INFORMATION-THEORETIC MEASURES OF MUSIC LISTENING BEHAVIOUR

Daniel Boland, Roderick Murray-Smith
School of Computing Science, University of Glasgow, United Kingdom
daniel@dcs.gla.ac.uk; roderick.murray-smith@glasgow.ac.uk

ABSTRACT

We present an information-theoretic approach to the measurement of users' music listening behaviour and selection of music features. Existing ethnographic studies of music use have guided the design of music retrieval systems however are typically qualitative and exploratory in nature. We introduce the *SPUD* dataset, comprising $10,000$ handmade playlists, with user and audio stream metadata. With this, we illustrate the use of entropy for analysing music listening behaviour, e.g. identifying when a user changed music retrieval system. We then develop an approach to identifying music features that reflect users' criteria for playlist curation, rejecting features that are independent of user behaviour. The dataset and the code used to produce it are made available. The techniques described support a quantitative yet user-centred approach to the evaluation of music features and retrieval systems, without assuming objective ground truth labels.

1. INTRODUCTION

Understanding how users interact with music retrieval systems is of fundamental importance to the field of Music Information Retrieval (MIR). The design and evaluation of such systems is conditioned upon assumptions about users, their listening behaviours and their interpretation of music. While user studies have offered guidance to the field thus far, they are mostly exploratory and qualitative [20]. The availability of quantitative metrics would support the rapid evaluation and optimisation of music retrieval. In this work, we develop an information-theoretic approach to measuring users' music listening behaviour, with a view to informing the development of music retrieval systems.

To demonstrate the use of these measures, we compiled 'Streamable Playlists with User Data' *(SPUD)* – a dataset comprising $10,000$ playlists from Last.fm [1] produced by 3351 users, with track metadata including audio streams from Spotify. [2] We combine the dataset with the mood and genre classification of Syntonetic's Moodagent, [3] yielding a range of intuitive music features to serve as examples.

We identify the entropy of music features as a metric for characterising music listening behaviour. This measure can be used to produce time-series analyses of user behaviour, allowing for the identification of events where this behaviour changed. In a case study, the date when a user adopted a different music retrieval system is detected. These detailed analyses of listening behaviour can support user studies or provide implicit relevance feedback to music retrieval. More broad analyses are performed across the $10,000$ playlists. A Mutual Information based feature selection algorithm is employed to identify music features relevant to how users create playlists. This user-centred feature selection can sanity-check the choice of features in MIR. The information-theoretic approach introduced here is applicable to any discretisable feature set and distinct in being based solely upon actual user behaviour rather than assumed ground-truth. With the techniques described here, MIR researchers can perform quantitative yet user-centred evaluations of their music features and retrieval systems.

1.1 Understanding Users

User studies have provided insights about user behaviour in retrieving and listening to music and highlighted the lack of consideration in MIR about actual user needs. In 2003, Cunningham et al. bemoaned that development of music retrieval systems relied on "anecdotal evidence of user needs, intuitive feelings for user information seeking behavior, and a priori assumptions of typical usage scenarios" [5]. While the number of user studies has grown, the situation has been slow to improve. A review conducted a decade later noted that approaches to system evaluation still ignore the findings of user studies [12]. This issue is stated more strongly by Schedl and Flexer, describing systems-centric evaluations that "completely ignore user context and user properties, even though they clearly influence the result" [15]. Even systems-centric work, such as the development of music classifiers, must consider the user-specific nature of MIR. Downie termed this the multi-experiential challenge, and noted that "Music ultimately exists in the mind of its perceiver" [6]. Despite all of this, the assumption of an objective ground truth for music genre, mood etc. is common [4], with evaluations focusing on these rather than considering users. It is clear that much work remains in placing the user at the centre of MIR.

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

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

[3] http://www.moodagent.com Last accessed: 30/04/14

1.2 Evaluation in MIR

The lack of robust evaluations in the field of MIR was identified by Futrelle and Downie as early as 2003 [8]. They noted the lack of any standardised evaluations and in particular that MIR research commonly had an "emphasis on basic research over application to, and involvement with, users." In an effort to address these failings, the Music Information Retrieval Evaluation Exchange (MIREX) was established [7]. MIREX provides a standardised framework of evaluation for a range of MIR problems using common metrics and datasets, and acts as the benchmark for the field. While the focus on this benchmark has done a great deal towards the standardisation of evaluations, it has distracted research from evaluations with real users.

A large amount of evaluative work in MIR focuses on the performance of classifiers, typically of mood or genre classes. A thorough treatment of the typical approaches to evaluation and their shortcomings is given by Sturm [17]. We note that virtually all such evaluations seek to circumvent involving users, instead relying on a 'ground truth' which is assumed to be objective. An example of a widely used ground truth dataset is *GTZAN*, a small collection of music with the author's genre annotations. Even were the objectivity of such annotations to be assumed, such datasets can be subject to confounding factors and mislabellings as shown by Sturm [16]. Schedl et al. also observe that MIREX evaluations involve assessors' own subjective annotations as ground truth [15].

1.3 User-Centred Approaches

There remains a need for robust, standardised evaluations featuring actual users of MIR systems, with growing calls for a more user-centric approach. Schedl and Flexer made the broad case for "putting the user in the center of music information retrieval", concerning not only user-centred development but also the need for evaluative experiments which control independent variables that may affect dependent variables [14]. We note that there is, in particular, a need for quantitative dependent variables for user-centred evaluations. For limited tasks such as audio similarity or genre classification, existing dependent variables may be sufficient. If the field of MIR is to concern itself with the development of complete music retrieval systems, their interfaces, interaction techniques, and the needs of a variety of users, then additional metrics are required. Within the field of HCI it is typical to use qualitative methods such as the think-aloud protocol [9] or Likert-scale questionnaires such as the NASA Task Load Index (TLX) [10].

Given that the purpose of a Music Retrieval system is to support the user's retrieval of music, a dependent variable to measure this ability is desirable. Such a measure cannot be acquired independently of users – the definition of musical relevance is itself subjective. Users now have access to 'Big Music' – online collections with millions of songs, yet it is unclear how to evaluate their ability to retrieve this music. The information-theoretic methodology introduced in this work aims to quantify the exploration, diversity and underlying mental models of users' music retrieval.

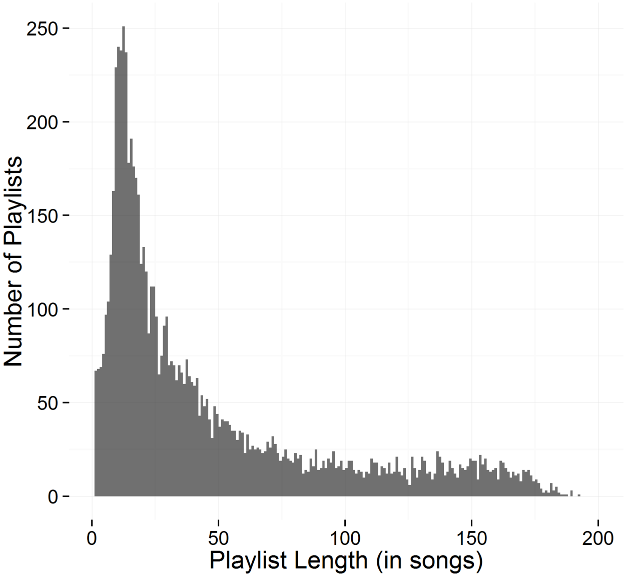

Figure 1. Distribution of playlist lengths within the *SPUD* dataset. The distribution peaks around a playlist length of 12 songs. There is a long tail of lengthy playlists.

2. THE *SPUD* DATASET

The *SPUD* dataset of $10,000$ playlists was produced by scraping from Last.fm users who were active throughout March and April, 2014. The tracks for each playlist are also associated with a Spotify stream, with scraped metadata, such as artist, popularity, duration etc. The number of unique tracks in the dataset is $271,389$ from 3351 users. The distribution of playlist lengths is shown in Figure 1. We augment the dataset with proprietary mood and genre features produced by Syntonetic's Moodagent. We do this to provide high-level and intuitive features which can be used as examples to illustrate the techniques being discussed. It is clear that many issues remain with genre and mood classification [18] and the results in this work should be interpreted with this in mind. Our aim in this work is not to identify which features are best for music classification but to contribute an approach for gaining an additional perspective on music features. Another dataset of playlists *AOTM-2011* is published [13] however the authors only give fragments of playlists where songs are also present in the Million Song Dataset (*MSD*) [1]. The *MSD* provides music features for a million songs but only a small fraction of songs in *AOTM-2011* were matched in *MSD*. Our *SPUD* dataset is distinct in maintaining complete playlists and having time-series data of songs listened to.

3. MEASURING MUSIC LISTENING BEHAVIOUR

When evaluating a music retrieval system, or performing a user study, it would be useful to quantify the music-listening behaviour of users. Studying this behaviour over time would enable the identification of how different music retrieval systems influence user behaviour. Quantifying listening behaviour would also provide a dependent variable for use in MIR evaluations. We introduce entropy as one such quantitative measure, capturing how a user's music-listening relates to the music features of their songs.

3.1 Entropy

For each song being played by a user, the value of a given music feature can be taken as a random variable X. The entropy $H(X)$ of this variable indicates the uncertainty about the value of that feature over multiple songs in a listening session. This entropy measure gives a scale from a feature's value never changing, through to every level of the feature being equally likely. The more a user constrains their music selection by a particular feature, e.g. mood or album, then the lower the entropy is over those features. The entropy for a feature is defined as:

$$H(X) = -\sum_{x \in X} p(x) \log_2[p(x)] \,, \tag{1}$$

where x is every possible level of the feature X and the distribution $p(x)$ is estimated from the songs in the listening session. The resulting entropy value is measured in bits, though can be normalised by dividing by the maximum entropy $\log_2[|X|]$. Estimating entropy in this way can be done for any set of features, though requires that they are discretised to an appropriate number of levels.

For example, if a music listening session is dominated by songs of a particular tempo, the distribution over values of a TEMPO feature would be very biased. The entropy $H(\text{TEMPO})$ would thus be very low. Conversely, if users used shuffle or listened to music irrespective of tempo, then the entropy $H(\text{TEMPO})$ would tend towards the average entropy of the whole collection.

3.2 Applying a Window Function

Many research questions regarding a user's music listening behaviour concern the change in that behaviour over time. An evaluation of a music retrieval interface might hypothesise that users will be empowered to explore a more diverse range of music. Musicologists may be interested to study how listening behaviour has changed over time and which events precede such changes. It is thus of interest to extend Eqn (1) to define a measure of entropy which is also a function of time:

$$H(X, t) = H(w(X, t)) \,, \tag{2}$$

where $w(X, t)$ is a window function taking n samples of X around time t. In this paper we use a rectangular window function with $n = 20$, assuming that most albums will have fewer tracks than this. The entropy at any given point is limited to the maximum possible $H(X, t) = \log_2[n]$ i.e. where each of the n points has a unique value.

An example of the change in entropy for a music feature over time is shown in Figure 2. In this case $H(\text{ALBUM})$ is shown as this will be 0 for album-based listening and at maximum for exploratory or radio-like listening. It is important to note that while trends in mean entropy can be identified, the entropy of music listening is itself quite a noisy signal – it is unlikely that a user will maintain a single music-listening behaviour over a large period of time. Periods of album listening (low or zero entropy) can be seen through the time-series, even after the overall trend is towards shuffle or radio-like music listening.

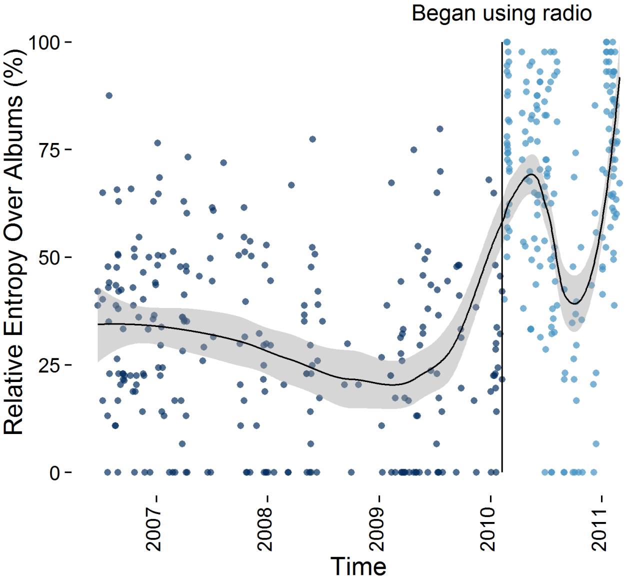

Figure 2. Windowed entropy over albums shows a user's album-based music listening over time. Each point represents 20 track plays. The black line depicts mean entropy, calculated using locally weighted regression [3] with 95% CI of the mean shaded. A changepoint is detected around Feb. 2010, as the user began using online radio (light blue)

3.3 Changepoints in Music Retrieval

Having produced a time-series analysis of music-listening behaviour, we are now able to identify events which caused changes in this behaviour. In order to identify changepoints in the listening history, we apply the 'Pruned Exact Linear Time' (PELT) algorithm [11]. The time-series is partitioned in a way that reduces a cost function of changes in the mean and variance of the entropy. Changepoints can be of use in user studies, for example in Figure 2, the user explained in an interview that the detected changepoint occurred when they switched to using online radio. There is a brief return to album-based listening after the changepoint – users' music retrieval behaviour can be a mixture of different retrieval models. Changepoint detection can also be a user-centred dependent variable in evaluating music retrieval interfaces i.e. does listening behaviour change as the interface changes? Further examples of user studies are available with the *SPUD* dataset.

3.4 Identifying Listening Style

The style of music retrieval that the user is engaging in can be inferred using the entropy measures. Where the entropy for a given music feature is low, the user's listening behaviour can be characterised by that feature i.e. we can be certain about that feature's level. Alternately, where a feature has high entropy, then the user is not 'using' that feature in their retrieval. When a user opts to use shuffle-based playback i.e. the random selection of tracks, there is the unique case that entropy across all features will tend towards the maximum. In many cases, feature entropies have high covariance, e.g. songs on an album will have the same artist and similar features. We did not include other features in Figure 2 as the same pattern was apparent.

4. SELECTING FEATURES FROM PLAYLISTS

Identifying which music features best describe a range of playlists is not only useful for playlist recommendation, but also provides an insight into how users organise and think about music. Music recommendation and playlist generation typically work on the basis of genre, mood and popularity, and we investigate which of these features is supported by actual user behaviour. As existing retrieval systems are based upon these features, there is a potential 'chicken-and-egg' effect where the features which best describe user playlists are those which users are currently exposed to in existing retrieval interfaces.

4.1 Mutual Information

Information-theoretic measures can be used to identify to what degree a feature shares information with class labels. For a feature X and a class label Y, the mutual information $I(X;Y)$ between these two can be given as:

$$I(X;Y) = H(X) - H(X \mid Y), \tag{3}$$

that is, the entropy of the feature $H(X)$ minus the entropy of that feature if the class is known $H(X \mid Y)$. By taking membership of playlists as a class label, we can determine how much we can know about a song's features if we know what playlist it is in. When using mutual information to compare clusterings in this way, care must be taken to account for random chance mutual information [19]. We adapt this approach to focus on how much the feature entropy is reduced, and normalise accordingly:

$$AMI(X;Y) = \frac{I(X;Y) - E[I(X;Y)]}{H(X) - E[I(X;Y)]}, \tag{4}$$

where $AMI(X;Y)$ is the adjusted mutual information and $E[I(X;Y)]$ is the expectation of the mutual information i.e. due to random chance. The AMI gives a normalised measure of how much of the feature's entropy is explained by the playlist. When $AMI = 1$, the feature level is known exactly if the playlist is known, when $AMI = 0$, nothing about the feature is known if the playlist is known.

4.2 Linking Features to Playlists

We analysed the AMI between the $10,000$ playlists in the *SPUD* dataset and a variety of high level music features. The ranking of some of these features is given in Figure 3. Our aim is only to illustrate this approach, as any results are only as reliable as the underlying features. With this in mind, the features ROCK and ANGRY had the most uncertainty explained by playlist membership. While the values may seem small, they are calculated over many playlists, which may combine moods, genres and other criteria. As these features change most between playlists (rather than within them), they are the most useful for characterising the differences between playlists. The DURATION feature ranked higher than expected, further investigation revealed playlists that combined lengthy DJ mixes. It is perhaps unsurprising that playlists were not well characterised by whether they included WORLD music.

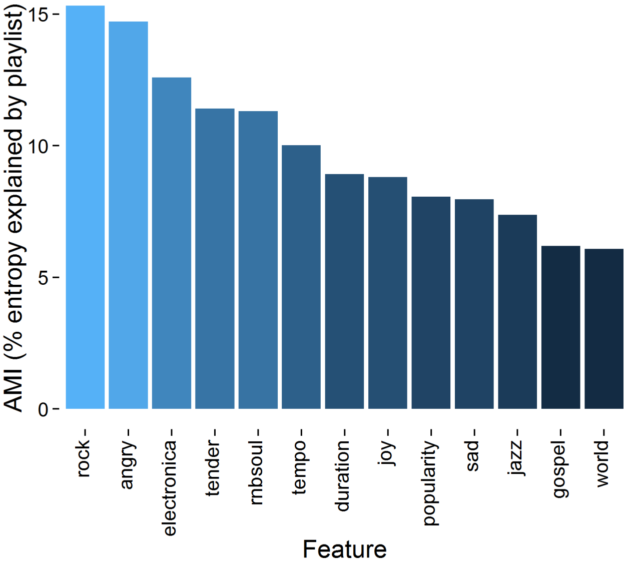

Figure 3. Features are ranked by their Adjusted Mutual Information with playlist membership. Playlists are distinguished more by whether they contain ROCK or ANGRY music than by whether they contain POPULAR or WORLD.

It is of interest that TEMPO was not one of the highest ranked features, illustrating the style of insights available when using this approach. Further investigation is required to determine whether playlists are not based on tempo as much as is often asumed or if this result is due to the peculiarities of the proprietary perceptual tempo detection.

4.3 Feature Selection

Features can be selected using information-theoretic measures, with a rigorous treatment of the field given by Brown et al. [2]. They define a unifying framework within which to discuss methods for selecting a subset of features using mutual information. This is done by defining a J criterion for a feature:

$$J(f_n) = I(f_n; C \mid S). \tag{5}$$

This gives a measure of how much information the feature shares with playlists given some previously selected features, and can be used as a greedy feature selection algorithm. Intuitively, features should be selected that are relevant to the classes but that are also not redundant with regard to previously selected features. A range of estimators for $I(f_n; C \mid S)$ are discussed in [2].

As a demonstration of the feature selection approach we have described, we apply it to the features depicted in Figure 3, selecting features to minimise redundancy. The selected subset of features in rank order is: ROCK, DURATION, POPULARITY, TENDER and JOY. It is notable that ANGRY had an AMI that was almost the same as ROCK, but it is redundant if ROCK is included. Unsurprisingly, the second feature selected is from a different source than the first – the duration information from Spotify adds to that used to produce the Syntonetic mood and genre features. Reducing redundancy in the selected features in this way yields a very different ordering, though one that may give a clearer insight into the factors behind playlist construction.

5. DISCUSSION

While we reiterate that this work only uses a specific set of music features and user base, we consider our results to be encouraging. It is clear that the use of entropy can provide a detailed time-series analysis of user behaviour and could prove a valuable tool for MIR evaluation. Similarly, the use of adjusted mutual information allows MIR researchers to directly link work on acquiring music features to the ways in which users interact with music. In this section we consider how the information-theoretic techniques described in this work can inform the field of MIR.

5.1 User-Centred Feature Selection

The feature selection shown in this paper is done directly from the user data. In contrast, feature selection is usually performed using classifier wrappers with ground truth class labels such as genre. The use of genre is based on the assumption that it would support the way users currently organise music and features are selected based on these labels. This has lead to issues including classifiers being trained on factors that are confounded with these labels and that are not of relevance to genre or users [18]. Our approach selects features independently of the choice of classifier, in what is termed a 'filter' approach. The benefit of doing this is that a wide range of features can be quickly filtered at relatively little computational expense. While the classifier 'wrapper' approach may achieve greater performance, it is more computationally expensive and more likely to suffer from overfitting.

The key benefit of filtering features based on user behaviour is that it provides a perspective on music features that is free from assumptions about users and music ground truth. This user-centred perspective provides a sanity-check for music features and classification – if a feature does not reflect the ways in which users organise their music, then how useful is it for music retrieval?

5.2 When To Learn

The information-theoretic measures presented offer an implicit relevance feedback for music retrieval. While we have considered the entropy of features as reflecting user behaviour, this behaviour is conditioned upon the existing music retrieval interfaces being used. For example, after issuing a query and receiving results, the user selects relevant songs from those results. If the entropy of a feature for those selected songs is small relative to the result set, then this feature is implicitly relevant to the retrieval.

The identification of shuffle and explorative behaviour provides some context for this implicit relevance feedback. Music which is listened to in a seemingly random fashion may represent an absent or disengaged user, adding noise to attempts to weight recommender systems or build a user profile. At the very least, where entropy is high across all features, then those features do not reflect the user's mental model for their music retrieval. The detection of shuffle or high-entropy listening states thus provides a useful data hygiene measure when interpreting listening data.

5.3 Engagement

The entropy measures capture how much each feature is being 'controlled' by the user when selecting their music. We have shown that it spans a scale from a user choosing to listen to something specific to the user yielding control to radio or shuffle. Considering entropy over many features in this way gives a high-dimensional vector representing the user's engagement with music. Different styles of music retrieval occupy different points in this space, commonly the two extremes of listening to a specific album or just shuffling. There is an opportunity for music retrieval that has the flexibility to support users engaging and applying control over music features only insofar as they desire to. An example of this would be a shuffle mode that allowed users to bias it to varying degrees, or to some extent, the feedback mechanism in recommender systems.

5.4 Open Source

The SPUD dataset is made available for download at:
`http://www.dcs.gla.ac.uk/~daniel/spud/`
Example R scripts for importing data from *SPUD* and producing the analyses and plots in this paper are included. The code used to scrape this dataset is available under the MIT open source license, and can be accessed at:
`http://www.github.com/dcboland/`

The MoodAgent features are commercially sensitive, thus not included in the *SPUD* dataset. At present, industry is far better placed to provide such large scale analyses of music data than academia. Even with user data and the required computational power, large-scale music analyses require licensing arrangements with content providers, presenting a serious challenge to academic MIR research. Our adoption of commercially provided features has allowed us to demonstrate our information-theoretic approach, and we distribute the audio stream links, however it is unlikely that many MIR researchers will have the resources to replicate all of these large scale analyses. The CoSound[4] project is an example of industry collaborating with academic research and state bodies to navigate the complex issues of music licensing and large-scale analysis.

6. CONCLUSION

This work introduces an information-theoretic approach to the study of users' music listening behaviour. The case is made for a more user-focused yet quantitative approach to evaluation in MIR. We described the use of entropy to produce time-series analyses of user behaviour, and showed how changes in music-listening style can be detected. An example is given where a user started using online radio, having higher entropy in their listening. We introduced the use of adjusted mutual information to establish which music features are linked to playlist organisation. These techniques provide a quantitative approach to user studies and ground feature selection in user behaviour, contributing tools to support the user-centred future of MIR.

[4] `http://www.cosound.dk/` Last accessed: 30/04/14

ACKNOWLEDGEMENTS

This work was supported in part by Bang & Olufsen and the Danish Council for Strategic Research of the Danish Agency for Science Technology and Innovation under the CoSound project, case number 11-115328. This publication only reflects the authors' views.

7. REFERENCES

[1] T Bertin-Mahieux, D. P Ellis, B Whitman, and P Lamere. The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval*, Miami, Florida, 2011.

[2] G Brown, A Pocock, M.-J Zhao, and M Luján. Conditional likelihood maximisation: a unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012.

[3] W. S Cleveland and S. J Devlin. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403):596–610, 1988.

[4] A Craft and G Wiggins. How many beans make five? the consensus problem in music-genre classification and a new evaluation method for single-genre categorisation systems. In *Proceedings of the 8th International Conference on Music Information Retrieval*, Vienna, Austria, 2007.

[5] S. J Cunningham, N Reeves, and M Britland. An ethnographic study of music information seeking: implications for the design of a music digital library. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, Houston, Texas, 2003.

[6] J. S Downie. Music Information Retrieval. *Annual Review of Information Science and Technology*, 37(1):295–340, January 2003.

[7] J. S Downie. The Music Information Retrieval Evaluation eXchange (MIREX). *D-Lib Magazine*, 12(12):795–825, 2006.

[8] J Futrelle and J. S Downie. Interdisciplinary Research Issues in Music Information Retrieval: ISMIR 2000 - 2002. *Journal of New Music Research*, 32(2):121–131, 2003.

[9] J. D Gould and C Lewis. Designing for usability: key principles and what designers think. *Communications of the ACM*, 28(3):300–311, 1985.

[10] S. G Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting*, San Francisco, California, 2006.

[11] R Killick, P Fearnhead, and I. A Eckley. Optimal Detection of Changepoints With a Linear Computational Cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

[12] J. H Lee and S. J Cunningham. The Impact (or Non-impact) of User Studies in Music Information Retrieval. In *Proceedings of the 13th International Conference for Music Information Retrieval*, Porto, Portugal, 2012.

[13] B McFee and G Lanckriet. Hypergraph models of playlist dialects. In *Proceedings of the 13th International Conference for Music Information Retrieval*, Porto, Portugal, 2012.

[14] M Schedl and A Flexer. Putting the User in the Center of Music Information Retrieval. In *Proceedings of the 13th International Conference on Music Information Retrieval*, Porto, Portugal, 2012.

[15] M Schedl, A Flexer, and J Urbano. The neglected user in music information retrieval research. *Journal of Intelligent Information Systems*, 41(3):523–539, 2013.

[16] B. L Sturm. An Analysis of the GTZAN Music Genre Dataset. In *Proceedings of the 2nd International ACM Workshop on Music Information Retrieval with User-centered and Multimodal Strategies*, MIRUM '12, New York, USA, 2012.

[17] B. L Sturm. Classification accuracy is not enough. *Journal of Intelligent Information Systems*, 41(3):371–406, 2013.

[18] B. L Sturm. A simple method to determine if a music information retrieval system is a horse. *IEEE Transactions on Multimedia*, 2014.

[19] N. X Vinh, J Epps, and J Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

[20] D. M Weigl and C Guastavino. User studies in the music information retrieval literature. In *Proceedings of the 12th International Conference for Music Information Retrieval*, Miami, Florida, 2011.