Investigating the Impact of Audio States & Transitions for Track Sequencing in Music Streaming Sessions

Aaron Ng
University College London
London, United Kingdom
aaron.ng.19@ucl.ac.uk

Rishabh Mehrotra
Spotify
London, United Kingdom
rishabhm@spotify.com

ABSTRACT

Music streaming is inherently sequential in nature, with track sequence information playing a key role in user satisfaction with recommended music. In this work, we investigate the role audio characteristics of music content play in understanding music streaming sessions. Focusing on 18 audio attributes (e.g. dancability, acousticness, energy), we formulate audio transitioning in a session as a multiple changepoint detection problem, and extract latent states of different audio attributes within each session. Based on insights from large scale music streaming data from a popular music streaming platform, we investigate questions around the extent to which audio characteristics fluctuate within streaming sessions, the heterogeneity across different audio attributes and their impact on user satisfaction. Furthermore, we demonstrate the promise of such audio-based characterizing of sessions in better sequencing tracks in a session, and highlight the potential gains in user satisfaction on offer. We discuss implications on the design of track sequencing models, and identify important prediction tasks to further research on the topic.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; **Music retrieval**.

KEYWORDS

Recommender Systems, Hidden Markov Model, Track Sequencing, Audio States

ACM Reference Format:
Aaron Ng and Rishabh Mehrotra. 2020. Investigating the Impact of Audio States & Transitions for Track Sequencing in Music Streaming Sessions. In *Fourteenth ACM Conference on Recommender Systems (RecSys '20), September 22–26, 2020, Virtual Event, Brazil.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3383313.3418493

1 INTRODUCTION

Music streaming sessions are increasingly being shaped by algorithmically generated sequences, aimed at creating engaging, seamless, and cohesive listening experiences for users. An ideal recommender system would not only recommend music that matches user preferences, but sequence musical tracks in a way so as to make the music "*flow smoothly*" from one song to the next. Sequencing tracks has been regarded as "*more of an art than a science*" [5], with the ordering of tracks and the quality of the transitions between them being fundamentally linked: it can be very difficult to create an enjoyable transition between songs that significantly differs in style, tempo, or key.

Audio attributes of music content, e.g. danceability, energy, instrumentalness, help quantify and describe the acoustic characteristics of the track. Such acoustic characteristics provide complimentary information to the more commonly used organizational (i.e. which tracks are present in a playlist) and consumption (i.e. which users consume which content) information, and can be used to understand how transitions in the audio attributes impact user perception of recommended music. Indeed, transitioning between a slow, smooth jazz piece and a high energy, fast electronic track, for example, will likely feel awkward or unnatural and create an abrupt change in the listening experience.

In this work, we focus on characterizing such transitions in music streaming sessions and investigate their impact on user satisfaction, based on a large scale analysis of music streaming sessions from over 1 million users. We propose to treat each sequence of audio feature values in a session as a time series and formulate audio transitions as a changepoint detection task: events happen at a rate that changes over time, driven by sudden shifts in the (unobserved) state of some system or process generating the data. Within a session, audio characteristics could stay relatively constant (i.e. in a single state), or vary drastically across a session (i.e. fluctuate between multiple states). We model the latent states of each sequence using Hidden Markov Models and propose a multiple changepoint detection technique.

Quantifying such transitions across a session enables us to analyze streaming sessions at scale to answer a number of key research questions. First, we begin by investigating how much and how common do audio characteristics change in listening sessions. Second, we investigate how these variations differ across the 18 different audio features. Third, we analyze how such variations are related to user satisfaction metrics. Finally, we consider the task of track re-ranking within sessions, to understand how we could use such variations in audio characteristics to better sequence tracks.

We observe that audio attributes do change across tracks in listening sessions, with over 95% sessions having at least one audio feature with two or more states. Such fluctuations are fairly common across all audio attributes. Moreover, we observe that satisfying sessions have fewer state transitions than dissatisfying sessions, with a high correlation between state transitions and track

RecSys '20, September 22–26, 2020, Virtual Event, Brazil

ACM ISBN 978-1-4503-7583-2/20/09.
https://doi.org/10.1145/3383313.3418493

skips. Lastly, we highlight that leveraging such information about state transitions holds promise, as it can help us improve key user satisfaction metrics by over 10%. Our preliminary findings have implications on the design of sequential recommendation techniques, and we identify important prediction tasks as future work.

2 RELATED WORK

A variety of features can be extracted from audio signals in tracks, such as rhythmic, timbre or harmonic characteristics. These features are used for multiple applications of Music Information Retrieval (MIR), including automatic genre classification and music recommendation [14]. The idea of using the extracted features to automatically generate listening experiences that flow seamlessly between tracks by sequencing musical styles and matching tempos have been explored in several works [2, 3, 8, 10, 15]. Traditionally, multi-faceted scoring methods combining track co-occurrence patterns, metadata and user preferences have been used in many systems [7, 8, 11]. In order to recommend similar tracks, one also need to develop measures of similarity between tracks, which have led to works examining both acoustic and subjective approaches [1]. Other works include automatically identifying genres via machine learning techniques [6], and through discovering latent structures using Gaussian Mixture Models [13] or Convolutional Neural Networks [12].

3 TRANSITIONS IN AUDIO CHARACTERISTICS ACROSS SESSIONS

We hypothesize that understanding how audio properties of tracks change within a session, and their interplay with user satisfaction can enable us to develop better track sequencing models. To this end, we focus on characterizing transitions in audio attributes of tracks across a large number of user sessions, and present a model to extract "*states*" which an audio attribute can belong to in a music session, and leverage the extracted states information to characterize music streaming sessions. We begin by stating key research questions we wish to address (Section 3.1), describe data context (Section 3.2), present the proposed state extraction model (Section 3.3), and present findings in Section 4.

3.1 Key Research Questions

To better understand the interplay between variations in audio characteristics of songs within a session, and user satisfaction, we investigate the following research questions:

- RQ1 *How varied are audio properties of tracks within a session?* We investigate the extent of variations in audio characteristics within sessions and highlight how common audio fluctuations are (Section 4.1).
- RQ2 *How do variations in audio characteristics differ across the different audio attributes?* We investigate attribute level heterogeneity and highlight audio attributes which have higher than average, and lower than average variations. We also relate such differences to the general distribution of the attribute (Section 4.2).
- RQ3 *Are variations in audio properties related to user satisfaction?* To gauge the interplay between such variations and user satisfaction, we investigate the correlation between states and state transitions with song skips (Section 4.3).
- RQ4 *Can insights about audio states help in track sequencing for improved user satisfaction?* We present preliminary results on track re-ranking based on audio states information and highlight the promise of such methods for developing better track sequencing methods (Section 4.4).

3.2 Data Context

In order to answer the above questions, we use the Music Streaming Sessions Dataset (MSSD) [4], which consists of 160 million listening sessions gathered over an 8 week period on Spotify and associated user interactions, audio features and metadata. Each session vary between 10 to 20 tracks, where sessions longer than 20 tracks are cut off. User interaction features provided include different types of skips, forwards, pauses, etc. Each track also comes with several metadata (popularity, duration, release year, etc) and various audio attributes (e.g. liveness, key, energy). The full list and definitions of the audio attributes can be found in [4].

Some preliminary analysis of the audio feature values can be found in Figure 1a. Looking at the distributions of audio characteristics, we notice that different features have different distributions. Some features (e.g. flatness, instrumentalness) have bimodal / heavily skewed distributions, which are likely to impact results when we analyze variations in feature values across different sessions. When we compute the Pearson correlation between audio characteristics, we also found that most features are not correlated with each other.

3.3 State Extraction via Change Point Detection

In order to perform state extraction and measure state transitions across sessions, we analyzed 50,000 listening sessions from MSSD. Each sequence of audio feature values in a session is treated as a time series (i.e. 18 sequences per session), and change point detection is used on each sequence to detect changes and extract states.

We model the latent states of each sequence using a Hidden Markov Model (HMM), using k discrete latent states $z_t \in \{1, 2, \dots, k\}$. To model movement between states, we define a simple transition model using a categorical distribution, such that the probability of staying in the previous state or transiting to another state is uniform $z_t|z_{t-1} \sim \mathrm{Cat}(\{\frac{1}{k}, \dots, \frac{1}{k}\})$. The emission probabilities are defined using a normal distribution $x_t \sim \mathcal{N}(\mu_{z_t}, \sigma^2_{\text{feat}})$ where μ_{z_t} is the mean of the trainable latent states, and σ^2_{feat} is the average standard deviation of the corresponding audio feature across all sessions. We also include a prior on the latent states $z_t \sim \mathcal{N}(\mu_{\text{feat}}, \sigma^2_{\text{feat}})$, using the average mean and standard deviation of the corresponding audio feature across all sessions.

To train the model, we run an Adam [9] optimizer with a learning rate of 0.1 to compute the Maximum a Posteriori (MAP) fit to the observed values:

$$\boldsymbol{\mu}_{MAP} = \mathrm{argmax}_{\boldsymbol{\mu}}\ p(z_{1:T}|x_{1:T}) \tag{1}$$

Once the model is fitted, we compute the marginal posterior distribution $p(Z_t = z_t|x_{1:T})$ over the states for each timestep using the forward-backward algorithm and assign the most likely state

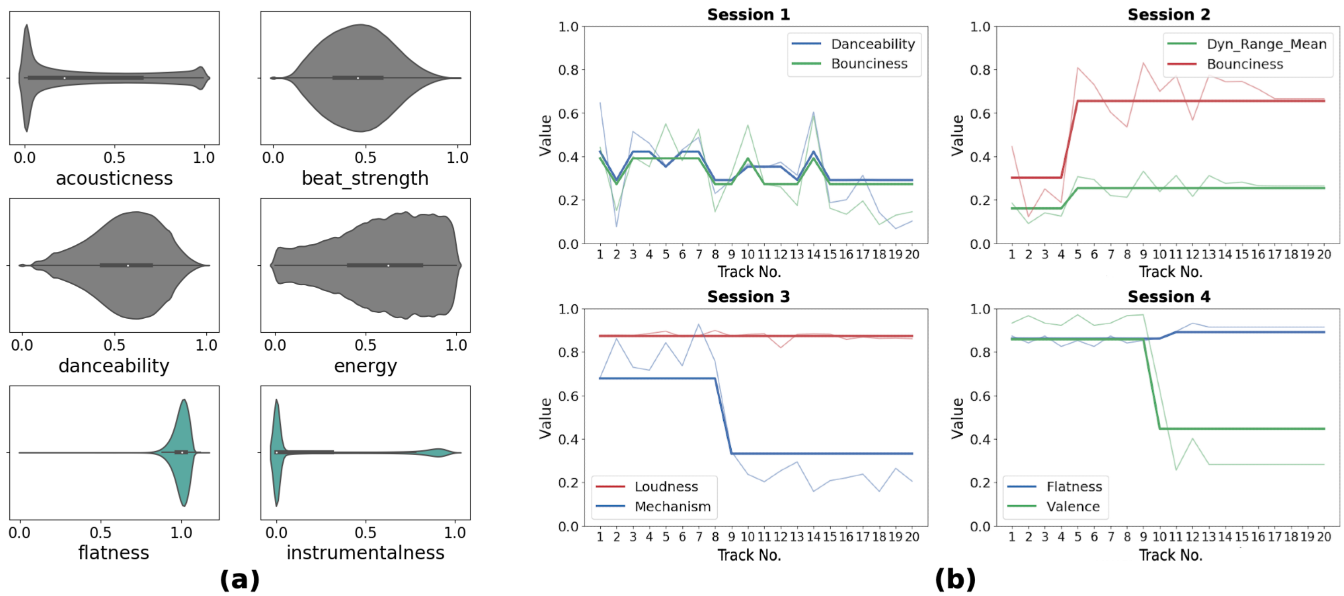

Figure 1: (a) Distribution of values for selected audio features in MSSD. Green plots have heavily skewed distributions compared to the rest. (b) Raw audio feature values (Normal line) and their corresponding states (Bold line) for 2 selected audio features in 4 sample listening sessions.

to each timestep:

$$z_t^* = \text{argmax}_{z_t}\ p(z_t|x_{1:T}) \tag{2}$$

In our experiments, we set $k = 10$, but states with similar means are merged together after inference. States with fewer than 3 tracks are also filtered out, since states with only 1 or 2 tracks are not meaningful. Since each session has a maximum of 20 tracks and each state should be meaningful, we will not have more than 10 latent states, thus k was set to 10. Examples of raw feature values and the inferred states generated by the HMM can be seen in Figure 1b.

Our state extraction mechanism captures both global and local variations of audio characteristics in listening sessions. User preferences would vary between different sessions, and our mechanism captures local variations as we fit a HMM per audio feature per session. However, global variations are also taken into account as the local HMM parameters are derived from global distributions of the audio feature values. Lastly, we define a state or session as satisfying (SAT) when the average number of skips ≤ 0.25, and dissatisfying (DSAT) when the average number of skips ≥ 0.75.

4 FINDINGS

4.1 RQ1: How varied are audio properties of tracks within a session?

4.1.1 Do audio characteristics change across tracks in listening sessions? Audio characteristics do change across tracks in listening sessions once we analyze the states extracted for each feature in each session. Over 95% of the sessions have at least 1 feature with ≥2 states, as seen in Figure 2a, which indicate the majority of sessions have at least 1 feature with variations. However, over 55% of sessions have fewer than 5 features with ≥2 states, which implies that most sessions are characterized only by a handful of features, rather than all the features.

4.1.2 Are fluctuations in audio characteristics common? Fluctuations in audio characteristics are also fairly common, as seen in Figure 1b, where 2 selected audio features and their corresponding states extracted using our model are plotted for 4 sample sessions. We can observe that the number of states vary across different features in different sessions, with many features exhibiting 1 state (Loudness in Session 3), 2 states (Sessions 2 & 4), and even 3 states (Danceability in Session 1). The number of transitions also vary, with some features transitioning only once (Sessions 2 & 4), while some features exhibit many transitions (Session 1).

4.2 RQ2: How do variations in audio characteristics differ across the different audio attributes?

Most features have at least 20% of the sessions with ≥2 states, as shown in the first 5 features of Figure 2c. However, a few features (e.g. instrumentalness, key, time_signature) have <10% of the sessions with ≥2 states. These coincides with their bimodal / highly skewed distributions.

When measuring transitions between states, we also observed that different features exhibit different number of state transitions, as shown in Figure 2d. By measuring the ratio of state transitions to all track transitions, features such as 'acousticness' exhibit a lot more fluctuations between states than others like 'key'. Thus, some features have more variations than others, both in terms of number of states and transitions.

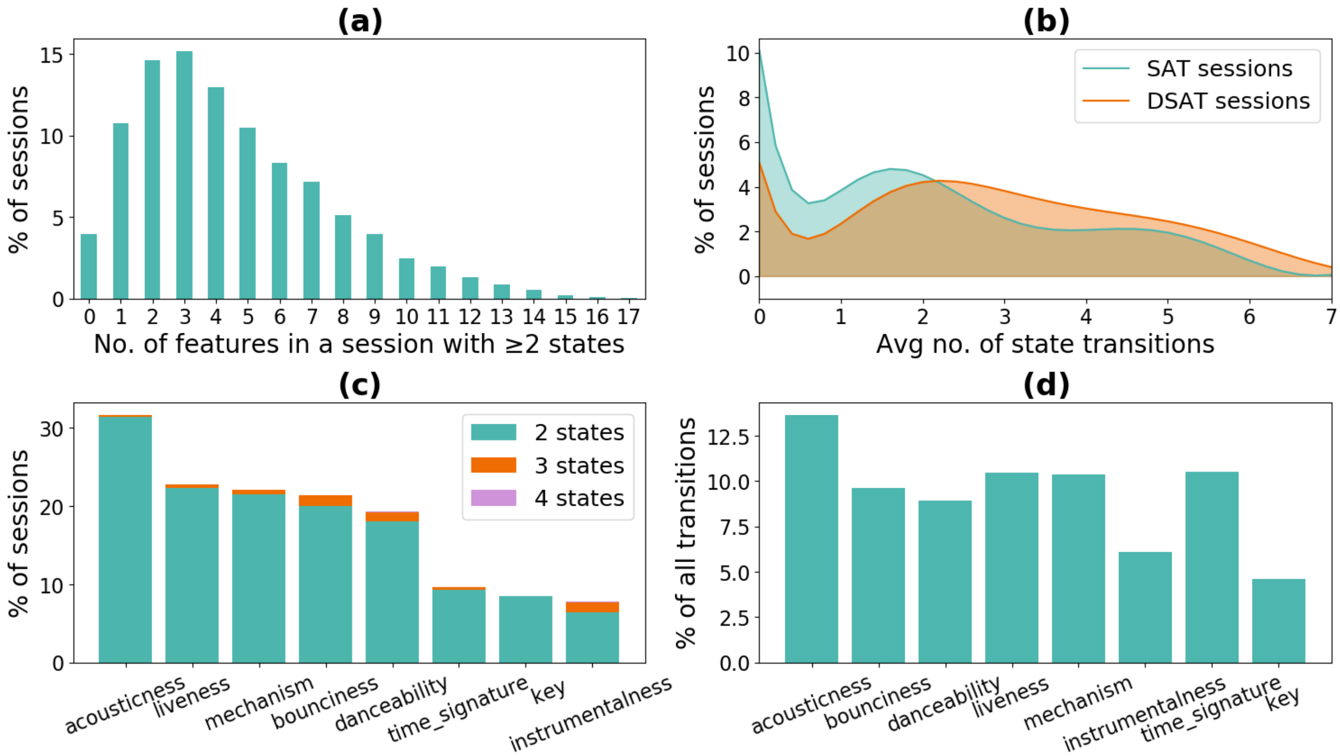

Figure 2: (a) Distribution of sessions with N features that have ≥2 states. (b) Distribution of sessions across average number of state transitions. (c) Percentage of sessions with 2,3,4 states respectively across selected audio features. (d) Percentage of all transitions that are state transitions across selected audio features.

Table 1: Percentage of states over all sessions that are SAT or DSAT for selected audio features

	acousticness	bounciness	danceability	key	liveness	mechanism	organism	tempo
SAT	23.70	21.73	21.90	23.23	20.07	22.53	23.05	23.97
DSAT	25.36	26.50	26.67	25.96	23.64	27.49	25.90	24.06

4.3 RQ3: Are variations in audio properties related to user satisfaction?

4.3.1 Are state transitions correlated with skips? State transitions are eventful as they may lead to users being satisfied or dissatisfied with the new states. While some features have more state transitions than others (Figure 2d), we found that there is a consistent trend across all features when we compare the ratio of state transitions that are correlated with skip/non-skip transitions, to that of all state transitions. That ratio is about 25% across all audio features, which is a significant portion of state transitions.

4.3.2 Are users typically okay with fluctuations? We hypothesize that the number of state transitions affect user satisfaction in sessions, as users may not be happy when the audio characteristics of the tracks in their listening sessions frequently changes. To measure that, we compared the average number of transitions in the top 5 features in SAT and DSAT sessions, and we found that SAT sessions do have fewer state transitions in general, as seen in Figure 2b. The top 5 features are selected based on the total number of transitions the feature has across all sessions.

4.3.3 Are the states of audio features related to user satisfaction? We observe that 40 to 50% of the extracted states in audio features have useful information in Table 1, as they exhibit correlation with SAT/DSAT. This could be leveraged in good ways, and potentially used to optimize sessions for SAT if we re-rank tracks with SAT states over tracks with DSAT states. However, it should be noted that this may be pure correlation – there might be other factors which results in a certain state being related with SAT.

4.4 RQ4: Can insights about audio states help in track sequencing for improved user satisfaction?

Having shown that the states of audio features in sessions are indeed correlated with SAT/DSAT, we attempt to leverage the features and its extracted states to optimize sessions for SAT. While the development of an audio-state aware track sequencing model is beyond the scope of the present study, we perform a counterfactual track re-ranking experiment to investigate the promise offered by the audio state information. Specifically, we use user interaction

Table 2: Ranking results on remaining candidate tracks in 15,000 sessions, after taking out the first $N = 5$ in-session tracks from each session. Rank by User Relevance: Rank by similarity between user vector and candidate track vector. Rank by Audio Similarity: Rank by cosine similarity between the first N in-session tracks and candidate track. Rank by Popularity: Rank by popularity score (us_popularity_estimate in MSSD). Random Top 3 Features: Rank using randomly selected audio features. Global Top 3 Features: Rank using audio features that appear most frequently across all sessions. Known Top 3 Features: Rank using the best audio features for each session.

	NDCG@10			AP@10			P@10			RR@10		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
Rank by User Relevance	0.667	–	–	0.615	–	–	0.530	–	–	0.628	–	–
Rank by Audio Similarity	0.700	–	–	0.642	–	–	0.538	–	–	0.690	–	–
Rank by Popularity	0.719	–	–	0.662	–	–	0.542	–	–	0.712	–	–
Random Top 3 Features	0.672	0.675	0.675	0.618	0.619	0.620	0.531	0.532	0.532	0.636	0.641	0.642
Global Top 3 Features	0.674	0.678	0.681	0.617	0.620	0.623	0.533	0.534	0.535	0.639	0.646	0.653
Known Top 3 Features	**0.730**	**0.734**	**0.734**	**0.681**	**0.682**	**0.682**	**0.544**	**0.547**	**0.546**	**0.706**	**0.719**	**0.721**

information from the entire session, and identify top audio feature which would have resulted in a potential track ordering which minimizes skips. Doing so enables us to understand the scope of potential improvement possible if we have access to such an oracle, which helps us identify appropriate audio feature for each session.

For each session, we pick the top few attributes that helps us re-rank the session best. For each audio feature, we tag its different states as "*good*" state or "*bad*" state, wherein we define the 'good'ness of a state by its *average number of skips* – the lower the average number of skips, the better the state is. Subsequently, each audio attribute can be used to re-rank the tracks in a session by ranking tracks in 'good' states higher than tracks in 'bad' states. Tracks within the same state are ordered by its original session position, since the tracks in a session in MSSD already have an implicit ranking based on user-track similarity. For example, if the 'acousticness' feature in a session of 15 tracks has 2 states, where state A has 5 tracks where 1 track is skipped, and state B has 10 tracks where 8 tracks are skipped, we can rank all the tracks in state A higher than state B, since state A has a lower average skip than state B.

However, when re-ranking a collection of unseen candidate tracks, we do not have information on whether the unseen tracks will be skipped, and consequently how 'good' the states are. We can attempt to predict how 'good' are the states in the top few features by incorporating skip feedback from the first N in-session tracks, then using that prediction to re-rank the remaining tracks. Feedback can be incorporated into the states by assigning a positive score for states that appear in non-skipped tracks, and a negative score for states that appear in skipped tracks. Subsequently, we can treat states with higher scores as better states, and use this to re-rank the candidate tracks.

We begin by analyzing the distribution of audio features amongst the identified top audio attributes. We observe that there exist a variety in the features that best optimize different listening sessions: while some features like 'acousticness' and 'beat_strength' appear more often than others, no single feature dominates all the sessions. This indicates that there is no global audio feature which will work best always, and further motivates the need to develop prediction models that are able to predict the top audio features per session, which can be used to optimize sessions for SAT.

In our experiments, we have found that we are indeed able to optimize sessions for SAT by leveraging audio attributes and their states, as shown in Table 2, where we outperform several baselines. Comparing the different track ranking approaches, we observe that ranking by audio similarity is better than ranking by user relevance. Further, we observe that exposing popular content is well received by users, despite it suffering from the filter bubble issue of biasing recommendations towards popular content.

If we use a random set of features, or a fixed set of top-3 audio attributes that appear most frequently, we do not improve user SAT by much. However, if we assume access to the oracle which helps us identify the top-3 audio features per session, then using those features we are able to improve the SAT metric by over 10% relative to ranking by user relevance. This is a very promising result, since it highlights that by developing and leveraging an accurate audio attribute prediction model, we can hope to increase satisfaction metrics by a significant amount. While this motivates the need for further research around the development of such a predictive model, it also highlights the fact that appropriately considering audio state information in track sequencing is a promising future direction to pursue.

5 IMPLICATIONS & CONCLUSION

Analyzing state transitions of different audio attributes within sessions highlights that fluctuations in audio properties are fairly common in music streaming sessions, and are related with the skipping behavior of users. More importantly, we highlight that leveraging the right audio attribute to re-rank tracks can result in increasing user satisfaction metrics by 10%. These findings bring to attention the importance of developing a number of important prediction and sequencing models. First, we advocate for the development of a real-time model which leverages immediately available user feedback, to predict dominant audio attributes to consider for track re-ranking. Such a system enables development of better contextualized recommendations on streaming platforms, including radio and playlist continuation products. Second, we advocate for track sequencing algorithms which help maintain the sequential aesthetics of music, by appropriately modulating audio properties of subsequent songs to be acoustically similar as a function of creative intent exhibited in the current user session. Such models suffer from less abrupt transitions, thereby leading to enhanced user engagement. Finally, we posit that the importance of audio attributes are conditioned on context, specifically, user intent behind consuming music. This motivates the need for development of intent extraction techniques as well as methods which identify the importance of a specific audio attribute to best support current intent.

REFERENCES

[1] Adam Berenzweig, Beth Logan, Daniel Ellis, Brian Whitman, and Cambridge A. 2003. A Large-Scale Evaluation of Acoustic and Subjective Music Similarity Measures. *Computer Music Journal* 28 (11 2003). https://doi.org/10.1162/014892604323112257

[2] Rachel M. Bittner, Minwei Gu, Gandalf Hernandez, Eric J. Humphrey, Tristan Jehan, Hunter McCurry, and Nicola Montecchio. 2017. Automatic Playlist Sequencing and Transitions. In *Proceedings of the 18th International Society for Music Information Retrieval Conference, ISMIR 2017, Suzhou, China, October 23-27, 2017*, Sally Jo Cunningham, Zhiyao Duan, Xiao Hu, and Douglas Turnbull (Eds.). 442–448.

[3] Geoffray Bonnin and Dietmar Jannach. 2014. Automated Generation of Music Playlists: Survey and Experiments. *ACM Comput. Surv.* 47, 2, Article 26 (Nov. 2014), 35 pages. https://doi.org/10.1145/2652481

[4] Brian Brost, Rishabh Mehrotra, and Tristan Jehan. 2019. The Music Streaming Sessions Dataset. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 2594–2600. https://doi.org/10.1145/3308558.3313641

[5] Sally Jo Cunningham, David Bainbridge, and Annette Falconer. 2006. 'More of an Art than a Science': Supporting the Creation of Playlists and Mixes. In *ISMIR 2006, 7th International Conference on Music Information Retrieval, Victoria, Canada, 8-12 October 2006, Proceedings.* 240–245.

[6] Enric Guaus. 2009. *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.* Ph.D. Dissertation. Universitat Pompeu Fabra.

[7] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2012. Context-Aware Music Recommendation Based on Latenttopic Sequential Patterns. In *Proceedings of the Sixth ACM Conference on Recommender Systems* (Dublin, Ireland) *(RecSys '12)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/2365952.2365979

[8] Dietmar Jannach, Lukas Lerche, and Iman Kamehkhosh. 2015. Beyond "Hitting the Hits": Generating Coherent Music Playlist Continuations with the Right Tracks. In *Proceedings of the 9th ACM Conference on Recommender Systems* (Vienna, Austria) *(RecSys '15)*. Association for Computing Machinery, New York, NY, USA, 187–194. https://doi.org/10.1145/2792838.2800182

[9] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[10] Beth Logan. 2002. Content-Based Playlist Generation: Exploratory Experiments. In *ISMIR 2002, 3rd International Conference on Music Information Retrieval, Paris, France, October 13-17, 2002, Proceedings.* 295–296.

[11] Brian McFee and Gert RG Lanckriet. 2012. Hypergraph Models of Playlist Dialects. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, Mosteiro S.Bento Da Vit 'o ria, Porto, Portugal, October 8-12, 2012*, Fabien Gouyon, Perfecto Herrera, Luis Gustavo Martins, and Meinard M "u ller (Eds.). FEUP Edi c c o es, 343–348.

[12] Aäron van den Oord, Sander Dieleman, and Benjamin Schrauwen. 2013. Deep Content-Based Music Recommendation. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 2643–2651.

[13] Andy Sarroff and Michael Casey. 2012. Modeling and predicting song adjacencies in commercial albums. In *Proceedings of the 9th Sound and Music Computing Conference.* 364—371.

[14] Claus Weihs, Dietmar Jannach, Igor Vatolkin, and Guenter Rudolph. 2016. *Music Data Analysis: Foundations and Applications* (1st ed.). Chapman & Hall/CRC.

[15] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining* (Seattle, Washington, USA) *(WSDM '12)*. Association for Computing Machinery, New York, NY, USA, 13–22. https://doi.org/10.1145/2124295.2124300