sempre:
Society for Education, Music
and Psychology Research

MUSIC &
SCIENCE

Research Article

Music & Science
Volume 2: 1–20
© The Author(s) 2019
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2059204319893179
journals.sagepub.com/home/mns
SAGE

Encouraging Attention and Exploration in a Hybrid Recommender System for Libraries of Unfamiliar Music

John R. Taylor and Roger T. Dean

Abstract

There are few studies of user interaction with music libraries comprising solely of unfamiliar music, despite such music being represented in national music information centre collections. We aim to develop a system that encourages exploration of such a library. This study investigates the influence of 69 users' pre-existing musical genre and feature preferences on their ongoing continuous real-time psychological affect responses during listening and the acoustic features of the music on their liking and familiarity ratings for unfamiliar art music (the collection of the Australian Music Centre) during a sequential hybrid recommender-guided interaction. We successfully mitigated the unfavorable starting conditions (no prior item ratings or participants' item choices) by using each participant's pre-listening music preferences, translated into acoustic features and linked to item view count from the Australian Music Centre database, to choose their seed item. We found that first item liking/familiarity ratings were on average higher than the subsequent 15 items and comparable with the maximal values at the end of listeners' sequential responses, showing acoustic features to be useful predictors of responses. We required users to give a continuous response indication of their perception of the affect expressed as they listened to 30-second excerpts of music, with our system successfully providing either a "similar" or "dissimilar" next item, according to—and confirming—the utility of the items' acoustic features, but chosen from the affective responses of the preceding item. We also developed predictive statistical time series analysis models of liking and familiarity, using music preferences and preceding ratings. Our analyses suggest our users were at the starting low end of the commonly observed inverted-U relationship between exposure and both liking and perceived familiarity, which were closely related. Overall, our hybrid recommender worked well under extreme conditions, with 53 unique items from 100 chosen as "seed" items, suggesting future enhancement of our approach can productively encourage exploration of libraries of unfamiliar music.

Keywords

Australian art music, model, music recommender systems, perception of affect

Submission date: 30 May 2019; Acceptance date: 14 November 2019

Introduction

Art music, such as historic or post-serial Western Classical composition and improvisation, is a minority interest. For example, Schedl et al. (2018) find that in a diverse (although mainly Australian sample), the median listening time per week to classical music is 1 hour, compared with 8 hours for other genres: as they summarise, "participants either love classical music and devote a lot of time to it, or do not listen to it at all" (p. 6). Similarly, we find here (Appendix S1 in Supplemental Materials) that both Classical-Historic and Classical-Contemporary have median familiarity ratings of only 3, on a 1–7 scale (where 1 = "not familiar" or "not likeable" and 7 = "familiar" or "likeable"). Yet, past history shows that music previously perceived as inaccessible (such as that of Xenakis, or Stravinsky's *The Rite of Spring*), often becomes the canonic

MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Penrith, New South Wales, Australia

Corresponding author:
John R. Taylor, MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Locked Bag 1797, Penrith, New South Wales 2751, Australia.
Email: j.taylor@westernsydney.edu.au

music of the future. Additionally, because artists usually strive to develop their own voices in expressing their view and responses to contemporary life, it is necessary for national artistic communities to promote their own work, and to encourage community access. Consequently, the International Association of Music Information Centres exists to represent numerous national collections of such music, as held in various countries' music information centers.

Another reason for attempting to elicit interest in such collections is the general observation that independent music of unfamiliar cultures can be pro-social. This can apply also to unfamiliar artistic allure within a given community. The music of these collections usually contains music that is deeply unfamiliar to most potential users. In a similar way, one can anticipate that specialised libraries of jazz, free improvisation, electroacoustic composition (Bailes & Dean, 2012), noise music or music of a particular culture, such as that of Iran, would be experienced as totally unfamiliar by the average new Western listener visiting the collection. Thus, there is a need to develop conditions within these music information centers and other collections that encourage exploration of their music—a main aim of our work here.

In terms of conventional recommender system approaches (for general reviews, see Aggarwal, 2016; Ricci, Rokah, & Shapira, 2015; Schedl, Knees, McFee, Bogdanov, & Kaminskas, 2015), because our corpus contains predominantly new items and genres, and because there are very few users of the database, we cannot rely on any familiarity with (or ratings of) the music, because there are very few item ratings or user history to make use of. Therefore, we must predict the relationship between users' stated preferences before exposure and their responses upon listening. We used stated user preferences amongst a small group of labelled genres, and a small group of musical features, to obtain an estimate of the diversity of each individual's musical taste and the likely acoustic features that might make a seed item (the first music to be auditioned) reasonably appealing. We then encouraged musical attention by requiring participants (in a lab-based setting) to register their continuous perception of affect (arousal and valence) expressed in the music over a 30-second sample and to indicate after each item the liking and familiarity they felt towards it, and also their choice as to whether the next item should be "similar" or "dissimilar" (note that the participants' impression of familiarity essentially relates to stylistic familiarity, since there is little likelihood that they would have heard these pieces before; see also the Content-Based Similarity Measures of Music section). For each individual, our system translated their final dominant affective response into an acoustic preference, and then selected a "similar" next item, if requested, on the basis of that preference, and conversely for a "dissimilar" request. Participants were not presented with any specific choice of item among potential next items, rather the single item was provided automatically on the basis of similar or dissimilar.

Our exemplar music library is the not-for-profit Australian Music Centre Ltd (AMC), which aims to promote and support domestic composers and their music (Australian Music Centre Ltd, 2019), and makes use of FRBR (Functional Requirements for Bibliographic Records) metadata to add value to its community of represented musical artists. AMC's financial resources are limited, and the AMC online music database comprises over 13,000 digital music files containing varying amounts of solely high-level metadata (mostly using descriptive categories based on instrumentation and period), thus limiting the possibilities to engage users. Despite this, the database has linkage, topic, and historical information, which allows the exploration of styles, composers, influences, periods, and ethnicities.

Consequently, in comparison with Spotify or iTunes, the AMC and related specialist libraries face four main challenges: firstly, almost all of the musical items available on the AMC database are unfamiliar (thus the music, even if presented on Spotify or iTunes, is in the "long tail" of usage (Celma, 2010), such as "sound art"); secondly, the AMC's diverse database contains substantial proportions of genres and musical forms that are even more unfamiliar than historic Western Classical music (such as improvised music and electroacoustic music); thirdly, the AMC database descriptors are limited; and finally, the AMC needs to recommend diverse music, even during the exploration of music that is already unfamiliar to listeners. This ideally requires an extended duration of engagement, even under the harshest of "cold start" conditions, where there is normally no historical usage data for either items or users. We attempt to combat the first to third challenges, above, and to accommodate the fourth.

Aim, Design and Hypothesis

For the purposes of music recommendation, we sought to predict a user's liking and familiarity responses to unfamiliar music from their prior preferences for genre and musical feature and their ongoing continuous affect assessment of each auditioned item. We assess all these data, together with acoustic features of the items as potential predictors in statistical models of user responses, specifically proposing that our system will have potential utility if:

H1: The use of information on participants' prelistening genre preferences will mitigate the cold-start problem and achieve seed item ratings comparable to later ratings, rather than dramatically worse.

H2: Familiarity (and to a lesser extent liking) will increase during a listening session as a person's item and style exposure increases. We expected to trace the early part of the normal inverted-U dependence of these parameters upon exposure, even though the items were likely to be deeply unfamiliar and potentially quite challenging for many listeners. Commonly, some increase in familiarity is required before there are increases in liking. Chmiel and Schubert (2018) have recently reviewed the psychology of exposure and familiarity in the context of music recommender systems.

H3: When a "similar" item is requested, liking and familiarity of the present and the subsequently provided item will be higher than when a "dissimilar" item is chosen (because of the mechanism by which we select items).

H4: Liking and familiarity responses to the items will be highly correlated and show positive mutual influences in statistical models of response sequences.

H5: Participants' expressed prior preferences for musical features (such as "bass") will predict their individualised liking and familiarity responses to the items (by virtue of our interpreting them during recommendation in terms of the items' acoustic features).

The use of acoustic features will be intrinsic to any success our model displays in relation to H1 to H5. Nevertheless, we also predict:

H6: Statistical models of sequential liking and familiarity will show additional roles of acoustic features as predictors.

To assess the efficacy of our approach, our experimental scenario required sequential responses, and thus it is clearly inappropriate to treat all responses as being independent and identically distributed, as is commonly done. Rather, each individual's responses have a potential time series dependency, which we consider in some models, using cross-sectional time-series analysis to maintain every series as a distinct data set, allowing assessment of both fixed and random effects in mixed effects models.

This paper is organised as follows: in the Related Work section we present a brief overview of recommender systems and of methods for obtaining content-based similarity measures of music. The Participants, Methods, Materials, and Procedures section describes our experimental approach. In the Results, we present the results of our experimentation and the associated analytical models, and finally, in the Discussion section, we draw conclusions and discuss potential future work.

Related Work

Recommender Systems

There are five predominant recommender system approaches: (1) collaborative filtering (CF), (2) content-based (CB), (3) utility-based, (4) knowledge-based, and (5) hybrid (Burke, 2002; Jannach, Zanker, Felfernig, & Friedrich, 2010). In CF techniques, recommendations are based upon aggregated user-purchase history and similarities between users' ratings or recommendations. Widely used, CF techniques can suffer from the cold-start problem (when there are sparse ratings), and from the "grey sheep" problem (e.g., user profiles that deviate from existing user classifications; Burke, 2002). Content-based recommendation systems (type 2) use the similarity between items which, for music, exploits measures of acoustic content (e.g., MPEG-7 descriptors), often combined with semantic labels such as those Spotify attempts to provide (e.g., danceability) or high-level tags (such as the words "happy" or "sad").

CB systems often omit user ratings data. Acoustic measures are used widely in music information retrieval (MIR) (Knees & Schedl, 2016; Lartillot, Toiviainen, & Eerola, 2008) as well as recommender systems (Bogdanov, Serra, Wack, Herrera, & Serra, 2010). Utility-based recommender systems (UBRS: type 3) and knowledge-based recommender systems (KBRS: type 4) evaluate whether the specification of a product satisfies the user's requirements (Burke, 2002; Huang, 2011). KBRS focus on satisfying customer requirements from item descriptions, whereas UBRS focus on the utility of the product to the user (Aggarwal, 2016). Neither suffer from the cold-start problem because they do not need historical usage data, although to infer relevance and similarity they need item and user requirement information.

Hybrid recommender systems (HRS: type 5) employ combinations of the systems described above. They perform better than the individual methods described above (Burke, 2002), making them a popular technique. The success of a hybrid approach is dependent on application, the items, the users, and the system's existing knowledge: ultimately, on the datasets used. HRS have successfully been augmented with large datasets of music preference and consumption patterns (such as the LFM-1b dataset; Schedl, 2017).

Currently, these recommender systems are often supplemented by "context-awareness", in which information about time, environment, user activity, and perhaps character is determined and used. The only aspect of context which could have been used in our study is that of a user's personality (beyond taste for musical feature or genre). Schedl et al. (2018), for example, use the standardised Ten Item Personality Instrument, and demonstrate some modest correlations (largest absolute magnitude 0.222) between these features and propensities for post-listening retrospective ratings among 11 emotion descriptors. Our intent was to use continuous affect responses (rather than discrete retrospective ratings), on the basis of ecological relevance and to focus attention during listening. It has been found that with both unfamiliar and familiar music presented in this way to non-musicians, trained musicians, and specialist electroacoustic musicians, inter-personal differences in responses are far greater than inter-group differences, and that inter-personal differences are just as pronounced in each expertise group (Dean, Bailes, & Dunsmuir, 2014a, 2014b). Therefore, we chose not to include a personality instrument in our study.

In our experimental situation, of a library entirely comprised of unfamiliar music to which the participants have not been exposed, and on which there is little prior usage information, a content-based approach is essentially the only applicable one from types 1 to 4, above. We hybridised this with pre-listening user preference data, to create a type 5 system. We also used the extremely limited current AMC data on item usage (simply the sum of view counts by item). As noted already, the library does provide some facilities for utility or knowledge-based interrogation, for example via individually specified composers (e.g., Peter Sculthorpe), or via topics such as indigenous music or environmental music. We do not pursue these here.

The recommender system types described above have limited application to the AMC online database: the AMC has limited users, and ratings data comes from web page hits and item purchases. The latter are unrepresentative of typical consumption patterns, because many items are mandatory in the Australian Music Examination Board (AMEB) syllabus (AMEB, 2019), and thus purchased for educational rather than general consumption purposes. Consequently, a CF technique solely using these data would recommend AMEB items rather than new unfamiliar music. Our approach to personalising the recommendation attempts to use a basic CF technique, by linking item view count from AMC data with the diversity/homogeneity of the user's musical taste in order to recommend relatively appealing seed items, given that our items are predominantly unfamiliar music. Thereafter, CF is not used in our prototype system.

In music recommender systems, CB approaches often fail because acoustic similarity measures are not universally comparable between songs/genres. We expect similar difficulties with the diverse AMC library, especially given the types of users (e.g., the content and context; Knees & Schedl, 2016). Nevertheless, we use auditory content information in order to repeatedly choose "similar" or "dissimilar" items. We aim with our content-based approach to acoustic similarity, to facilitate a noticeable improvement of liking and familiarity of the requested and auditioned "similar" items, compared to chosen "dissimilar" items (H3), although this assumes that a choice of "similar" by a participant indicates that they liked the present item more than average, which we can assess from our data.

Although the study of users and their reactions is beginning to attract attention, few suggestions specific to libraries of uniformly unfamiliar music can be gleaned from the literature (e.g., see the review by Weigl & Guastavino, 2011). Given that we aim to encourage universal exploration of an unfamiliar library (i.e., where most people are non-musicians), we assessed participants' prior preferences for genres and musical features (such as preferences for "bass" or "melody"') and used these to personalise the seed item. Using a musical sophistication scale would have been alienating for most participants and was not adopted. Thus, our system interprets information about prior preferences in terms of acoustic content to drive the seed recommendations. It also interprets users' continuous affective responses to a piece in acoustic terms in order to make the subsequent recommendation.

Music Genre and Feature Classification

Several of the recommender system types described above attempt to use self-identified musical preferences expressed by participants, alongside their demographic information. Genre taxonomies derived from the semantic web, such as those of DBpedia, offer numerous musical categories in hierarchies (Schreiber, 2016), while others offer rather few parent/root genre similarities (Sturm, 2013b; Tzanetakis & Cook, 2002). These inconsistencies lead to misclassification and confusion (Sturm, 2013a) and poor content-based recommendation (Bogdanov, Porter, Urbano, & Schreiber, 2017; Sturm, 2013b). An additional problem is that the taxonomy employed by the AMC is often vague and not explicitly focused on genre (e.g., orchestral music, which can appear as "instrumental" and "orchestral" and does not circumscribe a genre), and many of the diverse music genres in the AMC corpus, such as electroacoustic, art, choral, chamber, and jazz music among others are unfamiliar.

Thus, rather than employing an item taxonomy-based approach, we estimated each user's general music diversity. This was done by asking them to rate, on a Likert scale of 1-7, their enjoyability of and familiarity with the following genres: Acoustic, Blues, Classical–Contemporary, Classical–Historic, Country, Electronic, Experimental, Jazz/Improvisation, Pop, Rock, Urban, and World. These music genres were adapted from a taxonomy previously used in a large study of Australian cultural tastes in relation to socioeconomic groupings (Bennett, Emmerson, & Frow, 1999).

Content-Based Similarity Measures of Music

MPEG-7, the international standard for audio content description under ISO/IEC 15938:2002 (International Organization for Standardization (ISO), 2002) contains seventeen hierarchic spectral and temporal descriptors of music acoustics and instrumental timbres based on perceptual knowledge: such as acoustic intensity, spectral flatness and centroid, log attack time, and brightness (Casey, 2001; Dean & Bailes, 2011). This has led to many applications in MIR (Kim, Moreau, & Sikora, 2006) including audio analysis techniques and machine listening (Jehan, 2005); audio content matching and comparison (Allamanche et al., 2001); automatic classification (Tzanetakis & Cook, 2002); and music recommendation systems (Aggarwal, 2016; Celma, 2010). One predominant challenge in MIR and in psychoacoustics is adequately associating the perceived timbral aspects with the acoustic features of audio signals because of timbre's multidimensional nature. For instrumental classification, some acoustic features are more suitable, the extent of which can vary between genres and within songs (Tzanetakis & Cook, 2002).

Even with short sounds, substantial inter-participant differences of dissimilarity ratings depend on the relative salience of timbral features (Caclin, McAdams, Smith, & Winsberg, 2005). For example, the detection of musical transitions is related to the conspicuousness of the phrase (Bailes & Dean, 2007b), the segment length (Bailes & Dean, 2007a) and the speed of transition (Bailes & Dean, 2009). More recently, Olsen, Dean, and Leung (2016) showed substantial differences in how acoustic features predicted perceptions of segmentation in sound-based music extracts (that is, music primarily focused on continually varying timbres, such as noise, rather than instrumental note-based events; Landy, 2009) and in note-based music extracts (e.g., canonical classical, popular instrumental, vocal music).

Nevertheless, listeners' may be attracted to similar musical acoustic features irrespective of genre (Rentfrow, Goldberg, & Levitin, 2011), hence our hypotheses suggesting a predictive influence of acoustic features on liking and familiarity even across different genres. Participants' prior preferences for musical features may encompass those acoustic features (Hypothesis 5); furthermore, in using acoustic similarities in our similar/dissimilar recommendation step, we may transfer the predictive impact of this parameter somewhat onto the liking/familiarity and affect parameters themselves. Our core measure of acoustic feature similarity is the Mahalanobis (M) distance between each item and the mean acoustic feature set of the whole current corpus of extracts. *M* distance is a multivariate measure of distance between a single observation and a set of observations. For example, for a data matrix of musical items X (n × p), containing *n* items indexed by *i* with *p* acoustic measures (such that $x_{i,p}$ is the *p*th acoustic measure of the *i*th musical item), we can calculate the Mahalanobis distance $dist_M$ between the *i*th row vector x_i of X and the mean row vector $\bar{x}$ where C_x is the variance-covariance matrix, and ^{T} is the transposed vector as:

$$dist_M(x_i,\bar{x}) = \sqrt{(x_i - \bar{x})C_x^{-1}(x_i - \bar{x})^T} \quad \text{for } i = 1 \text{ to } n, \qquad (1)$$

(De Maesschalck, Jouan-Rimbaud, & Massart, 2000; see also Mahalanobis (1936) for a more detailed explanation). This approach takes account of the individual variabilities of all the acoustic dimensions, hence is suitable for our dataset of multiple acoustic features. *M* values range from 0 (identity) to an unbounded positive upper (extreme dissimilarity) (Komkhao, Lu, Li, & Halang, 2013).

Information compressibility can significantly impact pattern recognition, similarity measures, liking and familiarity (Hudson, 2011; Schmidhuber, 2009). Extreme musical pattern complexity is perceived as uninteresting, as compressibility is either impossible or trivial (Hudson, 2011). Schmidhuber (2009) posits that the brain compresses auditory information more efficiently for familiar stimuli (e.g., has perceived similarity to a prior listening experience), because of prior history in compressing similar information, although other psychological mechanisms might explain such an effect. Since this study's corpus is limited to domestic art music, we expect diverse patterns of complexity, and significant unfamiliarity. Hence our H2 suggests that liking and familiarity will be higher when a "similar" item is requested and proffered than a "dissimilar" item. Predicting a recommendation's success based upon acoustic similarity is difficult, particularly when songs and genres are unfamiliar. Thus, we also incorporate participants' continuous measures of perception of affect for recommendation, and in the longer term for understanding their acoustic preferences more comprehensively than they can self-describe. Note that we provide no guidance to participants as to the interpretation of 'familiarity': since no item is heard by an individual more than once, there can be no real measure of familiarity with an individual item, but participants may feel increasingly familiar with styles that recur in the dataset (e.g., minimalism), which is then reflected in a rising familiarity rating.

Perception of Affect and its Use for Recommendations

In the long run, we aim to interpret users' real-time continuous affect responses towards in depth prediction of their preferences and hence towards recommendation. With data on a large enough body of users and given that the continuous responses (sampled at 2 Hz) provide far more data per item than the simple post-audition ratings, this should allow a more powerful system even with data from a relatively small number of users. The continuous affective response reflects the variable contextual influences upon the perception of the acoustic features. As a first step towards this long term aim, here we use the two-dimensional circumplex model of affect (Russell, 1980) because of its suitability and common prior usage as a continuous self-report method (Schubert, 2010), particularly continuous ratings of perceived affect (Bailes & Dean, 2012; Olsen, Dean, & Stevens, 2014; Schubert, 2004). Work on continuous responses demonstrates that acoustic intensity is a significant modeling predictor that is also causal of listeners' perception of arousal (Dean, Bailes, & Schubert, 2011), and that acoustic features such as spectral flatness (Weiner entropy) modulate perception of structural change, arousal and valence (Bailes & Dean, 2012; Olsen et al., 2014).

Such continuous behavioural response measures do not seem to have been used in conjunction with recommender systems, though continuously measured facial expressions have been used to provide discrete measures then applied predictively through random forest and gradient boosting training (Tkalčič et al., 2019). Thus, here we employ continuous ratings of arousal and valence in a limited way to provide discrete measures to drive our RS.

Instrumental performance factors, such as articulation (e.g., staccato and legato) can be associated with contrasting perceptual effects (in this case, gaiety and solemnity) (Gabrielsson, 2016). Again, the perceptual relevance of acoustic features depends on context, in part due to the multidimensional nature of timbre. Thus, in previous work modeling continuous perception of musical phrases (segments) based upon acoustic features, dominant influences of the last 5 seconds of sound on overall phrase perception have been observed, as judged by time-dependent predictions using contemporary versions of Cox survival analysis (Olsen, Dean, & Leung, 2016). Correspondingly, in the present study the next item recommendations are partly based upon the terminal portion of an individual listener's continuous ratings. We used these to assess the likely dominant spectral features in the individual's perception, to recommend a "similar" next item with analogous feature and magnitude. Conversely, the "dissimilar" item was identified by evaluating whether the Mahalanobis (M) distance of the "similar" item was above or below the mean *M* for the current corpus and by selecting the item with the most antagonistic *M* value from the available corpus.

Participants, Materials, Methods, and Procedure

Participants

This experiment was approved by our University's Human Ethics Committee and participants provided informed written consent (approval number: H12015). Sixty-nine non-musicians were recruited via our University's online participation system SONA. First year students received course credit in return for participation, and participants conducted the test properly. Initially, participants completed a questionnaire (see Appendix S2) to obtain demographic and socioeconomic data together with music genre and feature preferences (these demographic, socioeconomic music genre and feature data are shown in Appendix S1, see online Supplemental Materials). The main demographic and socioeconomic data are not analysed in this study but were collected as they may be of use in further work.

The group was made up of 69.56% female, 30.43% male. The percentage of participants in each age group was (years): 17–21 (63.76%); 22–34 (27.53%); 35–44 (4.34%); 45–54 (4.34%), >54 (0). Ethnicity[1] percentages were Australian (62.31%); Arab (8.69%); South-East Asian–Vietnamese (4.34%); and South-Asian–Indian (4.34%)—the remaining 20.3% of respondents identified as either Aboriginal Australians (1.45%), Torres Strait Islander persons, New Zealand Peoples, and Other North African/Middle Eastern (all at 2.89%), or rest of the world (10.14% combined). Participants were prompted to select one option from the ethnic categories list (see Appendix S2), and the term "ethnicity" was not described to participants. Thus, although 90% of participants were aged between 17 and 34 years, they were otherwise diverse. The second part of the questionnaire concerned participants' musical tastes, asking them to rate their experience of enjoyability of (Q7) and familiarity with (Q8) different genres of music; and how important different features of music are to them (Q9) (all were Likert scales of 1, not very enjoyable/familiar/important, to 7, very enjoyable/familiar/important; midpoint 4). The questionnaire used the term enjoyability to avoid ambiguity with the contemporary usage of the word "like" in social media. Here we use the terms "liking" and "enjoyability" interchangeably.

Table 1 shows the musical features whose personal importance was evaluated by participants, and how we translated these features into acoustic parameters in our recommender system. Neither the genres nor musical feature terms were explained to participants. When more than one feature scored the same maximal value, the first in the list was used. This avoided adding further emphasis to loudness, which we used separately after the choice of the seed item in any case (see below). We considered the possible alternative approach to eliciting user pre-listening preferences proposed by Bogdanov et al. (2013), in which users present a small group or liked items which are then interpreted for semantic audio content cues to subsequent recommendations: we viewed it as highly unlikely, given the totally unfamiliar music collection, that this approach would be very helpful, and it was consequently not assessed.

Materials and Design

Stimuli. We randomly selected recordings from the AMC collection, so as to reflect the collection's distributions across instrumentation and year.[2] Extracts were 30 seconds in duration.

Acoustic Analysis. We performed acoustic analysis on our corpus. Previous work has found that some acoustic features contribute more toward continuous perceptions of arousal and valence than others (Bailes & Dean, 2012; Dean et al., 2011, Dean & Bailes, 2010; Olsen et al., 2014). Here we use the acoustic features to drive the recommendation system successfully, and also assess whether such features can predict liking and familiarity time-series (H6).

Seven acoustic features and two measures which aim to model perceptual parameters on the basis of acoustic information (roughness, and rhythmic density) were analysed. For simplicity we refer to this whole set of measures below as 'acoustic measures'. MaxMSP software (Cycling '74) was chosen for analysis because in the future we intend to run these analyses in real-time when a new (previously unused) item is introduced to a listening session. Our acoustic analysis used a combination of the Zsa.descriptors library for MaxMSP (Malt & Jourdan, 2008), CNMAT externals (University of California, Berkeley; Puckette, Apel, & Zicarelli, 1998), and Alex Harker's [descriptorsrt ~] object (Harker, 2017), to obtain window-by-window (sampling rate 2 Hz) measures of the following spectral features: spectral centroid, spectral flatness (Wiener entropy), spectral flux, inharmonicity, log kurtosis, log skewness, roughness, and rhythmic density. For rhythmic density, we used the MaxMSP [fzero ~] object, which detects new notes if either the peak amplitude or pitch changes more than a specified amount, to simulate the rhythmic density described in Olsen et al. (2016). We adopted this approach in light of our diverse corpus of Western classical music and sound art, for example: music with a higher number of onsets per 500 ms window (onsetRate) and with a higher current maximum number of onsets per 500 ms (maxOnsets) is suggestive of complex musical phrases associated with multi-instrument or vocal music, rather than sound art where phrase segmentation based on timbre rather than onsets; and music with less difference in running mean average of onsets per 500 ms (runningMeanOnsetRate) suggests onset pattern stability (although this should also take into account onsetRate and maxOnsets, as a zero value could also apply to both music with consistent onsets and no onsets).

We chose these spectral features based on their previous utility (compared to Mel Frequency Cepstral Coefficients (MFCC)) in studies of both sound- and note-based music (McAdams, 1999; Olsen et al., 2016). We calculated the absolute differences frame by frame for the acoustic features (bar spectral flux, which is already a measure of change between frames) of all items. Then we derived our item-level feature vectors as the absolute mean difference (absmeandiff) between successive samples of the resultant 2 Hz time series.[3] A detailed description of the acoustic analysis can be found in Appendix S3 (online Supplemental Materials).

The Prototype Recommender System Design

Our prototype recommender system comprises of two parts: firstly, each individual's two "diversity indices" (detailed in the next section), based on questionnaire data, to address the "cold-start" problem and provide a personalised seed recommendation; secondly, ongoing item recommendations, based on prior continuous affect responses and the (assumed) related acoustic features. This section briefly describes these aspects of our recommender system (see Appendix S4 for a detailed process-flow), although a comparison with other approaches is outside the scope of this study.

Table 1. Music descriptors for Q9 and their inferred acoustic parameter.

Musical Feature	Anticipated acoustic parameter relationship
Bass	Spectral centroid (lower values correspond to greater bass)
Brightness	Spectral centroid (higher values correspond to greater brightness)
Melody	Inharmonicity (higher values correspond to greater melodic content: e.g., passing notes and dissonances)
Noise	Spectral flatness (higher Wiener entropy values correspond to more noisy sounds)
Rhythm	Total onsets per unit time (higher onset rates correspond to greater rhythmic dynamism)
Loudness	Acoustic intensity (higher acoustic intensity corresponds to greater loudness)

The Diversity Indices and the Seed Item Recommendation. We inferred each participant's diversity of musical taste from their liking and familiarity ratings in the pre-experiment questionnaire, with higher ratings for multiple genres indicating more diverse listening habits than lower ratings. The 100 excerpts were sorted in descending order according to the individual's main musical feature preference (as in Table 1). For the *Diversity Index: Enjoyability* (DI:E), each participant's genre enjoyability ratings from the questionnaire were summed (to a potential maximum score of 84; 12 items receiving the maximum 7 rating) where greater diversity (above the midpoint score, 48) was used to increase the number of potential seed items available for random selection (and vice versa) (see Supplemental Materials). For the *Diversity Index: Familiarity* (DI:F), (maximum score again 84; 12 items receiving a maximum 7 rating), when a participant's summed genre familiarity ratings exceeded the midpoint, seed item choice was restricted to items in our corpus whose AMC website view count was less than our corpus mean of 1,227 views, and vice versa. This procedure sought to maximise the likelihood that users with low diversity scores were presented an acceptable seed item (serendipity), but also that users with high scores were exposed to items that are relatively infrequently accessed in the AMC dataset, to encourage these participants to experience the long-tail items even among the uniformly unfamiliar library. The liking and familiarity ratings that we achieved (see Results) confirmed that our seed recommendations were appropriate, even though we did not uniformly optimise the likelihood of high ratings, as just indicated.

The Subsequent Recurring Recommendation Algorithm. After the seed item was chosen, it was removed from the available dataset. Subsequent auditioned items were similarly removed, so that every item was heard just once (sampling without replacement). Unlike other music recommender system, our prototype uses each participant's continuous two-dimensional ratings of arousal and valence to provide customised recommendations during the whole procedure following the seed item presentation. Consequently, we used two sorted versions of the item database. One database was permanently sorted by mean energy, item by item (descending order), intending to represent the influence of acoustic intensity on perceived arousal. The other version of the database was sorted by the acoustic measure corresponding to the participant's chosen most important musical feature (Questionnaire Q9; Table 1), high to low, which we chose to represent the key influence on the perceived valence dimension. Where two musical features were identically rated, the first feature in the questionnaire was chosen to represent the valence dimension. An alternative (which we did not assess) is a random choice between the tied features. In the case of acoustic intensity being chosen as a valence parameter, both databases were sorted according to energy.

From the final 5 seconds of playback for each item, we took the user ratings for both the arousal and valence dimensions (sampled at 2 Hz), and then calculated the mean of for each dimension. To find the "similar" recommendation, we took the higher of the two mean values and in the remaining corpus chose the item with the closest acoustic parameter value (e.g., if valence had the higher mean, and for the particular participant we had determined that the valence dimension would be represented by "bass", then we found the item with the closest spectral centroid value). In the event that the mean values for valence and arousal were identical, valence was selected given we had relevant personal preferences for the related acoustic parameter. To select a "dissimilar" item, we evaluated whether the acoustic features Mahalanobis distance (M) (described above) of the chosen "similar" item was above or below the mean *M* for the current corpus (i.e., allowing for the fact that items are progressively removed from the available dataset as listening proceeds). The recommended dissimilar item was either the lowest or highest *M* value from the available corpus (respectively, when the "similar" item had an *M* value above or below the present corpus mean). To avoid repeat auditioning of excerpts, each item and its data was removed from the corpus after auditioning (i.e., the available corpus progressively contracted). The original dataset's mean *M* value was 11.88; the median 7.28. Using mean values avoided interpolating between two values to obtain the median when the remaining corpus count was even. The few items with very high *M* values were generally auditioned within the first ten sequence items, because of a predominance of requests for dissimilar items. When only one item remained in the database, that item was presented regardless of the user's request for similar or dissimilar. A detailed description of the system is shown in Appendix S4.

Linear Mixed Effects (LME) Modeling of Serial Responses to Items. LME cross-sectional time-series models of serial liking and familiarity responses to items were constructed in the *lme4* package in R, permitting assessment of both fixed effects, autoregression, other potential sequential effects, and random effects by participants and items, to reveal how these factors themselves varied. Cross-sectional time series analysis maintains the integrity of very individual time series of responses, rather than aggregating them, as is often done. It also avoids the misplaced assumption that the data are independently and identically distributed. Our analytical approach allowed the model predicting the familiarity response to item *n* to use its liking response, and vice versa for the liking model. Conversely, a purely predictive model would normally only permit information available prior to the event to be used. The data comprised the complete serial sequences of item responses (100 items, 69 participants) for each participant, analysed in single models for liking and familiarity. Ordinal Likert data were treated as continuous, as required by the *lme4* package. We compared two approaches to our models: decremental, starting from a maximal model containing all hypothesised and design-driven predictors and then removing unnecessary predictors, and additive, using previous best models as the foundation for a new model and then adding potentially effective predictors.

In both approaches, we refined the model based upon the following criteria. We removed statistically or quantitatively insignificant predictors progressively, seeking parsimony with the following provisions: minimising the Bayesian Information Criterion (BIC), while allowing for the complexities of defining the degree of freedom in random effects models. Models that differed in BIC by less than six were construed as not distinguishable from each other. We sought to minimise the RMSE (root mean square error) between the model predictions and the data, and subject to the BIC, chose the more parsimonious models for further assessment. Selection among the best performing models was achieved by the likelihood ratio method. The quality of the selected model was further assessed by confirming that its residuals retained no autocorrelation and by graphical checks, including checking the distributions using quantile–quantile plots.

Procedure for Real-Time Continuous Perceived Affect Responses and Post-Listening Liking and Familiarity Responses. Listening to each item, participants used a computer mouse to continually represent their perception of valence and arousal in a two-dimensional "emotion space" (Bailes & Dean, 2009; Dean & Bailes, 2010; Gabrielsson, 2016; Schubert, 1999, 2004). The emotion space axes were labelled "expressing" or "not expressing" arousal and expressing "positive" or "negative" valence, to emphasise our concern with perceived, rather than felt, emotion (Gabrielsson, 2001). The mouse coordinates and delta values on the emotion space were logged at 2 Hz as mouse pixel locations in MaxMSP relative to the main window (0,0 being top left; both axes ranging for 0.0 to 1.0, with the centre of the main window at 0.5, 0.5).

Participants first received a verbal description of the study, and verbal instructions on conducting the study, followed by further onscreen instructions for continuously rating each item played to them. Finally, participants were given one practice item to familiarise themselves with the rating process and to experience a musical item. These three strategies were aimed to mitigate any primacy effect. Prior to the beginning of each item, a "GO!" button appeared at the centre point of the emotion space, to centre the cursor on both axes, and begin a countdown of 3 seconds, to ready them for the next item. After each item, participants rated their familiarity with, and liking of the item (Likert scales where 1 = "not familiar" or "not likeable" and 7 = "familiar" or "likeable"). The post-item ratings of liking and familiarity were not used for the recommendation: for this, as described above, acoustic features were used as recommendation selectors, driven by participants' continuous representations of perception of arousal and valence.

After each item, the participant was then offered two choices of music: "similar" or "dissimilar", and the recommender system presented the selected item using the process described above. Participant responses to both the rating of the previous item and their choice of similar/dissimilar for the next item were saved. This process was repeated until all 100 items had been auditioned once. The experiment lasted approximately 1 hour, including questions, practice, and auditioning the items.

Results

Liking, Familiarity, and Influences of Time and Seeding: Mitigating the Cold-Start Problem

Figure 1 shows aggregated time courses for all participants' ratings of liking and familiarity. The first striking observation is that all the ratings are very low—well below the midpoint (4) of the scales. This immediately confirms how different our conditions are from those of most recommender systems, even those in which exploration of a long tail is encouraged (Celma, 2010). In most systems studied, mean liking ratings are between 4 and 5 on a 1 to 5-point scale (our scale is 1 to 7, so these would correspond to 5.6 to 7). Figure 1 also shows that there is hardly any cold-start issue (arguably supporting H1), as the data remain "cold" throughout. The first few observations are not the lowest rated, though the lowest values do occur within the first dozen or so. This is considered further below.

There is apparently a slight progressive increase in both ratings across the experiments, with a modest positive linear regression coefficient between liking or familiarity and sequence item number (partially supporting H2). Note again that each sequence item rating represents responses from 69 people to a maximum of 69 different items. Regressions indicated a significant moderately positive predictive influence of sequence item number on familiarity (F(1,98) = 23.06, p <.001), with an R2 of 0.182, and a less positive insignificant relationship for liking (F(1,98) = 2.995, p = .086), with an R2 of .019. This suggests that exposure to each item marginally increases mean familiarity (ß = .003), with the same (as yet) little effect on liking (ß = .001). A Spearman two-sided correlation test between liking and familiarity considered here by sequence item number is r = .51, p <.001. Further, Spearman two-sided correlation tests between familiarity and sequence item number found a stronger positive correlation (r = .05, p = <.001) than between liking and sequence item number (r = .02, p = .07).

While liking and familiarity rise slightly in Figure 1 and are significantly correlated, when the post-item mean liking and mean familiarity is calculated by item (instead of by sequence item number), the correlation between them is much stronger (r = .94, p <.001). The items are irregularly distributed in time; thus, this result strongly supports H4. We investigated this relationship further by calculating the mean expanding window average of post-test liking and familiarity ratings by sequence number. This also allows us to assess more closely H1, that ratings of the seed item chosen using participant profiles are comparable with those of subsequent items (i.e., mitigating the cold-start problem). This analysis is shown in Figure 2.

The grand average sequence kinetics show three phases: an opening phase of ~7 items where liking, and familiarity drop rapidly, followed by a rapid recovery to approximately item 20, and finally a long subsequent phase in which both gradually rise. Figure 2 shows that despite this initial sharp drop, H2 is generally supported, insofar as there is an upward trend in familiarity (and to a lesser extent liking). Furthermore, the responses to the seed items (which include 53% of all items) are competitive with the long-term responses. This additional evidence is again consistent with H4, that liking, and familiarity are closely related.

We then performed similar mean expanding window averages of post-test liking and familiarity ratings by sequence number, separated by whether the user previously requested a similar or dissimilar item. This is shown in Figure 3.

Figure 3 reveals the origins of the trends in Figure 2 more clearly, by indicating the distinctive behaviours following "similar" versus "dissimilar" user requests. Similar requests show an immediate rise in familiarity and liking (though followed in this case by a transient drop), reaching overall maximal values within 25 items (suggesting that we quite rapidly identify the items a particular user will find most appealing). Consequently, there is a slow drop in both familiarity and liking ratings for the "similar" items thereafter, plateauing at about sequence item 50. Dissimilar request items show the initial drop already apparent in Figure 2 (being the dominant response

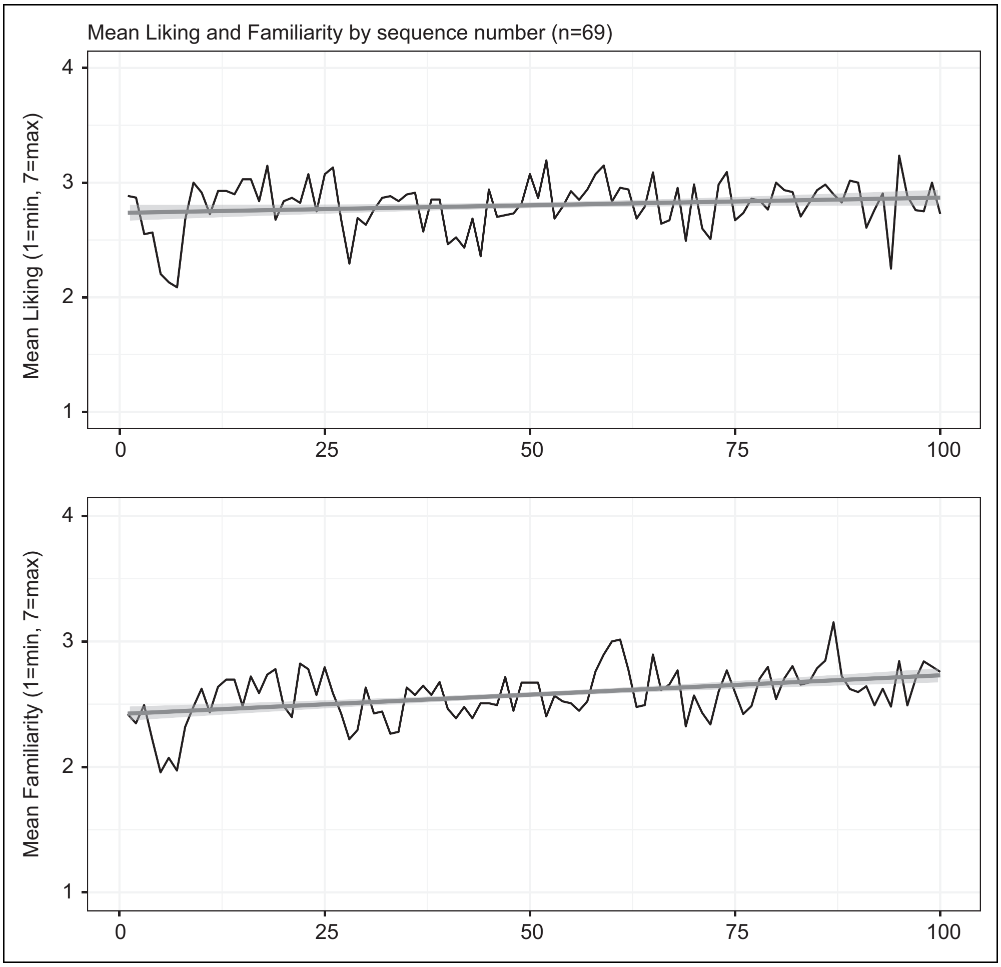

Figure 1. Regressions between sequence item number and mean liking and familiarity ratings. The shading around the regression line represents 95% confidence interval.

choice throughout). Whereas familiarity in the dissimilar request time series eventually rises to ratings comparable to those at the outset, liking ratings rise only to a lower value. These results confirm the limited relevance of the cold start concept here, because every item and user is relatively "cold", and confirm that liking ratings commonly lag behind those for familiarity. Overall, we cautiously interpret Figures 2 and 3 as revealing the complex underpinnings in the early stage of inverted-U responses for both familiarity and liking (cf., Chmiel & Schubert, 2018) in our unusually and uniformly unfamiliar dataset, as we next assess further.

A changepoint (*cpt* package in R software) analysis based on joint changes in mean and variance of the data in Figure 2 (asymptotic penalty value = 0.05, AMOC), revealed changepoint locations of 20 for liking and 21 for familiarity, thus appropriately amalgamating phases 1 and 2 described above. Spearman correlations for the post-changepoint segment, 21–100, for both liking and familiarity with sequence item number are shown in Table 2.

This analysis shows the second changepoint segment encompassing 80% of items and is strongly coherent with H2 (that familiarity and, to a lesser extent, liking will increase during a listening session, that is, with extent of exposure), as the L ~ S and F ~ S correlation coefficients are high and significant. Familiarity for the seed item was greater than for all later windowed averages, and competitive for liking, and not exceeded until at least 20 items had been auditioned (i.e., the start of the post-changepoint segment). Our data are consistent with H1, since our initial (seed) recommendation (based on participant diversity indices and corresponding acoustic choices) attracted quite favorable responses, and consequently, our attempt to reduce the cold-start effect was beneficial. These results are also consistent with repeated dissimilar item choices at outset, which combined with the item selection algorithm meant that items with an *M* value closest to the mean (e.g., less acoustically extreme items) were presented later on. The progressive increase in liking and familiarity across the sessions also support H5/H6, that choice of acoustic

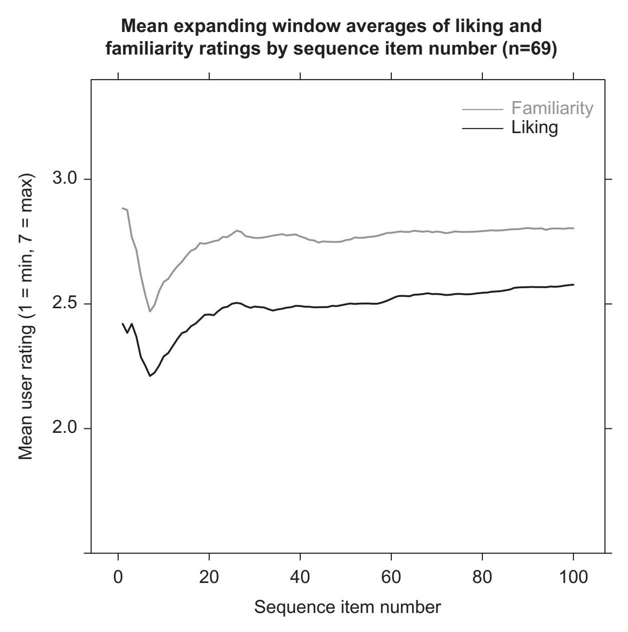

Figure 2. Mean expanding window (cumulative) averages by sequence item number (1–100) for liking and familiarity, including all (similar and dissimilar) choices.

features, as implied pre-listening preferences, allows us to enhance the liking and familiarity responses.

To assess this further, we focused on the 53 items (among the overall 100) which appeared as a seed item (see also Appendix S5 for some summary statistics on these seed items) and used a Wilcoxon rank test to determine whether their ratings as seed differed from their ratings in the post-seed periods (either 2–100 or 20–100). These tests (Table 3) showed no significant difference in the ranking distributions which would agree with H1, that we reduced the cold-start problem and that seed items were not rated unfavorably compared to their rating as a non-seed item.

Similar Versus Dissimilar: Recurrent Recommendations After the Seed Item

The explicit prediction of H3, that an item provided as "similar" will have higher ratings than one provided as "dissimilar" (based on acoustic features), is supported by the data in Table 4. Wilcoxon unpaired rank sum tests of these liking and familiarity ratings with respect to items provided as "similar" versus "dissimilar" were significant (both tests p <.001), confirming support for H3. However, note that all mean values in Table 4 are below the midpoint of the Likert scale (i.e., participants felt unfamiliar with and did not like all items).

Over 75% of participants requested "dissimilar" successors to the first seven items, concomitant with the descent in the first phase of the moving window averages for liking and familiarity. This strategy is unsurprising, as it attempts to express continued aversion to the material and should ensure a rapid awareness of the full range of the material. Appendix S6 shows the similar/dissimilar responses by the excerpt eliciting the response, and by sequence item number for each participant. We see that the aversive behaviour is very strong across our participants, despite the fact that many of the individual excerpts elicited a similar response (e.g., they liked the excerpt and wanted a similar one). The choice of a forthcoming dissimilar item remained predominant across all 100 sequence item numbers, again consistent with the low ratings. For all participants and all sequence items, similar items were only chosen 32.8% of the time. Likewise, for phase 2 sequence item numbers 21–100, "similar" was chosen 33.8% of the time. Despite the upward trend of liking and familiarity in this segment, there was no trend for participants to choose "similar"' items more often. This may reflect continued optimism by participants that given their ratings of liking were low, there remained the possibility of finding more appealing items, which would logically be expected to be dissimilar to the previous item.

Implicit in H3 is that an item eliciting a request for a "similar" next item will itself be liked more than when the request is for a "dissimilar" item. Correspondingly, the mean liking and familiarity ratings for the items which elicited similar versus dissimilar requests are shown in Table 5: Wilcoxon unpaired rank sum tests of these liking and familiarity ratings with respect to items eliciting similar versus dissimilar requests were both significant (both tests p <.001), confirming further support for H3.

Time Series Models of the Liking and Familiarity Response Series

The analyses so far suggest that our recommender system was beneficial, even though liking and familiarity remained below the median value throughout. The recommendations were based on acoustic features, either translated from musical feature preferences of the users indicated in the questionnaire, or from their affective responses during listening. Thus, the value of using acoustic features in recommendation, even in these negative conditions, is strongly supported. In this section, we model the response process itself, to assess possible cognitive influences of the sequential ratings choices themselves and of the user preferences (and other features), and to determine whether additional specific acoustic influences remain important.

We established previously that there are close correlations between liking and familiarity responses (H4). Here, we investigate the influences of factors such as autoregression (the commonly critical sequential influence of modelled responses upon themselves), the user request (0 = "dissimilar", 1 = "similar"), exposure (i.e., sequence item number), and acoustic features upon models of liking and familiarity, using linear mixed effects (LME) cross-sectional time-series analyses. This allows maintaining the integrity of all individual response time series. The

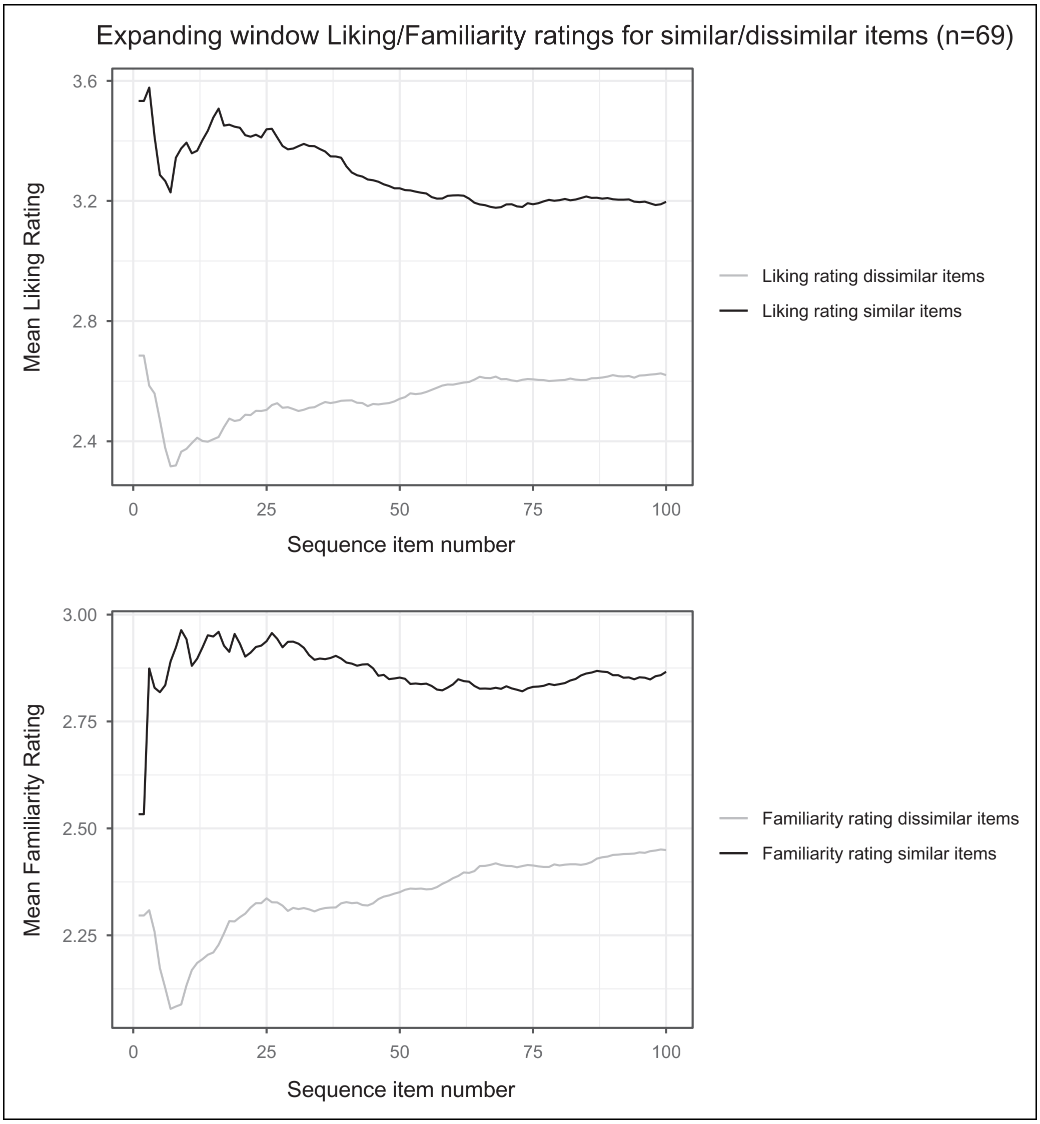

Figure 3. Mean expanding window (cumulative) averages by sequence item number (1–100) for liking and familiarity, and by previous choice of similar or dissimilar item. (Since the seed item does not have an eliciting user's "similar" or "dissimilar" choice, we have used the choice it elicited to separate the values for the seed, item 1).

Table 2. Spearman correlations tests of post-changepoint segment of the mean expanding window averages for liking, familiarity and sequence item number. L = liking, F = familiarity, S = sequence item number. Note that S = 20–100 where L ~ S, S = 21–100 where F ~ S, and both L and F are length = 21 when L ~ F.

Correlation test	Statistic	*p*-value	Estimate (rho)
L ~ S	15,878	<.001	0.81
F ~ S	4,444	<.001	0.95
L ~ F	9,714	<.001	0.89

analyses permit the delineation of fixed effects (such as those aforementioned) as well as random effects (the influences of inter-individual participant and inter-item differences upon responses).

The pacf (partial autocorrelation function) across a random selection of individual response time series showed significance in lags 1 to 5 for both liking and familiarity, although varying by participant. This informed our initial model, which considered autoregression, the preceding user request, and sequence item number. We refined and assessed for quality: see methods for more detail on model selection. The resultant selected models are shown in Table 6.

Table 6 shows that liking and familiarity were both autoregressive and mutually predictive. Given the autoregression, and the user request predictors, sequence item number was not a predictor: in other words, the dependence of ratings upon exposure described above was explicable in terms of the other factors. We found a significantly positive influence of the previous response request for a "similar"

Table 3. Results of the Wilcoxon paired one-tailed rank sum tests for the mean liking and familiarity ratings for the seed item against the same items in the later changepoint phases (2–100, and 20–100 for liking; 21–100 for familiarity). Thus, our exploitation of user preferences, and the resultant diversity index and feature importance rating enhances listener responses to the seed item.

Liking/ Familiarity	Condition 1	Mean	Condition 2	Mean	*p*-value
Liking	Seed item	2.89	Mean 2–100	2.80	.3443
Liking	Seed item	2.89	Mean 20–100	2.80	.3034
Familiarity	Seed item	2.42	Mean 2–100	2.57	.7794
Familiarity	Seed item	2.42	Mean 21–100	2.57	.7575

Table 4. Mean familiarity and liking responses for items following the similar/dissimilar recommendations, with *SD* shown in parentheses.

	Liking *M* = 2.80 (*SD* = 1.79)	Familiarity *M* = 2.57 (*SD* = 1.70)
Similar	3.18 (1.87)	2.85 (1.80)
Dissimilar	2.61 (1.73)	2.44 (1.64)

Table 5. Mean (*SD*) familiarity and liking responses for the items eliciting requests for "similar" or "dissimilar" recommendations.

	Liking *M* = 2.80 (*SD* = 1.79)	Familiarity *M* = 2.57 (*SD* = 1.70)
Similar	3.90 (1.86)	3.32 (1.91)
Dissimilar	2.26 (1.49)	2.20 (1.45)

item on liking. For Familiarity, the influence of the previous response request was negative, and apparently inconsistent with earlier observations. But we note that the models (see methods) of both liking and familiarity included a Lag0 contribution from each other, with a high coefficient: in other words, some information from the item whose response is being predicted, is already included. Furthermore, uniquely in the familiarity model, Lag1 of both liking and familiarity is included, corresponding to the item eliciting the "previous response" request of the participant: "similar" or "dissimilar" (note again that this results in the recommender system providing items based either on a single acoustic feature for "similar" items, or on *M* values for "dissimilar" items). Thus, the selected familiarity model has an overlap of information sources from the previous item (both its liking and familiarity, and the request that it elicits). This overlap of information accounts for the negative coefficient on previous response in the familiarity model: when all Lag0 and Lag1 information is removed from the Familiarity model (worsening the model), the previous response coefficient becomes positive and of a similar order to the Liking model. Therefore, the negative coefficient is applicable only in the context of the larger set of additional predictors, and the earlier observations are not challenged, rather enhanced by these LME models.

Our second set of LME time series models appended the participant ratings for musical feature importance as possible predictors, to investigate whether pre-listening feature preference could enhance the models above the previous models of Table 6. The resultant selected models are shown in Table 7.

The models of Table 7 improved BICs (compared with Table 6), without degradation in the RMSE values. Preference for rhythm, and additional lags for liking (4) and familiarity (1) were retained after model selection as a significant predictor of liking, and preference for noise contributed to the familiarity model. Other autoregressive and predictive features were retained from Table 6 with only slight modification. Likelihood ratio tests compared the models of Table 7 with the corresponding ones of Table 6 (though this, and subsequent tests required the omission of the data from the three participants whose musical feature preferences were lacking): the later models were highly preferred (p <.001 in both cases). Pre-listening musical feature preferences were thus useful predictors of responses (upholding H5).

Our third set of LME models considered as predictors acoustic features of the items in addition to the those included in Table 7. In this additive approach, our best liking model included spectral kurtosis, although the BIC (20,881.38) was significantly worse than the previous best liking model (MLMF13; BIC = 20,867.67). The two models showed the same RMSE. Our best familiarity model included roughness although the BIC was significantly worse (18,197.19) than the previous best model (MFMF10; BIC = 18,185.57), but again with very similar RMSE. Likelihood ratio tests on both liking and familiarity models confirmed that these models with acoustic features did not improve upon the previous best models in Table 7. This approach did not support to H6; but it needs to be recalled that the recommender system already uses acoustic information as part of its item selection process, and its success is already an indication of the impact of that information.

To confirm the validity of these model selection processes, we also undertook a decremental modeling approach (see methods in Participants, Materials, Methods, and Procedure section), progressively refining an initial model that included all putative predictors. This supported our conclusions. Correspondingly, the best models (Table 7) accounted well even for participants who failed to complete the experiment (characterised by predictions, responses and residuals that account for only a portion of the 100 sequence items), and for a few cases of monotonous responses (where liking and/or familiarity responses were rated as consistently low). Figures 4 and 5 show the actual liking and familiarity responses for participants 21–23, chosen as

Table 6. Parameter estimates and fit statistic of the selected LME models for Liking and Familiarity. Random effects are shown in brackets. *SD* = Standard deviation, ID = Participant ID. Note: User request (previous response) denotes the participant choice of similar item (1) or dissimilar item (0). Lags are shown as L1Liking = Liking with a Lag of 1, etc. The nomenclature of the models comprises absolute mean differenced data (M), as well as Liking (L), Familiarity (F), as well as later in the results, genre preferences (G) acoustic features (A) and musical feature preferences (MF). Sequence item number was not statistically significant in either of these models.

		Coefficient					
Model designation/	Effect	Estimate					BIC
Response variable	(random)	(variance)	*SD*	*t*-value	*p*-value	sig	(RMSE)
ML10	(Excerpt no.)	(0.06)	0.24				22,169.6 (1.314)
Liking	(ID)	(1.74)	1.32				
	User request	0.149	0.042	3.564	<.001	***	
	L3Liking	0.058	0.011	5.385	<.001	***	
	L0Familiarity	0.551	0.013	41.005	<.001	***	
MF6	(Excerpt no.)	(0.05)	0.23				19,262.36 (1.045)
Familiarity	(ID)	(0.50)	0.70				
	User request	−0.206	0.037	−5.610	<.001	***	
	L0Liking	0.35	0.009	39.301	<.001	***	
	L1Liking	0.108	0.010	10.310	<.001	***	
	L1Familiarity	0.098	0.012	7.917	<.001	***	
	L2Familiarity	0.069	0.010	6.423	<.001	***	
	L3Familiarity	0.11	0.010	10.291	<.001	***	
	L4Familiarity	0.061	0.011	5.807	<.001	***	

Note. ‘ *** ’ $p < 0.001$, ‘ ** ’ $p < 0.01$, ‘ * ’ $p < 0.05$, ‘ . ’ $p < 0.1$, ‘ ’ $p < 1$.

Table 7. Parameter estimates and fit statistic of the best model (LME, random plus fixed effects) to estimate Liking and Familiarity with lags based upon acf and pacf assessment. *SD* = Standard deviation, ID = Participant ID. Note: User request (previous response) denotes participant choice of similar item (1) over dissimilar item (0). Lags of Liking and Familiarity are shown as L1Liking = Liking with a Lag of 1, etc. Sequence item number was not statistically significant in either of these models.

		Coefficient					
Model designation/	Effect	Estimate					BIC
Response variable	(random)	(variance)	*SD*	*t*-value	*p*-value	sig	(RMSE)
MLMF13	(Excerpt no.)	(0.06)	0.25				20,867.67 (1.285)
Liking	(ID)	(0.36)	0.60				
+	User request	0.199	0.044	4.550	<.001	***	
Feature importance	L3Liking	0.044	0.011	3.998	<.001	***	
	L4Liking	0.03	0.011	2.759	.00582	**	
	L0Familiarity	0.568	0.014	40.769	<.001	***	
	L1Familiarity	-0.039	0.014	-2.732	.00631	**	
	Rhythm preference	0.195	0.016	12.336	<.001	***	
MFMF10	(Excerpt no.)	(0.04)	0.21				18,185.57 (1.024)
Familiarity	(ID)	(0.34)	0.58				
+	User Request	-0.217	0.037	-5.859	<.001	***	
Feature importance	L0Liking	0.366	0.009	40.284	<.001	***	
	L1Liking	0.076	0.011	7.046	<.001	***	
	L4Liking	-0.035	0.010	-3.484	<.001	***	
	L1Familiarity	0.108	0.013	8.497	<.001	***	
	L2Familiarity	0.059	0.011	5.430	<.001	***	
	L3Familiarity	0.099	0.011	9.085	<.001	***	
	L4Familiarity	0.077	0.012	6.342	<.001	***	
	Noise preference	0.131	0.020	6.521	<.001	***	

Note. ‘***’ $p < 0.001$, ‘ ** ’ $p < 0.01$, ‘ * ’ $p < 0.05$, ‘ . ’ $p < 0.1$, ‘ ’ $p < 1$.

representative of a variety of response types we observed, together with our modelled liking and familiarity and the corresponding residuals for these individuals. Such comparisons are among our routine assessments of model quality, together with confirmation that residuals essentially lack autocorrelation.

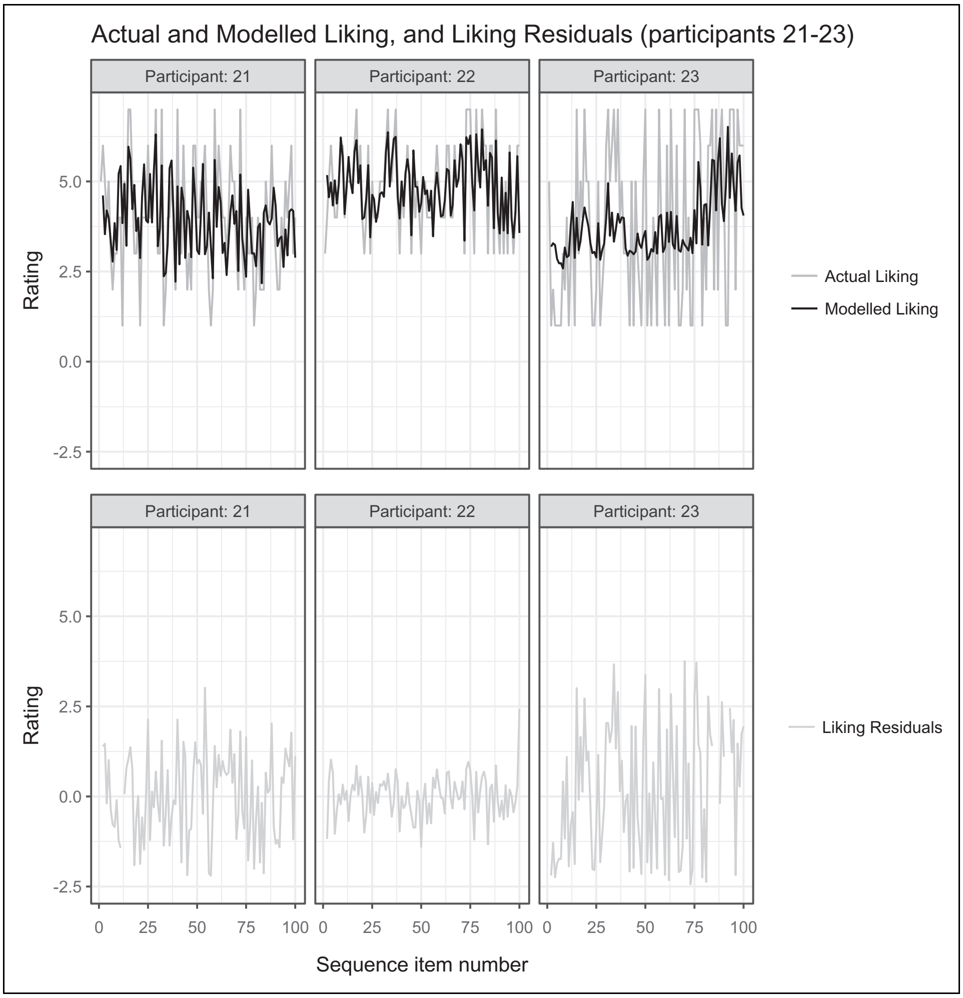

Figure 4. Actual and predicted liking, and liking residuals for model MLMF13.

Discussion

The prototype recommender system seems to successfully use acoustic features that are "translated" from users' pre-exposure preferences, and from their within-experiment continuous affect responses, so as to make effective predictions. This can be judged by the relative lack of a cold-start effect of our recommended seed item, even given our drastically "cold" and uniformly unfamiliar and unliked material; the progressive increases in familiarity and liking even in these circumstances; and the more favourable responses to items which elicit a request for the next item to be "similar", as well as to the items provided in response to a similar request (compared to corresponding items eliciting "dissimilar" requests and for their responding recommendations).

Note that practical considerations (such as cost) prevented us from including a control condition, in which participants received random items, indifferent to their choice of "similar" or "dissimilar", and so there are necessary limitations on the interpretation of our data. While our system is in no way yet optimised, it nevertheless behaves differently from and better than what would be expected of a random recommendation system. In the circumstances of our experiment, random recommendations would mean that, across participants, every sequence item number has the same likelihood of receiving any of the items. Thus in contrast to what we observe, and given a large enough participant group, there could be no utility to the genre and feature preferences of the users, and no dependence of familiarity and liking on the user requests (whether for "similar" or "dissimilar"), nor in all probability would there be the complex two-phase kinetics we observe. On average, the ratings at each sequence step would be the mean of all ratings, though they might conceivably still change slightly as exposure increased. Modest exposure was accompanied here by an increase in ratings in our data, but our observations can best be explained by the autoregression of the ratings themselves, together with the user choices, and not by the sequence item number per se, so there is no obvious reason to expect any upward trend in ratings given random item presentation. Indeed, in a second study on this music library (submitted), with recommendations uninfluenced by user responses, we

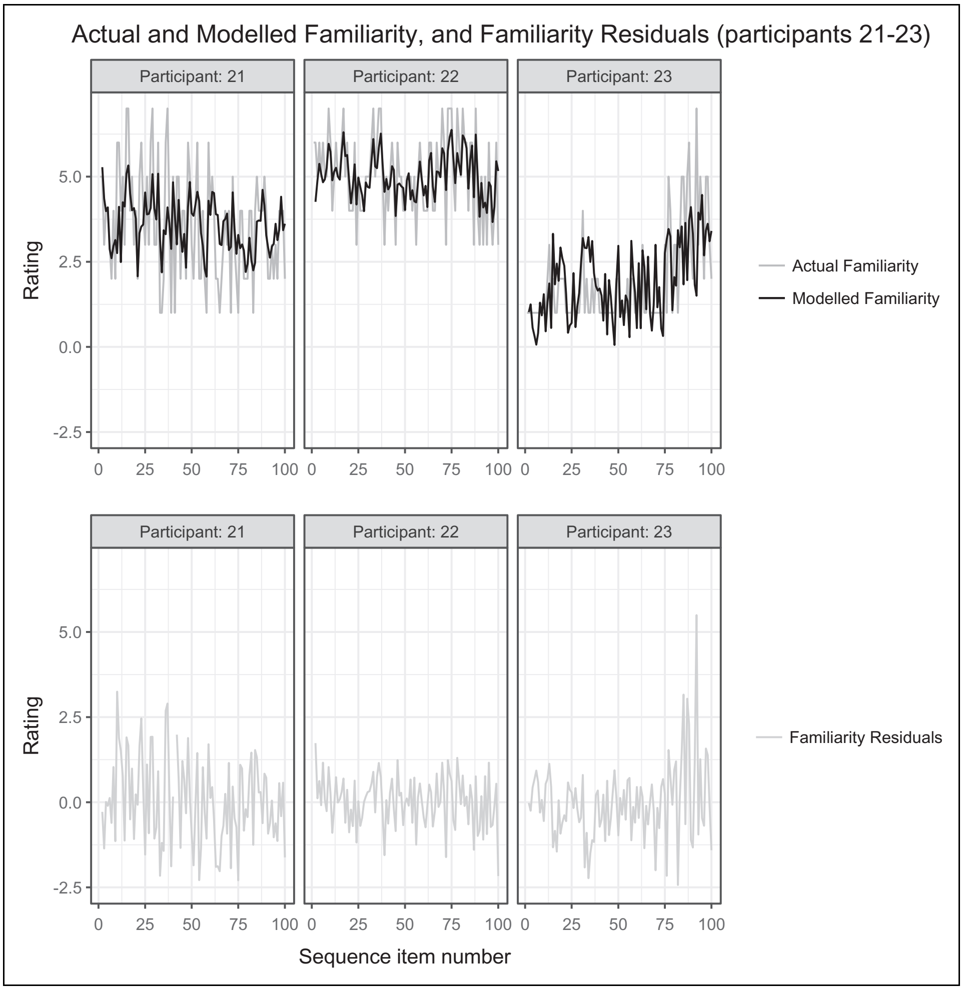

Figure 5. Actual and modelled familiarity, and familiarity residuals for model MFMF10.

found no significant change in familiarity and liking with respect to sequence item number. Altogether, it is clear that our system achieves recommendations with some utility. In the second study we compare four different exposure conditions (cf., Weigl & Guastavino, 2011) for their effects on user's duration of attention to presented items, as part of the process of enhancing our system.

Next, we consider our six hypotheses in turn in more detail: most are supported, some are not. Clearly the number of participants in our study is substantial by the standards of interventional psychology experiments, but it is skewed strongly towards the undergraduate age range (63% of participants were aged 17–21 years). Thus, we suggest caution with the results on demography and pre-exposure user taste (see Supplemental Materials), and some cases of a lack of clear-cut result may be due to the sample nature.

H1 proposed that information about participants' pre-listening preferences, translated into acoustic features, could mitigate the cold-start problem and achieve seed item ratings comparable to later ratings. Our approach using diversity indices was largely successful as the ratings for the seed item were generally higher than many of those for the rest of the items, even though most items were poorly rated (see Figure 2). This was particularly the case for liking and occurred even though we balanced our efforts towards providing an acceptable item to participants with low diversity preferences, with the provision of low AMC-user access items to those of our participants with high diversity preference. Thus, H1 was supported. Future work could further develop the diversity indices by including additional variables, such as socioeconomic data (as collected in our questionnaire) and testing whether these new variables are influential in addition to the current diversity indices. When enough data becomes available in the future for collaborative filtering, such approaches will be entirely appropriate.

H2, which was strongly supported, suggested that familiarity (and to a lesser extent liking) will increase during a listening session, with item exposure. At the beginning of the exposure, a notable brief decline (over around seven successive items) in liking and familiarity was subsequently reversed and overcome. This implies a short-term effect of exposure in a listening session distinct from the enhancement of ratings after longer term exposure.

Interestingly, we found a predictive contribution of rating lags in our mixed effect models of liking and familiarity of the auditioned items. Not only do liking and familiarity increase over a listening session, but our models have significant autoregressive lags of up to order 4, suggesting that we may be able to use more information about acoustic features and individual participant perceived affect from the immediately preceding four items as part of the recommendation.

H3 proposed that when listeners request a "similar" item, liking and familiarity of the subsequently provided item will be higher than when a "dissimilar" item is requested; it also implied that the same could be expected of the item which elicits the "similar" request. H3 was confirmed in both respects. This supports our approach to recommending "similar" and "dissimilar" items, making use of users' continuous response ratings and acoustic parameters (most notably Mahalanobis distance). Further work may allow us to improve the recommendations. Firstly, our recommendations used listeners' affect responses to each item, but only averages over the last 5 seconds of a 30-second item, rather than taking coefficients from a full time series model of the relation between acoustic features and perceived affect, which we will address in another study. Secondly, our recommendation approach prevented participants actively ending item auditions, and yet the optimum point at which we measure a response may depend on the engagement of the user and/or item (Olsen et al., 2014). Finally, as implied by the serial autoregressive effects noted above, the sequence of choices of "similar"/ "dissimilar" itself might have predictive power: for example, the more previous successive user requests for "similar", the more likely the next request will be "dissimilar". However, we did not demonstrate such effects over lags beyond 1. We note also that a participant may be influenced by their "similar"/"dissimilar" choice per se in their response to the next item, such that a "similar" request tends to generate a more positive response regardless of the proffered item, as might be implied by the serial dependency just mentioned. We cannot presently separate this possibility entirely from the intended influence of the item selection itself.

H4, that familiarity and liking are closely related, was supported by their strong correlations, and by LME models confirming they are mutually positively predictive. Encouragingly, this suggests that a participant can be persuaded to become familiar with music, and eventually like it, although the items' generally low ratings illustrates the continuing difficulties faced in developing a recommender system for unfamiliar music.

H5 suggested that participants expressed prior preferences for musical features (such as 'bass') might predict individualised liking and familiarity responses to items. In two cases this was upheld: with the feature rhythm, for item liking, and with noise, for item familiarity. We also found an interesting phenomenon whereby the feature preference "noise" as a predictor of familiarity, could be replaced by pop music familiarity ratings with almost identical effect (not shown). Although not easily explained by our data, there may be one plausible explanation: our population is mainly of the 17 to 34 years age range; a generation whose popular music has been characterised largely by "loudness wars" or a reduction in dynamic range (Robjohns, 2014). Consequently, we may find that participants closely associate such a loss of dynamic range with "noise," rather than volume (implied by the absence of the loudness parameter in our selected LME models).

In our autoregressive models these preference predictors rendered the acoustic predictors ineffective (contrary to H6). This is not surprising since both in the seed and the subsequent recommendations, user responses (prior preferences or current perceived affect respectively) were interpreted in terms of acoustic features that then drove the item recommendation, so that the influence of acoustic features had already been built in. The success of the Mahalanobis distance as a basis for recommendation can be understood in the light of the fact that it is a relative measure of all acoustic features, rather than a single feature, and our list of musical features may comprise, or be interpreted by listeners as conglomerated acoustic features. This again will tend to over-ride the potential predictive modeling influence of acoustic features. Clearly the significant impact of the acoustic features in our system is as yet poorly characterised.

Most of our hypotheses were supported by the results presented, and our prototype recommender system already shows utility. This is despite arbitrary system aspects, constructed a priori by necessity: notably, seed item choice and recommendation precision from participant request, whose empirical interrogation can be done in future work. Moreover, the full depth of the time series continuous affect responses remains to be mined. Previous cross-sectional time-series analyses have shown powerful relationships between acoustic features and these responses (Bailes & Dean, 2012; Olsen, Dean, Stevens, & Bailes, 2015). By extending analyses to obtain model parameters specific to each participant (perhaps on an ongoing basis during exposure) we could then formulate more precise predictors for retrospective liking and familiarity responses and thus the choice between "similar" and "dissimilar" requests (cf., Zhao, 2014).

Future work should include developing a similar online version of our system to enlarge and enhance the interpretation of this study, perhaps overcoming some of the identified limitations. Such an online system could adopt a similar approach to that we have taken here, by asking new users to complete a questionnaire when signing up to the online database (note that currently an account is required to purchase items from the music library). This, however, may prove to be a barrier for the initial engagement of new users, as this may be perceived as too burdensome to complete. This problem could be alleviated by providing incentives to users to complete the questionnaire (such as the AMC offering discounts to purchases in their collection), although this strategy is likely to prove unsustainable in the long term and indeed largely unnecessary: the AMC could evaluate user preferences and purchase data as they become available, with a view to moving towards a more conventional recommender system type, such as collaborative filtering. One possible way this could be achieved, could be to make use of social collaborative filtering (Sedhain, Sanner, Braziunas, Xie, & Christenson, 2014) where users link a social media account to their AMC account, so that similarity measures can be at least partly derived from side information (basic demographics, "Like" information, etc.).

However, there are still limitations with such an approach as there may not be any acoustic preference data available to link with items. A simple and rapid way of gaining an impression of the acoustic preferences of a user may be to ask them to nominate a few composers/artists whose work they most like (and then gather the corresponding acoustic information, even in real time from iTunes or Spotify). Additionally, a small group of questions addressing techniques of consumption (not referring to delivery platforms, but to modes of approach to finding music, related or unrelated to prior consumption), may be very valuable in providing recommendation predictors and in reducing the demands of a questionnaire. Work from our group by Chambers (submitted) provides support for both real-time acoustic data analysis, and consumption data, in the specific context of Australian art music that we focus on here.

Further useful areas of work may be to evaluate the effect of the continuous ratings task on participant engagement, or whether this can be enhanced in other ways: a major barrier to exploration of unfamiliar music is that people only engage for a few seconds. In a succeeding experiment, we are consequently assessing the influence of different experimental conditions on participants' (voluntary) listening time, whereas listening to the whole of each extract was enforced during the present study.

Author contribution

Experiment design: JRT and RTD; modelling: JRT and RTD; writing: JRT and RTD.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We are grateful for support towards this project from the Australian Research Council (grant no: LP150100487).

ORCID iD

John R. Taylor https://orcid.org/0000-0002-4435-0657

Action Editor

Dr. Frank Hentschel, Universitat zu Koln Musikwissenschaftliches Institut, Germany.

Peer Review

Peter Knees, TU Wien, Faculty of Informatics.
Two anonymous reviewers.

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The ethnicity categories are as per classification 1249.0 – Australian Standard Classification of Cultural and Ethnic Groups (ASCEG) (Australian Bureau of Statistics, 2011).
2. The AMC repertoire navigator is available at https://www.australianmusiccentre.com.au/search/search?type=ish&if[browse]=true
3. We chose absolute mean difference between successive samples as the feature vector measure after assessing the impact of using mean and absolute mean values in our LME models, and as seen in previous studies (McAdams, 1999; Olsen et al., 2016). The flux value is a mean value as this already represents the absolute difference in change.

References

Aggarwal, C. C. (2016). *Recommender systems: The textbook*. Cham, Switzerland: Springer.

Allamanche, E., Herre, J., Hellmuth, O., Froba, B., Kastner, T., & Cremer, M. (2001). *Content-based identification of audio material using MPEG-7 low level description*. Proceedings of the 2nd International Symposium on Music Information Retrieval (ISMIR), Bloomington, IN.

AMEB. (2019, November 22). *Australian Music Examinations Board (AMEB), Exams*. Retrieved November 22, 2019, from https://www.ameb.edu.au/exams/exams.html

Australian Bureau of Statistics. (2011). *1249.0 – National Standard Classification of Cultural and Ethnic Groups (ASCCEG), 2011*. Canberra, Australia: Commonwealth of Australian, Australia Bureau of Statistics.

Australian Music Centre Ltd. (2019). *Australian music centre*. Retrieved November 22, 2019, from http://www.australianmusiccentre.com.au/about

Bailes, F., & Dean, R. T. (2007a). Facilitation and coherence between the dynamic and retrospective perception of segmentation in computer-generated music. *Empirical Musicology Review*, *2*, 74–80. http://doi.org/10.18061/1811/28854.

Bailes, F., & Dean, R. T. (2007b). Listener detection of segmentation in computer-generated sound: An exploratory experimental study. *Journal of New Music Research*, *36*, 83–93. http://doi.org/10.1080/09298210701755123.

Bailes, F., & Dean, R. T. (2009). Listeners discern affective variation in computer-generated musical sounds. *Perception*, *38*, 1386–1404. http://doi.org/10.1068/p6063.

Bailes, F., & Dean, R. T. (2012). Comparative time series analysis of perceptual responses to electroacoustic music. *Music Perception: An Interdisciplinary Journal*, *29*, 359–375. http://doi.org/10.1525/mp.2012.29.4.359.

Bennett, T., Emmison, M., & Frow, J. (1999). *Accounting for tastes: Australian everyday cultures*. Cambridge, UK: Cambridge University Press.

Bogdanov, D., Haro, M., Fuhrmann, F., Xambó, A., Gómez, E., & Herrera, P. (2013). Semantic audio content-based music recommendation and visualization based on user preference examples. *Information Processing and Management*, *49*, 13–33.

Bogdanov, D., Porter, A., Urbano, J., & Schreiber, H. (2017). *The MediaEval 2017 AcousticBrainz Genre Task - Content-based Music Genre Recognition from Multiple Sources*. In Proceedings of the MediaEval 2017 Workshop. CEUR-WS.org, 2017. 13–15 September 2017, Dublin, Ireland.

Bogdanov, D., Serra, J., Wack, N., Herrera, P., & Serra, X. (2010). Unifying low-level and high-level music similarity measures. *IEEE Transactions on Multimedia*, *13*, 687–701. http://doi.org/10.1109/TMM.2011.2125784.

Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, *12*, 331–370. http://doi.org/10.1023/A:1021240730564.

Caclin, A., McAdams, S., Smith, B. K., & Winsberg, S. (2005). Acoustic correlates of timbre space dimensions: A confirmatory study using synthetic tones. *Journal of the Acoustical Society of America*, *118*, 471–482. http://doi.org/10.1121/1.1929229

Casey, M. (2001). General sound classification and similarity in MPEG-7. *Organised Sound*, *6*, 153–164. http://doi.org/10.1017/S1355771801002126.

Celma, Ò. (2010). *Music recommendation and discovery - The long tail, long fail, and long play in the digital music space*. Berlin, Heidelberg: Springer-Verlag.

Chmiel, A., & Schubert, E. (2018). Using psychological principles of memory storage and preference to improve music recommender systems. *Leonardo Music Journal*, *28*, 77–81.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*, 1–18. http://doi.org/10.1016/S0169-7439(99)00047-7.

Dean, R. T., & Bailes, F. (2010). Time series analysis as a method to examine acoustical influences on real-time perception of music. *Empirical Musicology Review*, *5*, 152–175. http://doi.org/10.1371/journal.pone.0018591.

Dean, R. T., & Bailes, F. (2011). Modelling perception of structure and affect in music: Spectral centroid and Wishart's red bird. *Empirical Musicology Review*, *6*, 90–137. http://doi.org/10.1002/0471743984.vse6611.

Dean, R. T., Bailes, F., & Schubert, E. (2011). Acoustic intensity causes perceived changes in arousal levels in music: An experimental investigation. *PLoS ONE*, *6*, e18591. http://doi.org/10.1371/journal.pone.0018591.

Dean, R. T., Bailes, F., & Dunsmuir, W. T. (2014a). Time series analysis of real-time music perception: Approaches to the assessment of individual and expertise differences in perception of expressed affect. *Journal of Mathematics and Music*, *8*, 183–205.

Dean, R. T., Bailes, F., & Dunsmuir, W. T. (2014b). Shared and distinct mechanisms of individual and expertise-group perception of expressed arousal in four works. *Journal of Mathematics and Music*, *8*, 207–223.

Gabrielsson, A. (2001). Emotion perceived and emotion felt: Same or different? *Musicae Scientiae*, *5*, 123–147. http://doi.org/10.1177/10298649020050s105.

Gabrielsson, A. (2016). The relationship between musical structure and perceived expression. In S. Hallam, I. Cross, & M. Thaut (Eds.), *The Oxford handbook of music psychology* (pp. 1–1). Oxford, UK: Oxford University Press.

Harker, A. (2017). *Software: AHarker externals*. Retrieved from http://www.alexanderjharker.co.uk/software/AHarker_Distribution_v1.0.zip

Huang, S.-L. (2011). Designing utility-based recommender systems for e-commerce: Evaluation of preference-elicitation methods. *Electronic Commerce Research and Applications*, *10*, 398–407. http://doi.org/10.1016/j.elerap.2010.11.003.

Hudson, N. J. (2011). Musical beauty and information compression: Complex to the ear but simple to the mind? *BMC Research Notes*, *4*, 9. http://doi.org/10.1186/1756-0500-4-9.

International Organization for Standardization. (2002). *Information Technology – Multimedia content description interface* (ISO/IEC Standard No. 15938). Retrieved from https://www.iso.org/standard/34228.html

Jannach, D., Zanker, M., Felfernig, A., & Friedrich, G. (2010). *Recommender systems*. Cambridge, UK: Cambridge University Press. http://doi.org/10.1145/2891406.

Jehan, T. (2005). *Creating music by listening*. Unpublished PhD Thesis, Massachusetts Institute of Technology, Cambridge, MA.

Kim, H.-G., Moreau, N., & Sikora, T. (2006). *MPEG-7 audio and beyond*. Chichester, UK: John Wiley & Sons.

Knees, P., & Schedl, M. (2016). *Music similarity and retrieval - an introduction to audio- and web-based strategies. The information retrieval series* (Vol. *36*). Berlin Heidelberg: Springer-Verlag.

Komkhao, M., Lu, J., Li, Z., & Halang, W. A. (2013). Incremental collaborative filtering based on Mahalanobis distance and fuzzy membership for recommender systems. *International Journal of General Systems*, *42*, 41–66. http://doi.org/10.1080/03081079.2012.710437.

Landy, L. (2009) Sound-based music 4 all. In R. T. Dean (Ed.), *The Oxford handbook of computer music* (pp. 518–535). Oxford, UK: Oxford University Press.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A MatLab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data* *analysis, machine learning and applications* (pp. 261–268). Freiburg, Germany: Springer.

Mahalanobis, P. C. (1936). On the Generalized Distance in Statistics. *Proceedings of National Institute of Sciences of India*, *2*, 49–55.

Malt, M., & Jourdan, E. (2008). *Zsa Descriptors: A library for real-time descriptors analysis*. Presented at 5th Sound and Music Computing (SMC) Conference, Berlin, Germany, 31 July – 3 August, pp. 134–137.

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23*, 85–102. http://doi.org/10.1162/014892699559797.

Olsen, K. N., Dean, R. T., & Leung, Y. (2016). What constitutes a phrase in sound-based music? A mixed-methods investigation of perception and acoustics. *PLoS One*, *11*, *e0167643*. http://doi.org/10.1371/journal.pone.0167643.

Olsen, K. N., Dean, R. T., & Stevens, C. J. (2014). A continuous measure of musical engagement contributes to prediction of perceived arousal and valence. *Psychomusicology: Music, Mind and Brain*, *24*, 147–156. http://doi.org/10.1037/pmu0000044.

Olsen, K. N., Dean, R. T., Stevens, C. J., & Bailes, F. (2015). Both acoustic intensity and loudness contribute to time-series models of perceived affect in response to music. *Psychomusicology: Music, Mind and Brain*, *25*, 124–137. http://doi.org/10.1037/pmu0000087.

Puckette, M. S., Apel, T., & Zicarelli, D. (1998). *Real-time audio analysis tools for Pd and MSP*. Proceedings of the International Computer Music Conference (ICMC), 1-6 October, Ann Arbor, MI.

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology*, *100*, 1139–1157. http://doi.org/10.1037/a0022406.

Ricci, F., Rokah R., & Shapira B. (Eds.). (2015). *Recommender systems handbook* (2nd ed.). New York, NY: Springer.

Robjohns, H. (2014). The end of the loudness war? *Sound On Sound*. Retrieved from https://www.soundonsound.com/techniques/end-loudness-war

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*, 1161–1178. http://doi.org/10.1037/h0077714.

Schedl, M. (2017). Investigating country-specific music preferences and music recommendation algorithms with the LFM-1b dataset. *International Journal of Multimedia Information Retrieval*, *6*, 71–84. http://doi.org/10.1007/s13735-017-0118-y.

Schedl, M., Gómez, E., Trent, E. S., Tkalčič, M., Eghbal-Zadeh, H., & Martorell, A. (2018). On the interrelation between listener characteristics and the perception of emotions in classical orchestral music. *IEEE Transactions on Affective Computing*, *9*, 507–525.

Schedl, M., Knees, P., McFee, B., Bogdanov, D., & Kaminskas, M. (2015). Music recommender systems. In F. Ricci, R. Rokach, & B. Shapira (Eds). *Recommender systems handbook* (2nd ed., pp. 453–492). New York, NY: Springer.

Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Journal of SICE*, *48*, 21–32.

Schreiber, H. (2016). *Genre ontology learning - Comparing curated with crowd-sourced ontologies*. Proceedings of the 17th ISMIR, New York City, USA, August 7-11, pp. 400–406. http://doi.org/10.1109/cw.2017.52.

Schubert, E. (1999). Measuring emotion continuously: Validity and reliability of the two-dimensional emotion-space. *Australian Journal of Psychology*, *51*, 154–165. http://doi.org/10.1080/00049539908255353.

Schubert, E. (2004). Modeling perceived emotion with continuous musical features. *Music Perception: An Interdisciplinary Journal*, *21*, 561–585. http://doi.org/10.1525/mp.2004.21.4.561

Schubert, E. (2010). Continuous self-report methods. In P. N. Juslin & J. A. Sloboda (Eds.), *Handbook of music and emotion: Theory, research, applications* (pp. 223–253). New York, NY: Oxford University Press.

Sedain, S., Sanner, S., Brazunias, D., Xie, L., & Christenson, J. (2014). *Social collaborative filtering for cold-start recommendations*. Presented at the 8th ACM Conference, New York, NY, pp. 345–348.

Sturm, B. L. (2013a). Classification accuracy is not enough. *Journal of Intelligent Information Systems*, *41*, 371–406. http://doi.org/10.1007/s10844-013-0250-y.

Sturm, B. L. (2013b). *On music genre classification via compressive sampling* (pp. 1–6). Presented at the Proceedings of IEEE International Conference on Multimedia and Expo (ICME), 15-19 July, San Jose, CA, pp. 1–6. http://doi.org/10.1109/ICME.2013.6607468.

Tkalčič, M., Maleki, N., Pesek, M., Elahi, M., Ricci, F., & Marolt, M. (2019) Prediction of music pairwise preferences from facial expressions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces ACM*, New York, NY, USA, 2019, pp. 150–159. New York, NY: ACM.

Tzanetakis, G., & Cook, P. R. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, *10*, 293–302. http://doi.org/10.1109/TSA.2002.800560.

Weigl, D. M., & Guastavino, C. (2011). *User studies in the music information retrieval literature*. In *Proceedings of the 12th International Society for Music Information Retrieval Conference*, ISMIR, Miami, United States, 24–28 October, 2011, pp. 335–340.

Zhao, S. (2014). *A personalized hybrid music recommender based on empirical estimation of user-timbre preference*. Unpublished Master of Science Thesis, Tampere University of Technology, Finland.