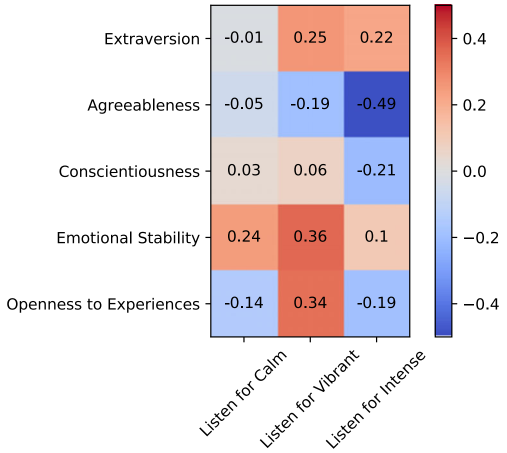


Figure 10 Correlation between our respondents' five personality traits and their frequency in listening to music while engaging in one of the three music archetypes

ing to music in situations they are not used to, making their music consumption more serendipitous.

Acknowledgements
We thank all the participants who participated in the user study.

Funding
Not applicable.

Abbreviations
BMRI, Brunel Music Rating Inventory; BPM, beats per minutes; RH, rhythm histogram; RP, rhythm pattern; HPCP, harmonic pitch class profile; DFA, Detrended Fluctuation Analysis; EBU, European Broadcast Union.

Availability of data and materials
The datasets generated and/or analysed during the current study are available on the project's site http://social-dynamics.net/pepmusic.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
YK conducted the data collection and analysis. LMA and DQ worked on the experimental design. All authors contributed to drafting the paper. All authors read and approved the final manuscript.

Author details
[1]Northwestern University, Evanston, USA. [2]Nokia Bell Labs, Cambridge, UK. [3]CUSP, King's College London, London, UK.

Endnotes

[a] https://developer.spotify.com/documentation/web-api

[b] It is possible for a participant who listened to a lot of songs for any given day, or depending on when a participant participated in the study, we may have sampled songs that they listened during specific activities. However, as our main goal is not to ensure the coverage of songs for varying activities but to evaluate the recommended songs for our three music archetypes, we believe this process will not affect our results.

[c] We limited the selection to 10 songs to ensure that the results were promptly shown to the user. The feature extraction for a larger number of songs could take a few minutes, which was too long to retain volunteers.

[d] http://social-dynamics.net/pepmusic

Received: 29 May 2019 Accepted: 24 January 2020 Published online: 24 May 2020

References

1. Park M, Thom J, Mennicken S, Cramer H, Macy M (2019) Global music streaming data reveal diurnal and seasonal patterns of affective preference. Nat Hum Behav 3(3):230
2. Karageorghis CI, Priest D-L (2012) Music in the exercise domain: a review and synthesis (part I). Int Rev Sport Exerc Psychol 5(1):44–66
3. Sloboda JA, O'Neill SA (2001) Emotions in everyday listening to music. In: Music and emotion: theory and research, pp 415–429
4. Elliott GT, Tomlinson B (2006) Personalsoundtrack: context-aware playlists that adapt to user pace. In: CHI'06 extended abstracts on human factors in computing systems. ACM, New York, pp 736–741
5. De Oliveira R, Oliver N (2008) Triplebeat: enhancing exercise performance with persuasion. In: Proceedings of the 10th international conference on human computer interaction with mobile devices and services. ACM, New York, pp 255–264
6. Baltrunas L, Kaminskas M, Ludwig B, Moling O, Ricci F, Aydin A, Lüke K-H, Schwaiger R (2011) Incarmusic: context-aware music recommendations in a car. In: International conference on electronic commerce and web technologies. Springer, Berlin, pp 89–100
7. Yakura H, Nakano T, Goto M (2018) Focusmusicrecommender: a system for recommending music to listen to while working. In: 23rd international conference on intelligent user interfaces. IUI'18. ACM, New York, pp 7–17. http://doi.acm.org/10.1145/3172944.3172981
8. Schedl M (2013) Ameliorating music recommendation: integrating music content, music context, and user context for improved music retrieval and recommendation. In: Proceedings of international conference on advances in mobile computing & multimedia. ACM, New York, p 3
9. Schedl M, Breitschopf G, Ionescu B (2014) Mobile music genius: reggae at the beach, metal on a friday night? In: Proceedings of international conference on multimedia retrieval. ACM, New York, p 507
10. Cheng Z, Shen J (2014) Just-for-me: an adaptive personalization system for location-aware social music recommendation. In: Proceedings of international conference on multimedia retrieval. ACM, New York, p 185
11. Schedl M, Hauger D, Farrahi K, Tkalčič M (2015) On the influence of user characteristics on music recommendation algorithms. In: European conference on information retrieval. Springer, Berlin, pp 339–345
12. Wang X, Rosenblum D, Wang Y (2012) Context-aware mobile music recommendation for daily activities. In: Proceedings of the 20th ACM international conference on multimedia. MM'12. ACM, New York, pp 99–108. http://doi.acm.org/10.1145/2393347.2393368
13. Yadati K, Liem CCS, Larson M, Hanjalic A (2017) On the automatic identification of music for common activities. In: Proceedings of the 2017 ACM on international conference on multimedia retrieval. ICMR'17. ACM, New York, pp 192–200. http://doi.acm.org/10.1145/3078971.3078997
14. Karageorghis CI, Terry PC, Lane AM (1999) Development and initial validation of an instrument to assess the motivational qualities of music in exercise and sport: the Brunel music rating inventory. J Sports Sci 17(9):713–724
15. Kahneman D, Krueger AB, Schkade DA, Schwarz N, Stone AA (2004) A survey method for characterizing daily life experience: the day reconstruction method. Science 306(5702):1776–1780
16. Quercia D, Aiello LM, Schifanella R (2018) Diversity of indoor activities and economic development of neighborhoods. PLoS ONE 13(6):0198441
17. Juslin PN, Liljeström S, Västfjäll D, Barradas G, Silva A (2008) An experience sampling study of emotional reactions to music: listener, music, and situation. Emotion 8(5):668
18. Scherer KR, Zentner MR (2001) Emotional effects of music: production rules. In: Music and emotion: theory and research, vol 361, p 392
19. Rentfrow PJ, Goldberg LR, Levitin DJ (2011) The structure of musical preferences: a five-factor model. J Pers Soc Psychol 100(6):1139
20. Atkinson G, Wilson D, Eubank M (2004) Effects of music on work-rate distribution during a cycling time trial. Int J Sports Med 25(08):611–615
21. Szabo A, Small A, Leigh M (1999) The effects of slow-and fast-rhythm classical music on progressive cycling to voluntary physical exhaustion. J Sports Med Phys Fit 39(3):220
22. Lehtiniemi A (2008) Evaluating supermusic: streaming context-aware mobile music service. In: Proceedings of the 2008 international conference on advances in computer entertainment technology. ACM, New York, pp 314–321
23. Kaminskas M, Ricci F (2011) Location-adapted music recommendation using tags. In: Proceedings of the 19th international conference on user modeling, adaption, and personalization. UMAP'11. Springer, Berlin, pp 183–194. http://dl.acm.org/citation.cfm?id=2021855.2021872
24. Kaminskas M, Ricci F, Schedl M (2013) Location-aware music recommendation using auto-tagging and hybrid matching. In: Proceedings of the 7th ACM conference on recommender systems. RecSys'13. ACM, New York, pp 17–24. http://doi.acm.org/10.1145/2507157.2507180
25. Cunningham S, Caulder S, Grout V (2008) Saturday night or fever? Context-aware music playlists. Proc Audio Mostly
26. Cai R, Zhang C, Wang C, Zhang L, Ma W-Y (2007) Musicsense: contextual music recommendation using emotional allocation modeling. In: Proceedings of the 15th ACM international conference on multimedia. ACM, New York, pp 553–556
27. Cheng Z, Shen J, Zhu L, Kankanhalli MS, Nie L (2017) Exploiting music play sequence for music recommendation. In: IJCAI, pp 3654–3660
28. Lidy T, Rauber A (2005) Evaluation of feature extractors and psycho-acoustic transformations for music genre classification. In: ISMIR, pp 34–41
29. Panteli M, Dixon S et al (2016) On the evaluation of rhythmic and melodic descriptors for music similarity. In: ISMIR
30. Bogdanov D, Wack N, Gómez Gutiérrez E, Gulati S, Herrera Boyer P, Mayor O, Roma Trepat G, Salamon J, Zapata González JR, Serra X (2013) Essentia: an audio analysis library for music information retrieval. In: Britto A, Gouyon F, Dixon S (eds) 14th Conference of the International Society for Music Information Retrieval (ISMIR); 2013 Nov 4-8; Curitiba, Brazil. [place Unknown]: ISMIR; 2013, pp 493–498. International Society for Music Information Retrieval (ISMIR)
31. Streich S, Herrera P (2005) Detrended fluctuation analysis of music signals: danceability estimation and further semantic characterization. In: Proceedings of the AES 118th convention

32. Schubart CFD, DuBois T (2004) On the human voice and the characteristics of the musical keys. N Engl Rev (1990) 25(1–2):166–171
33. Young JO (1991) Key, temperament and musical expression. J Aesthet Art Crit 49(3):235–242
34. Mattheson J. Der Vollkommene Capellmeister. C. Herold (1739)
35. Nielsen. Time with tunes: how technology is driving music consumption. https://www.nielsen.com/us/en/insights/article/2017/time-with-tunes-how-technology-is-driving-music-consumption Accessed 2017-11
36. Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. In: Proceedings of the 2000 ACM conference on computer supported cooperative work. ACM, New York, pp 241–250
37. Zhang YC, Séaghdha DÓ, Quercia D, Jambor T (2012) Auralist: introducing serendipity into music recommendation. In: Proceedings of the fifth ACM international conference on web search and data mining. ACM, New York, pp 13–22
38. Rentfrow PJ, Gosling SD (2003) The do re mi's of everyday life: the structure and personality correlates of music preferences. J Pers Soc Psychol 84(6):1236
39. Gosling SD, Rentfrow PJ, Swann WB Jr (2003) A very brief measure of the big-five personality domains. J Res Pers 37(6):504–528