AUGMENTING MUSIC LISTENING EXPERIENCES ON VOICE ASSISTANTS

Morteza Behrooz[1] **Sarah Mennicken**[2] **Jennifer Thom**[2]
Rohit Kumar **Henriette Cramer**[2]
[1] University of California Santa Cruz, Santa Cruz, USA
[2] Spotify USA, Somerville, MA, USA
morteza@ucsc.edu, sarahm@spotify.com

ABSTRACT

Voice interfaces have rapidly gained popularity, introducing the opportunity for new ways to explore new interaction paradigms for music. However, most interactions with music in current consumer voice devices are still relatively transactional; primarily allowing for keyword-based commands and basic content playback controls. They are less likely to contextualize content or support content discovery beyond what users think to ask for. We present an approach to dynamically augment the voice-based music experience with background information using story generation techniques. Our findings indicate that augmentation can have positive effects on voice-based music experiences, given the right user context and mindset.

1. INTRODUCTION

Voice-enabled devices, such as "smart speakers" like Amazon's Echo, Apple's HomePod, Google Home, or Sonos One, have reached the mainstream. In particular, listening to music is a popular use case for such devices [23, 25]. Finding music to listen to and discovering music on these devices can be a challenge as the interactions supported by voice-enabled speakers are relatively limited by current interaction models.

Prior research suggests listeners employ music search to learn and explore about new content to consume. Listeners seek background information to stay informed about their favorite artists, genres, and songs, and use it as a relationship builder with others [19]. This exploratory mindset, however, is relatively rare on music streaming apps because catalog-based entity search does not support this user need well [13]. Augmenting listening experiences and conversational interactions have the potential to support these exploratory user goals but leveraging them for a good user experience remains a challenge.

Learning about background information is sometimes a part of the listening experience itself. Often, such information is presented together with the music playback to contextualize the content. For example, user interfaces of several music streaming services, such as Apple Music, Pandora, and Spotify, include a section for additional information beyond basic track metadata for artists, albums, and playlists. Sometimes, songs are contextualized further by displaying the lyrics, stories, or background information associated with certain parts of the songs (e.g., "Behind the Lyrics" feature on Spotify [27]).

In this paper, we provide a method for how voice-based content consumption can be automatically augmented with background information and present the development and study of a prototype inspired by story generation methods.

We make the following contributions:

- Introduction of a type of content augmentation to contextualize voice-based content consumption with background information in Section 3.
- Detailed design of an approach taking playlists as input and utilizing weighted graphs to generate textual music augmentations, inspired by story generation in Section 3.
- Identification of best practices for using augmentation and conversation in voice-based music consumption in Section 5.

2. RELATED WORK

2.1 Listener Information Needs and Music Search

When listeners search for music, they have multiple information needs that they may be trying to fulfill. These user needs help to shape how listeners approach their music search goals. For instance, listeners may be in the mindset of looking for something specific or they may be in the mindset where they are open to multiple types of music-related information. Prior research has suggested that users of a streaming music service have distinct mindsets when they are searching for music [13]. In a focused mindset, users have one particular item in mind. Catalog, entity-based search interfaces favor this particular mindset and queries that align with the structure of available metadata. In an analysis of Google Answers queries, Bainbridge et al. [3] found that users typically (81.3% of the

time) expressed needs through bibliographic queries, using performer, title of work, or date of recording. Li et al. [17] also observed that typed searches on a streaming music platform are typically focused, suggesting that the modality and design of the current feature supported this type of mindset.

Listeners also have broader information needs that are not met by catalog-based entity searches commonly supported in online music services. Lee et al. [15] observed that people use cloud music services that store listeners' music libraries to listen to music that they were unfamiliar with, suggesting that music discovery and exploration is an important user need. In addition, listeners indicated they search for information about the artists and music for learning purposes [14]. Users of a streaming music platform, however, tended not to use the search feature to deeply learn about a specific type of music and left the platform to fulfill that need [13].

2.2 Voice Assistants and Music Consumption

Voice-enabled speakers currently allow music listeners to search for content (e.g., by saying "play Jazz" or "play Time by Pink Floyd") and control the music playback (e.g., play/pause/skip and volume controls). In fact, these basic playback controls form the most common category of user commands [25]. While many of these speakers can be used in conjunction with a secondary device that has a graphical user interface (GUI), voice interaction is increasingly becoming a primary modality for consuming music [25], which increases the importance of evolving and improving the music experience through voice. Notably though, the voice-only smart speaker experience does not offer much in the way of discovery or background information, and such lack of contextualization and grounding can reduce music discovery and listener's emotional investment [26].

Our work focuses on contextualizing the voice-based music experience with relevant background information. This idea shares similarities to music radio shows, where the hosts provide relevant information about the content they play and add other talking points in between songs. In [4], radio's interaction of speech and content is framed as a special kind of narrative, in which the DJ or radio host is the narrator. One of the main challenges in creating an experience like radio shows is maintaining the "flow" of the music, and balancing the spoken words and songs, as this is one of the main skills of the radio hosts [2]. Our user study seeks to learn more about how to achieve a balance between this flow and providing background information.

2.3 Story Generation

Story generation is the problem of automatically selecting a sequence of events that meet a certain criteria and can be narrated as a story [18]. Story generation and our approach to augment the music listening experience share the goal to generate sequences of textual content given specific constraints. While there are many different approaches to generate stories [12, 20, 28], ours is similar to planning-based approaches which also commonly use graph representations to map the space of story events and the possible constraints of a valid or optimal progression of the storyline. In [18] such constraints are reflected by logical precedence rules, while our method utilizes edge weights and pathfinding to extract a preferred storyline. Similar to PlotShot [9] we apply this a graph-based approach to generate a sequence of text for a given form of input media. Inspired by these different approaches, we take playlists as input and utilize graphs with edge weights that denote content preferences to generate textual music augmentations.

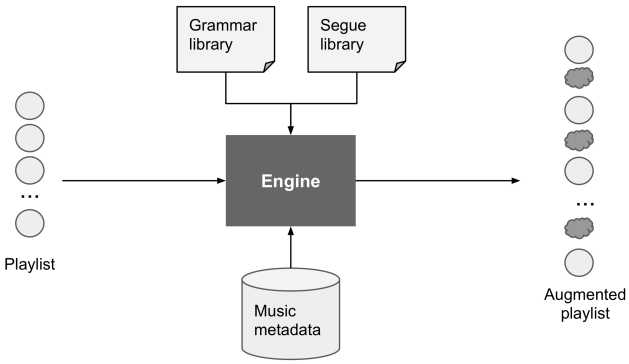

Figure 1. The system architecture of our prototype.

3. APPROACH AND PROTOTYPE

Our prototype takes ordered playlists as input, finds relevant background information and relationships for the songs it contains, and chooses a subset of that information for being used in the output (illustrated in Fig. 1). We call every piece of information that comes in between two consecutive songs a *segue*. Every segue describes a predefined property, such as some information or characteristic, of the next song or a relationship between the current and the next song.

While our approach is not limited to playlists as input, we decided to use them as a starting point given that songs in a playlist typically contain more variety in the metadata as opposed to an individual artist's album. Moreover, songs in a playlist often have an implicit reason for having been grouped together (e.g., being of the same genre, suiting a specific mood or situation [21], artist similarity, etc.) We kept the original order of playlist songs to preserve possible semantic reasons behind curation by playlist creators.

3.0.1 Music Metadata

Our prototype uses a set of metadata and background information about songs, artists, and albums. Table 1 shows some sample entries of qualitative facts about songs and artists consisting of short extractions from publicly available sources of background information (e.g., Wikipedia).

3.0.2 Segue Library and Grammar Library

Table 2 contains a few examples for segues. Each segue has a natural language generation (NLG) template resulting in a "segue text" when realized. Our segue library contains 21 segues. The authoring effort for creating new

Table 1. Songs' metadata examples.

Property	Applicable Entities	Example
Name	Artist, Album, Song	Drake, Scorpion, Wild Thoughts
Genre	Album, Song	Hip Hop
Qualitative fact	Artist, Song	Rihanna's real name is Robyn Rihanna Fenty.

segues simply depends on the complexity of the segue logic.

Inspired by the story generation concept of grammars [7], we defined a simple construct in our system to allow prioritizing authored sequences of segues that are presumed to be interesting. For instance, by preferring a sequence of `ArtistFact`, `ArtistOriginJump`, and `ArtistFact`, an augmentation can focus on the background of songs and their artists. Grammars are an instrument for professional authorship and editorial opinion to be reflected in the system.

3.1 Generating a Sequence of Segues

First, our prototype accesses available metadata about songs, artists, and albums appearing in the playlist. Then it finds all the matching segues for every two consecutive songs which results in a list of segue options for each such position. For the entire playlist, we get a list of these lists, which we call the *story possibility space*. Given that the choice of a segue at each position in this space is independent of other positions, the story possibility space forms a graph and the search problem for finding a sequence of augmentations becomes a problem of finding the best path in this graph. To do so, we use a set of heuristics and preferences which are reflected in a *weighting function*. These scores are assigned as weights to the edges that represent those transitions in our graph.

$$\begin{aligned} weight(s_1, s_2) = {} & diff(s_1, s_2) \\ & + s_1^{pref} + s_2^{pref} \\ & - lengthiness + silence\_reward \\ & + playlist\_reward \\ & + positional\_preference \end{aligned}$$

Several variables enable weighting absolute and relative preferences. $diff(s_1, s_2)$ enables avoiding repetition between consecutive segues. Static "segue preference scores" s_i^{pref} give specific segues authored preference. For example, pointing out a change of genre between two consecutive songs might be more interesting than simply stating title and artist of the next song. Terse responses are often preferred in conversational interactions [10], hence $lengthiness$ punishes a segue if it has a long text and $silence\_reward$ rewards a graph edge if the previous segue is long but the next segue is `NullSegue`. $playlist\_reward$ represents that some segues fit better to a specific type of playlist, such as ArtistQualFact in artist-focused playlists. $positional\_preference$ is used for segues that make only sense at a specific part of a playlist. For example, a playlist introduction with a short authored description only makes sense at the beginning.

Given a weighted graph, we first look for and choose any possible grammar matches. A grammar is a match if there exists a path in a sub-graph of the story possibility space, where the sequence of nodes in that path matches the grammar's sequence of segue types. Edge weights do not have a role in finding a grammar match. If two grammars overlap, we choose the path representing one of them at random.

For the portions of the story possibility space where no grammar match is found, we use the edge weights to find the best path, one with the heaviest sum of weights. If a given portion of the overall graph that needs pathfinding is larger than 5 playlist positions, we find the path step by steps in windows of size 5. In doing so, we ensure that each such window does not contain any segue types that exist in the previous window, and hence avoiding local repetition of segue types.

To exemplify conversational interactions, we identify possible interaction points in which we could trigger a short dialog and let the user response determine which segue option comes next. We do so by checking against simple logic definitions, e.g., if there are specific types of segues in the next list of segue options (see Table 3.)

After the full graph path is determined, we use the realized segue text of the segues in the chosen path, and insert these segue texts in between the songs. An example excerpt of an augmented playlist is shown in Table 4. Our prototype can generate augmentations for any given playlist as long as it has access to the metadata for the songs in that playlist. For our evaluation, we decided to focus on three popular types of playlists to start with those based on an artist, a genre, or listener popularity.

4. EVALUATION

To better understand how our method of adding contextual information to smart speaker experiences affects music listening, we conducted a two-phased study within Spotify.

In phase 1, we gathered feedback from two professional writers who are familiar with the music domain to elicit expert feedback on the content of the segues. They received the written output of our prototype generated for one representative example of each playlist type: artist, genre, and listener popularity. While we invited them to provide any type of feedback, we specifically asked them to share their views on the contents of individual segues and describe how they would approach writing similar content from a professional perspective as writers. After they returned their comments, we conducted a semi-structured interview with both of the writers which took about 45 minutes.

In phase 2, we conducted an internal evaluation with nine Spotify employees (four female, five male) from various parts of the organization to identify potential future improvements and establish a first understanding of user needs. Participants were in their early 20s to late 40s from non-technical functions (such as design, marketing, or operations) and located in various locations across the United

Table 2. Examples for segues, their logic description, and samples for their realized text.

Segue Type	Logic Description	NLG template	Realized text
`NullSegue`	Always a match regardless of the songs.	N/A	N/A
`MundaneSegue`	Always a match regardless of the songs.	Next song is `next_song_name` by `next_song_artist_name`.	Next song is Time by Pink Floyd.
`ArtistOriginJump`	Musical origin of the previous song's artist is different than the next one's.	From `prev_city` where `prev_artist_name`'s musical origins are, to `next_city` where `next_artist_name`'s are.	From Los Angeles where Tupac's musical origins are, to New York City, where Biggie's are.
`SameYearSameArtist`	Previous and next song share the same artist and release year.	Just like the last song, the next song is from `next_song_release_year` by `next_song_artist_name`.	Just like the last song, the next song is from 2007 by Rihanna.

Table 3. Examples for conversational augmentations.

Voice Prompt	User Response	Voice Response
From when do you think this last song was?	Correct	That's right. But the next song, called Shook Ones, Pt. II takes us into a different era. All the way to 1995. (`DifferentEraSegue`)
	Wrong	Actually, it's from 2007. The next song called Shook Ones, Pt. II and [...]. (`DifferentEraSegue`)
Question! Are you more interested in the artist's background or the genre?	Genre	The genre of the upcoming song is called "Latin Trap". (`NextGenreSegue`)
	Artist	Next song is by Cardi B. Here's a fun fact about their biography... (`ArtistQualFact`)

Table 4. Example excerpt of an augmented playlist.

•••

Juicy by The Notorious B.I.G.

Here's The Next Episode by Dr. Dre and Snoop Dogg.

The Next Episode by Dr. Dre

Now switching from the 2001 (Explicit Version) album to one called The Best of 2Pac - Pt. 1: Thug.

California Love by 2Pac

The last song was from 2007. The next song called Shook Ones, Pt. II takes us into a different era. All the way to 1995.

Shook Ones, Pt. II by Mobb Deep

Just like the last song, this song was released in 1995.

Gangsta's Paradise by Coolio

The last and the upcoming song both are described as dark groovy.

•••

States. Each session included a semi-structured interview in which we asked participants about their previous experience with voice assistants and whether or how they look for additional content around music. Each participant was asked to listen to a demo audio file for one of our three playlist types. After answering a short questionnaire, they also interacted with our envisioned conversational experience in a short Wizard-of-Oz (WoZ) demo where an experimenter controlled which content to play. Each playlist type was presented to three users who were randomly assigned to a condition.

The demo consisted of ten shortened songs (first and last 15 seconds) and ten segues (one intro segue, nine transition segues) which were generated using our proposed method and then read by a text-to-speech (TTS) engine. Overall, they had a duration of 5:30-5:50 minutes. The short WoZ section to convey the conversational experience covered three songs only, but between the songs the TTS voice prompted the participant with a potential question such as *"Question! Are you more interested in the artist's background or the genre?"*. Depending on the answer the experimenter chose the next audio file to play to continue the experience. Table 3 shows two examples.

We recorded and transcribed all sessions. Two of the co-authors went independently through the transcripts, first categorizing them for their relevance to the stated research questions and then doing an affinity analysis [5] moving relevant quotes between the high-level categories to derive our findings.

5. FINDINGS

We identified various factors that influence the perception and usefulness of including contextual information in music listening experiences.

5.1 Addressing Listener Needs and Contexts

Music is consumed in vastly different situations, playing a different role for listener's needs in each one. We found that listeners' perceived usefulness of the voice-based augmentations heavily depends on the situation and its unique needs.

Augmentations enable music discovery and education. Augmentations are well received when listeners are in an exploratory mindset. Our participants expressed special interest in voice augmentations to learn about content that is new to them. P4 said: *"[Talking about a playlist containing new songs] I'm like 'Wait, what band is this?' [...] 'What other songs can I listen to from them.' "* P6 described their interest in being able to learn about (niche) genres through such augmentations: *"I feel like metal would work really well for this because a lot of bands have a lot of history behind them [...] it's the opposite of trendy [...] people are still listening to music that was written and performed 20 years ago consistently."*

Similarly, editor 2 saw them as a way of discovering less-known artists by providing information about them: *"[When choosing music automatically] you might end up*

skewing the information toward [...] the top-selling artists of all time; yet obviously there are hugely influential artists that have not sold a lot of records but have impacted other artists and bands." Lastly, P1 brought up the need to identify the right occasions for adding information: *"I like that it's just another way to get to know an artist that you already like and I would potentially like if it was getting to know an artist that you don't know. What I wouldn't like is if it's in between."* This highlights a potential for leveraging listener's level of affinity for an artist they already know, or the predicted level of affinity for a new artist, in determining the quantity or focus of the augmentations.

Activities determine needs for and appropriateness of augmentations. Music often supports a specific listener activity. We found that activities with low cognitive load, such as doing chores or cooking, were commonly mentioned as appropriate contexts for voice-based augmentations. P8 said: *"The perfect experience [is] if I'm at home doing something fun like cooking or something not fun like cleaning."* Activities that require a higher level of focus but that listeners consciously choose to support with background music were perceived to be less suitable. Participants mentioned several examples where the music is serving such an activity-supporting focus like working out, studying, or relaxing and felt that any addition to the music could get into the way of that primary activity. *"I need [the music] to keep the motivation going, keep the music going."* (P8)

5.2 Selection of Appropriate Content for Augmentations

The next category of our findings relates to the content of the augmentation and what it focuses on.

Personalizing augmentations improves the experience. The level of affinity with an artist or genre varies significantly across listeners, and the same is true for the level of familiarity with background information. For example, using a sub-genre to describe a song might be very interesting to someone familiar with the general genre, but vague and uninteresting to someone who is only a casual listener of that type of content. P3 said: *"[...] a high, medium, low, [or] novice/expert setting [would be good], because I'm not an expert on this, so I don't understand [some of the segues]."* Similarly, P1 saw an opportunity to point out to them if they are listening to an artist for the first time: *"Say it's the first time I've listened to an artist, I think it would be cool to learn more about that artist."*

Another frequently mentioned interest for personalization was to allow the listener to adjust the topics that the augmentations focus on (e.g., artist life or genre information). P9 said: *"If I could somehow customize like what's being said by the voice to choose like facts or historical whatever, I think that'd be cool."* Editors had similar views. Editor 2 said: *"we've got one end of the spectrum there is music nerds. They've already put their hands in the air and said, 'Please give me more as much as you can."'* The same editor then drew a parallel between customization of content and augmentation: *"can I add another layer of personalization to this which is, please make [the augmentation] minimal [or] please tell me as much as you can about this artist or this genre."*

Augmentations could explain recommendations or present relevant news. Our study subjects mentioned other types of information that would be useful for them to hear. Music listeners increasingly delegate their choice of music content to streaming services, which use various algorithmic and machine learning methods to choose songs that they believe the listener might enjoy. However, listeners usually do not get any explanation for why a particular set of content is chosen for them. P4 said: *"It kind of guides you to know how they're piecing together this playlist for you. It's like, 'this is why we're playing this song for you',"* and P3 mentioned: *"A lot of times for [automatically generated playlists], I'm like why do I have this song, it would be great if [the voice] could tell me."* Alluding to the same point, Editor 1 noted: *"With just the bare information the name, the title, and to give more information and background obviously [one can] provide a much deeper experience for users and give users the reason why they should continue listening."*

Contextual needs of music listeners often extend to their awareness of the current happenings in the music world. Most prominently, our participants expressed interest in hearing about tours and relevant news headlines. P1 said: *"If they were on tour in my area, that's something I'd want to know,"* and P5 mentioned: *"There's a lot of news always with musicians, whether it's a controversy or other things [...] if you had some of that, like why is this song popular right now or what's going on with this song."* Editor 1 brought up the same point, and discussed the following as an example: *"Let's say [an artist] passes away [...], and you insert a little nugget of information to inform people about that. And then the next song is [by the same artist]. I mean [...] that might make it even more important for maybe someone to listen."*

5.3 Appropriate Presentation of Augmentations

Our last category of findings offers insights about the delivery and presentation of the augmentation content.

TTS voice needs to be trustworthy, high quality and fitting. The synthesized voice in which the segues are presented to the listener was one of the most common topics brought up by our participants; most prominently, the quality of the TTS voice as described by P5: *"With the DJ kind of idea, I think the sound of the thing makes a big difference; so [...] that computery voice takes me out of the moment"*. Despite the quality of the TTS, participants seemed to establish a connection with the agent behind the voice, and explicitly expressed a preference for knowing or at least being able to trust the agent. For instance, P2 said: *"Using someone's voice who is an authority on the genre*

or playlist [is better] [...] there's a difference between that voice telling me little tidbits and somebody like [reference to a Jazz musician]."

Participants mentioned they would like specific properties of the voice, such as gender and accent, to be personalized, either based on the current content or their general preferences. For instance, P4 said: *"I like [it] when people have the Google or Waze, the driving apps, and you can change the accent."* P5 noted: *"I think it would be cool if it was kind of genre-based [...] yea if it's tied to genre or playlist type of thing."* Editor 1 pointed out to voice's gender as well, saying: *"It'd be really jarring to hear like a very male voice [on] Ani DiFranco or Riot grrrl playlists or a very feminist playlist"*.

Augmentations should not be frequent. Participants expressed a preference for segues that connect the previous and the next songs (e.g., by highlighting similarities or differences) over segues that focus solely on the next song. For example, P3 said: *"I like this [...] it tells me a little bit about what I just listened to [...] and then it sets me up into what the next song is going to be,"* and P5 mentioned: *"I [liked] that some of them attempt to link the previous song to the next song."* While semantic continuity is valuable, the frequency of augmentations should not be too high, and segues should not come in between every two songs. We included a representation of an intentional skip (`NullSegue` in Table 2), but it formed either zero or just one out of the 10 generated segues that each participant experienced. Five of our participants (P1, P2, P4, P5, and P8) believed the augmentation was too frequent. P1, for instance, said: *"I definitely in no scenario want [to hear the segues] after every single song"*. Lee et al. [16] found that different user personas have a varying desire for engagement when interacting with music information retrieval systems, which needs to be taken into additional consideration when designing such augmented listening experiences.

Participants enjoyed the conversational augmentations. Our conversational augmentations showcased the ability to ask about the music that is being played, and this was well received by our participants. Most of them (seven out of nine) counted the conversational demo as more fun and interesting than the non-conversational case. When probed on the reasons, participants frequently pointed out the ability to interact. P4, for instance, said: *"You kind of feel like there's this other entity that you're having a conversation with."* In another example, P7 said: *"I think I like this better [than the non-conversational demo]. It was more fun [because of] the interaction aspect of it."* However, two of the participants (P1 and P3) could not imagine themselves using the conversational experience in any situation and generally disliked it. Both participants attributed this dislike to usually preferring a "leaned-back" music consumption mindset, as P3 said: *"I don't want it asking me questions. I actually hated it. It wasn't lean back and was trying to get me to interact..."*

6. LIMITATIONS AND FUTURE WORK

Our results indicate that augmenting voice-based music consumption with background information addresses some of the listener needs that are commonly ignored in current experiences [1, 6]. But similar to how different listening situations affect musical preferences [22], we need to investigate situation-specific preferences for augmentations to understand when music listening is a passive [8], flow-like [11] experience which should not be interrupted.

Our augmentations did not have a "narrative coherence" [24], i.e., a coherent story about a particular topic. In the music context, such narratives could be based on a variety of topics, such as recent events, genres, or artists, all of which were mentioned by our study participants as well. For instance, a dynamically generated augmentation about the history of a genre could focus on songs that represent the turning points of it or have other musical significance. Access to more metadata and large semantic models that capture music-related relationships between various entities can help a story generator in achieving this goal.

In terms of presentation, our evaluation suggests that the quality of the TTS engine seems to be particularly important for music listening experiences. We suspect that the imperfections of the TTS might be more apparent due to a general focus on the audio quality, both for music and voice output. In other use cases for voice assistants, the focus is often more on retrieving the requested information; however, this hypothesis requires further research. Changing the voice's accent or gender, based on explicit listener preference, was stated as an interest by several participants; doing so automatically, such as based on a listener model or audio content, is not only very difficult, but also poses the risk of reinforcing stereotypes of societal and cultural associations for certain types of music.

To minimize negative effects of breaking the audio flow of the music experience, a smoother transition between augmentations and music content is needed. For example, by matching audio properties of the augmentation with those of the surrounding music content, similarly to the techniques used by radio show hosts to match the nearby songs in their ending and beginning [2].

7. CONCLUSIONS

Consuming music via voice assistants is currently a limited experience. While enabling transactional requests for catalog search and basic playback controls, listeners may miss out on context or history of the consumed content. To better support an exploratory mindset for discovering music, we introduce a method that uses story generation techniques to augment the voice-based music experience with relevant background information in between the songs. Our results indicate that adding contextual information to voice-based music interactions can improve smart speaker experience and meet listener needs for music discovery and background information. However, there are also limitations to when the augmentations are desirable based on playback context and listener's mindset and activities.

8. ACKNOWLEDGEMENTS

We thank Keunwoo Choi and Aparna Kumar for their valuable feedback on this paper. Dedicated to Mila Kim, a constant presence during this project.

9. REFERENCES

[1] Music Ally. Everybody's talkin': Smart speakers & their impact on music consumption. http://musically.com/wp-content/uploads/2018/03/SmartSpeakersFinal.pdf, 2018. Accessed: 2018-07-20.

[2] Anupriya Ankolekar, Thomas Sandholm, and Louis Lei Yu. Evaluating mobile music experiences: Radio on-the-go. In *International Conference on Mobile Computing, Applications, and Services*, pages 56–73. Springer, 2018.

[3] David Bainbridge, Sally Jo Cunningham, and J. Stephen Downie. How people describe their music information needs: A grounded theory analysis of music queries. In *ISMIR*, pages 221–222, 2003.

[4] Jody Berland. Radio space and industrial time: music formats, local narratives and technological mediation 1. *Popular Music*, 9(2):179–192, 1990.

[5] Hugh Beyer and Karen Holtzblatt. *Contextual design: defining customer-centered systems*. Elsevier, 1997.

[6] Billboard. From mood playlists to metadata: How smart speakers are the next frontier – and challenge – for the music business. https://www.billboard.com/articles/business/8263197/smart-speaker-challenges-music-business, 2018. Accessed: 2018-07-20.

[7] William F Brewer and Edward H Lichtenstein. Event schemas, story schemas, and story grammars. *Center for the Study of Reading Technical Report; no. 197*, 1980.

[8] Steven Brown, Michael J Martinez, and Lawrence M Parsons. Passive music listening spontaneously engages limbic and paralimbic systems. *Neuroreport*, 15(13):2033–2037, 2004.

[9] Rogelio E Cardona-Rivera and Boyang Li. Plotshot: Generating discourse-constrained stories around photos. In *Proceedings of the 12th AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*, 2016.

[10] Ana Paula Chaves and Marco Aurelio Gerosa. Single or multiple conversational agents?: An interactional coherence comparison. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 191. ACM, 2018.

[11] Frank M Diaz. Mindfulness, attention, and flow during music listening: An empirical investigation. *Psychology of Music*, 41(1):42–58, 2013.

[12] Belén Díaz-Agudo, Pablo Gervás, and Federico Peinado. A case based reasoning approach to story plot generation. In *European Conference on Case-Based Reasoning*, pages 142–156. Springer, 2004.

[13] Christine Hosey, Lara Vujovic, Brian St. Thomas, Jean Garcia-Gathright, and Jennifer Thom. Just give me what I want: How people use and evaluate music search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2019.

[14] Jin Ha Lee, Hyerim Cho, and Yea-Seul Kim. Users' music information needs and behaviors: Design implications for music information retrieval systems. *Journal of the association for information science and technology*, 67(6):1301–1330, 2016.

[15] Jin Ha Lee, Yea-Seul Kim, and Chris Hubbles. A look at the cloud from both sides now: An analysis of cloud music service usage. In *ISMIR*, pages 299–305, 2016.

[16] Jin Ha Lee and Rachel Price. Understanding users of commercial music services through personas: Design implications. In *ISMIR*, pages 476–482, 2015.

[17] Ang Li, Jennifer Thom, Praveen Chandar, Christine Hosey, Brian St. Thomas, and Jean Garcia-Gathright. Search mindsets: Understanding focused and non-focused information seeking in music search. In *Proceedings of the 30th International Conference on World Wide Web Companion*. International World Wide Web Conferences Steering Committee, 2019.

[18] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. Story generation with crowdsourced plot graphs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[19] Adam J Lonsdale and Adrian C North. Why do we listen to music? a uses and gratifications analysis. *British Journal of Psychology*, 102(1):108–134, 2011.

[20] Lara J Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark O Riedl. Event representations for automated story generation with deep neural nets. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[21] Esa Nettamo, Mikko Nirhamo, and Jonna Häkkilä. A cross-cultural study of mobile music: retrieval, management and consumption. In *Proceedings of the 18th Australia conference on Computer-Human Interaction: Design: Activities, Artefacts and Environments*, pages 87–94. ACM, 2006.

[22] Adrian C North and David J Hargreaves. Situational influences on reported musical preference. *Psychomusicology: A Journal of Research in Music Cognition*, 15(1-2):30, 1996.

[23] Edison Research NPR. The smart audio report. https://www.nationalpublicmedia.com/smart-audio-report/, 2017. Accessed: 2018-07-20.

[24] Mark O Riedl. A comparison of interactive narrative system approaches using human improvisational actors. In *Proceedings of the intelligent narrative technologies III workshop*, page 16. ACM, 2010.

[25] Alex Sciuto, Arnita Saini, Jodi Forlizzi, and Jason I Hong. Hey alexa, what's up?: A mixed-methods studies of in-home conversational agent usage. In *Proceedings of the Designing Interactive Systems Conference*, pages 857–868. ACM, 2018.

[26] Elliot Jay Stocks. Music consumption in the era of smart speakers. https://medium.com/@elliotjaystocks/music-consumption-in-the-era-of-smart-speakers-b88d04a18746, 2017. Accessed: 2018-07-20.

[27] The Verge. Spotify and genius are collaborating on info-rich behind the lyrics playlists. https://www.theverge.com/2016/1/12/10750990/spotify-genius-behind-the-lyrics-playlists-iphone, 2016. Accessed: 2018-07-25.

[28] R Michael Young, Stephen G Ware, Brad A Cassell, and Justus Robertson. Plans and planning in narrative generation: a review of plan-based approaches to the generation of story, discourse and interactivity in narratives. *Sprache und Datenverarbeitung, Special Issue on Formal and Computational Models of Narrative*, 37(1-2):41–64, 2013.