UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

A simple method for user-driven music thumbnailing

van Nieuwenhuijsen, A.N.; Burgoyne, J.A.; Wiering, F.; Sneekes, M.

DOI
10.5281/zenodo.4245410
Publication date
2020
Document Version
Final published version
Published in
Proceedings of the 21st International Society for Music Information Retrieval Conference
License
CC BY

Link to publication

Citation for published version (APA):
van Nieuwenhuijsen, A. N., Burgoyne, J. A., Wiering, F., & Sneekes, M. (2020). A simple method for user-driven music thumbnailing. In J. Cuming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, & T. de Reuse (Eds.), *Proceedings of the 21st International Society for Music Information Retrieval Conference: ISMIR MTL2020, Montréal, Québec, Canada, Virtual Conference, 11 to 16 October 2020* (pp. 223-230). ISMIR. https://doi.org/10.5281/zenodo.4245410

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:25 Jan 2022

UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

A simple method for user-driven music thumbnailing

van Nieuwenhuijsen, Arianna; Burgoyne, J.A.; Wiering, F.; Sneekes, Mick

Publication date
2020
Document Version
Final published version
Published in
Proceedings of the 21st International Society for Music Information Retrieval Conference
License
CC BY

Link to publication

Citation for published version (APA):
van Nieuwenhuijsen, A., Burgoyne, J. A., Wiering, F., & Sneekes, M. (2020). A simple method for user-driven music thumbnailing. In J. Cuming, J. H. Lee, B. McFee, M. Schedl, J. Devaney, C. McKay, E. Zangerle, & T. de Reuse (Eds.), *Proceedings of the 21st International Society for Music Information Retrieval Conference: ISMIR MTL2020, Montréal, Québec, Canada, Virtual Conference, 11 to 16 October 2020* (pp. 223-230). ISMIR. https://www.ismir.net/conferences/ismir2020.html

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Aug 2021

A SIMPLE METHOD FOR USER-DRIVEN MUSIC THUMBNAILING

Arianne N. van Nieuwenhuijsen[1] **John Ashley Burgoyne**[2] **Frans Wiering**[1] **Mick Sneekes**[1]
[1] Utrecht University, The Netherlands [2] University of Amsterdam, The Netherlands
anvannieuwenhuijsen@gmail.com, j.a.burgoyne@uva.nl, f.wiering@uu.nl, me@micksneekes.nl

ABSTRACT

More and more music is becoming available digitally, increasing the need to navigate through large numbers of audio tracks easily. One approach for improving the browsing experience is *music thumbnailing*: the procedure of finding a continuous fragment that can represent the whole musical piece. This paper proposes a human-centred approach to creating thumbnails based on listeners' perception, directly asking listeners to identify the most characteristic fragment. We carried out a user study to assign representativeness scores to multiple fragments from a selection of popular music tracks. To strengthen the results, we performed a replication of the same user study with new participants and a different set of music. Thereafter, we used audio features, a segmentation algorithm, and participants' overall familiarity with the songs to predict representativeness scores. The results suggest that neither segmentation nor familiarity have a significant impact on users' thumbnail preferences: even segments with starting points that pay no regard to song structure can be suitable thumbnails. Three high-level audio characteristics, however, do impact the perceived representativeness of a fragment: Raw Intensity, Melodic Conventionality, and Conventionality of Intensity. Based on these findings, we propose a new, easy-to-apply method for music thumbnailing.

1. INTRODUCTION

With the rise of the digital age, more and more music is becoming available; streaming services and websites make music readily accessible to the public. The availability of so much music increases the need to navigate through large numbers of audio tracks easily, e.g., the results of search queries or long playlists. One approach to improve the browsing experience is to create music thumbnails. Music *thumbnailing*, or audio thumbnailing, is the procedure of finding a continuous segment within a musical piece which represents the whole piece [1–4]. By using these shorter fragments of audio, music thumbnails allow users to explore large quantities of music without spending too much time listening to or seeking within complete musical pieces [2, 5]. Audio thumbnailing should not be confused with music *summarisation*, which combines snippets of different parts of the song [2, 6], or audio *fingerprinting*, which creates a simpler representation of musical piece in the form of a vector or sequence [7, 8].

One practical example of music thumbnails even outside the major streaming services is Muziekweb, a Dutch music library that aims to make music and information about music available to everyone. [1] On their website, excerpts can be played to get a sense of the musical pieces on offer. To be able to assess the musical pieces, representative music thumbnails are a must. Currently, however, Muziekweb simply chooses its thumbnails randomly, which makes it likely that these excerpts do not represent the musical pieces very well.

There is no consensus about what approach works best to create good music thumbnails, and even the concept of music thumbnails is ambiguous [3–5, 9–11]. Approaches in previous studies include identification of the most repeated part [1, 4], finding the segment which is the most similar to the average sound [2], chorus detection [3–5, 9, 11], and structural identification of the "main part" [10]. Nonetheless, there is overlap between these approaches as the chorus is often the most repeated part in pop music [4] and is also likely to be the most memorable [3].

This paper proposes a user-driven approach by using the listeners' perception to improve upon Muziekweb's current thumbnailing method. Previous research has dis-

[1] https://www.muziekweb.nl/

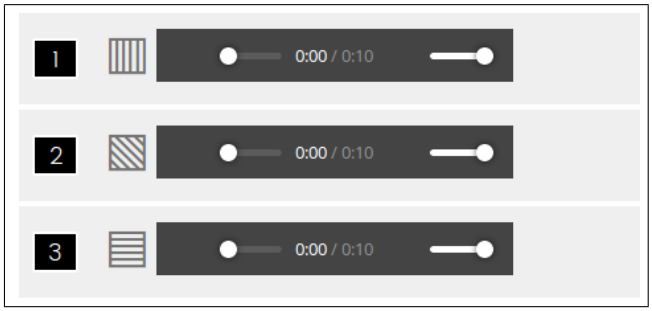

Figure 1. Example of three playable audio fragments of the same tune as displayed in the user study. To be able to distinguish the fragments, the players are displayed with differently filled squares. The numbers show the ranking chosen by the participant.

© Arianne N. van Nieuwenhuijsen, John Ashley Burgoyne, Frans Wiering, Mick Sneekes. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Arianne N. van Nieuwenhuijsen, John Ashley Burgoyne, Frans Wiering, Mick Sneekes, "A Simple Method for User-Driven Music Thumbnailing", in *Proc. of the 21st Int. Society for Music Information Retrieval Conf.*, Montréal, Canada, 2020.

cussed that the best thumbnail could be the segment containing the most memorable and distinguishable part of the musical piece [1, 3]. This aligns with the cognitive definition of *hooks*: hooks are the most salient segments in a musical piece, making them the most recognisable part of a song [12]. Suggestions have already been made about the potential of hooks and catchiness for music search engines [13]. Therefore, the method here is inspired by previous studies on catchiness to identify the most representative part as a music thumbnail.

To use listeners' perception for thumbnailing, we set up a user study to gain information on the representativeness of different fragments of the same tunes. The main task in this user study asked participants to rank three segments of the same song with respect to how well they conveyed a general idea of the song (see Figure 1). Thereafter, features were extracted from the audio fragments with the CATCHY toolbox [14]. To increase the interpretability of these features, we conducted an exploratory factor analysis for dimensionality reduction. These factors, the segmentation method, and the participant's familiarity with the songs were used to create an approximation of the scores from the user study with a linear model. Finally, we combined the feature loadings of the factors and the parameters of the linear model to create a function that can rate the relative representativeness of fragments within a song. The best-rated fragment in any set of candidates would be chosen as the audio thumbnail. To confirm our findings, we repeated the user study with new participants and a different music set and found a very similar result. Based on the findings, we propose a new user-driven method for music thumbnailing. Although we are certainly not the first researchers to test a thumbnailing algorithm on users, to our knowledge, this is the first published study to derive an algorithm for music thumbnailing algorithm directly from user preferences.

2. METHOD

2.1 Music Selection

Consistent with previous studies on catchiness [12, 14, 15], our study focused on popular music. The music came from lists of the 100 most-played songs on Muziekweb's website in 2017 and 2018, to ensure the data consisted of well-known music. Where the lists contained more than one song in languages other than English or Dutch (the two languages that would be most familiar to Muziekweb users), we retained only the most-played song. The resulting list was further reduced by removing songs with low play counts from artists or albums that appeared multiple times on the lists, in order to keep the music as diverse as possible. This resulted in a set of 60 songs, which Muziekweb provided to us as FLAC files. [2]

[2] The song list, segment start times, computed features, and analysis code can be found at `https://github.com/arianne-n/ISMIR-2020-User-Driven-Music-Thumbnailing`

2.2 Segmentation

Because hooks mostly occur at the start of structural sections [12, 13], we used a boundary detection algorithm to identify the start of these structural sections. Specifically, we used an algorithm that identifies boundaries based on structural features and time series similarity [16], as implemented in the Python package MSAF [17]. [3] We used Pitch Class Profiles (PCPs) as the underlying time series for segmentation, as the audio features we use to analyse the results are also mostly harmonic. The other harmonic time series available in MSAF were either too slow or resulted in too few boundaries to be feasible. Moreover, using PCPs aligns with previous thumbnailing studies that describe the importance of chroma [1, 11].

Thumbnails are by their nature short, and as such, our user study used only short excerpts from the original audio: 9.95 seconds, starting from one of the detected boundaries. Muziekweb is only allowed to make 29.9 seconds of music per song available on their site due to copyright, and excerpts of this length allow users to compare three segments from each song without causing copyright violations. Previous studies have assumed the middle of the song to be the most characteristic [3, 5], while others have noted the intro can also serve as a hook [18, 19]; given the conflicting opinions in the literature, we simply chose four segments at random among all the detected boundaries.

To check whether the segmentation method impacts the representativeness of fragments, we also created two extra baseline segments per song. The first is based on Muziekweb's current method: it picks any random point in a song as the start of the segment. The second baseline segment starts at the 1-minute mark in the song, skipping the intro, but staying away from the end. This resulted in six segments for each of the 60 songs.

2.3 User Study Design

The aim of the user study was to provide scores of the representative power for each of the six segments of the 60 songs. The study was carried out as a web-based survey, accessible between 3 April 2019 and 27 May 2019. Participants were recruited via social media and the Muziekweb website. Consent from the institutional ethics committee was acquired prior to collecting any data.

The main task in the survey was similar to the prediction task in the Hooked on Music study of catchiness [12], but rather than asking participants to make a binary choice, participants needed to provide an ordered ranking. Each question would display the title and artist of a song along with three audio fragments (see Figure 1). The participants were asked to rank the fragments on how well they helped them to get a sense of what the song is about (*"een idee van het nummer"*). This phrasing was intended to trigger participants to follow their gut feeling about the song, without thinking too much; asking for a ranking was intended to encourage participants to provide finer-grained distinctions than we might have obtained from a traditional rating

[3] `https://msaf.readthedocs.io/`

scale. The participants were also asked whether they were familiar with the song with a simple yes–no question.

The survey was implemented in the online survey platform Qualtrics. [4] Qualtrics offers a built-in drag-and-drop option whereby users can drag alternatives and place them in their preferred order. To help users distinguish among the different options visually, we gave each fragment one of six differently filled squares as a drag handle. These fill patterns have no apparent ordering in and of themselves, so as to avoid any bias during the ranking.

The survey started with a short explanation of the task, an informed consent form, and two practice songs. Then, the 60 songs were presented to each participant in random order, with a random selection of three of its segments also initially presented in random order. We elected for three segments instead of all six in order to keep the task manageable for participants. Participants were allowed to participate only once to reduce chances of bias. Participants were not required to complete all 60 songs and could end the survey whenever they wished.

2.4 Measures

Based on the data from the user study, we compute a representativeness score for each fragment. The data are in the form of partial rankings: for each song, we know each participant's relative ordering of three segments, but we have no information about their perception of the other three segments. The easiest way to model data in this form is to use a type of discrete-choice model known as the *Plackett–Luce* or *exploded logit model* [20]. We used the variant of the model implemented in the R package **PlackettLuce** [21]. The model is similar to a softmax function or a sort of logistic regression for rankings: specifically, it estimates the probability that, were participants given a choice among *all* six segments of a song, they would choose a particular segment as the most representative thumbnail. This probability is called the segment's *worth*.

In the user study, participants were also asked whether they were familiar with the songs they ranked. We convert these ratings to a continuous familiarity score by dividing the number of responses that indicated that a participant was familiar with the song by the number of responses where a participant was not familiar. As a continuity correction, we add one extra count to the numerator and to the denominator. Finally, we take the log of this ratio, and the standard score (z) of the result:

$$\text{familiarity} = z\left\{\log\frac{\text{known}+1}{\text{unknown}+1}\right\}. \qquad (1)$$

2.5 Audio Features

We evaluate the core measures from the user study with the help of audio features from the CATCHY toolbox [14]. This toolbox can compute psychoacoustic features such as loudness, roughness, and sharpness as well as more common MIR features such as MFCCs, melodic pitch height estimates, and chroma based on HPCPs. Additionally, the CATCHY toolbox introduces three higher-dimensional harmonic and melodic features that attempt to bring some of the concepts available in symbolic music processing to audio. The first is the *Harmonic Interval Co-occurrence* (HIC), which describes the distribution of triads based on their interval representation. The *Melodic Interval Bigram* (MIB) indicates how often triples of successive melodic pitches occur in the melody. Lastly, the *Harmonic Interval* (HI) measures how often pitches in the melody are accompanied by harmonic pitches measured in the chroma.

The last feature of the toolbox is the implementation of *first-order* and *second-order* features for audio. First-order features are computed using the intrinsic content of the music or audio itself, such as the average note duration within the melody [14, 15, 22]. Second-order features reflect the characteristics of the music in context of a corpus. This means that corpus-based second-order features describe the *commonality* of a segment as it describes the segment in the context of the complete corpus. Song-based second-order features outline the *recurrence* of the segment within the song as it measures characteristics of a segment in relation to the whole song.

Like most MIR toolboxes, CATCHY creates a larger set of features than desirable for interpretability, and there is substantial overlap among some subsets of features. We conducted an Exploratory Factor Analysis (EFA) on all the features as a means of dimensionality reduction similar to [14]. Closely related to PCA, EFA looks for shared variance to identify a smaller underlying latent structure responsible for a larger set of observed features [23]. We used Spearman rank correlations instead of Pearson correlations as the basis for our EFA to avoid problems with the non-normality of some CATCHY features. Given a correlation matrix, several algorithms for EFA are in wide use; we chose the standard minimum residual method, which is commonly used for exploratory and descriptive analyses [24]. In order to maximise interpretability, we then rotated the latent factor space using Varimax, a common orthogonal rotation that pushes as many loadings as possible either toward 0 or toward the extremes (correlation of -1 or 1 with a latent factor) [23].

2.6 Regression

The last step is to combine the features to obtain insights into what contributes to the representativeness of segments. We model a segment's representativeness with a log-linear regression implemented as a generalised linear model (GLM) [25, 26]. In this case, the independent variables are the features derived from the audio, the familiarity score, and the segmentation method; the non-linear dependent variable is the Plackett–Luce worth. Although more complex models would be possible, given the applied nature of this research, we are aiming for simplicity as much as accuracy: linear relations among independent variables make the models both easy to interpret and easy to implement for non-experts. By using the resulting model to assign a worth to an unseen fragment, new fragments can be evaluated and the fragment of a song with the

[4] https://www.qualtrics.com

highest value can thereafter be used as a music thumbnail.

2.7 Replication Study

As a final check, we ran a bilingual replication study with a new set of data from Muziekweb. The data set for this study consisted of an arbitrary list of 32 songs derived from the Dutch "Top 2000" of 2019. Although it is less directly connected to Muziekweb users, it should nonetheless also represent music that would be well known to its users. The only substantial difference in the replication was that we did not ask participants explicitly whether they were familiar with the songs. Results of the original study indicated that familiarity had no impact on how segments were perceived, and we hoped to encourage participants to rate more songs by reducing the number of extraneous questions. Ethical consent was also acquired for the replication study and the survey was available from 24 March 2020 until 30 April 2020.

3. RESULTS

3.1 Number of Responses

The original survey received 148 responses, of which 76 participants quit without completing more than the example questions, 14 participants completed the survey completely (i.e., all 60 songs), and 58 partially. The mean number of songs ranked per participant was 25 (SD = 21). This resulted in each segment being ranked by a mean of 15 participants (SD = 3). Segments in the replication study were ranked 17 times on average (SD = 3).

3.2 Dimensionality Reduction

To aid interpretability, we conducted an EFA based on all the CATCHY features computed for the two studies combined. As there is much disagreement in the literature about how to choose the optimal number of factors in EFA, we used a simple heuristic that each factor had to have at least three features with high loadings (correlation higher in magnitude than 0.4) to facilitate easier interpretation of factors [27]. This led to a maximum of five factors. Underfactoring is more harmful than overfactoring, and as four factors started to have more overlap between factors, we retained all five factors to improve identifiability.

Table 1 shows the CATCHY features which had loadings of magnitude greater than 0.4 for one of the five latent factors. To get a sense of what these factors are measuring, we consider the features with the highest loadings per factor.

3.2.1 Harmonic and Melodic Entropy

The first factor consists of second-order features describing harmony and melody. High absolute entropy is combined with high cross-entropy with respect to segments' own songs and with respect to the entire corpus of songs we considered. Sharpness also positively influences the factor, but does so with a far lower loading. This factor thus describes unpredictability or lack of motivic repetition in the harmony and melody; we call it Harmonic and Melodic Entropy.

	Factors				
Feature	1	2	3	4	5
HI Entropy	**0.90**	0.20	0.05	-0.01	0.05
MIB Entropy	**0.90**	0.35	-0.03	0.00	0.04
HI × Song Entropy	**0.89**	-0.29	0.13	-0.02	-0.01
MIB × Corpus Entropy	**0.88**	0.26	-0.05	0.00	0.04
MIB × Song Entropy	**0.88**	0.31	-0.05	-0.01	0.01
HI × Corpus Entropy	**0.87**	-0.33	0.13	-0.01	0.01
HIC Entropy	**0.86**	-0.29	0.09	0.02	0.02
HIC × Corpus Entropy	**0.85**	-0.31	0.14	0.00	0.01
HIC × Song Entropy	**0.85**	-0.28	0.14	-0.01	0.00
Sharpness	**0.46**	0.17	0.23	0.16	0.31
HI | Song	-0.33	**0.52**	0.01	0.15	0.06
HIC | Corpus	-0.20	**0.47**	0.19	0.17	0.13
HI | Song	-0.29	**0.47**	0.00	0.18	0.13
MIB | Corpus	0.11	**0.45**	0.01	0.25	0.09
HIC | Song	-0.21	**0.41**	0.15	0.16	0.15
Loudness	0.08	-0.01	**0.92**	0.00	-0.03
Roughness	0.30	0.01	**0.82**	0.07	0.05
Melodic Pitch Height	0.12	-0.08	**0.50**	-0.02	-0.10
MFCC Variance	0.21	-0.13	**-0.49**	0.02	0.00
MFCC Mean | Corpus	0.16	0.21	**0.45**	0.14	0.25
Loudness SD	0.32	-0.07	**0.43**	0.10	0.07
MIB Entropy | Corpus	-0.02	0.10	0.04	**0.79**	-0.02
HI Entropy | Corpus	-0.03	0.09	0.04	**0.77**	0.03
HI Entropy | Song	-0.01	-0.04	0.03	**0.55**	0.15
MIB Entropy | Song	-0.03	-0.04	0.01	**0.53**	0.08
Loudness | Corpus	0.10	0.09	-0.25	0.02	**0.55**
Loudness | Song	0.08	0.08	-0.05	0.08	**0.50**
Roughness | Song	0.08	0.08	0.28	0.06	**0.50**
Roughness | Corpus	0.13	0.05	**0.41**	0.02	**0.42**

Table 1. Factor loadings for Minimum Residual EFA for the features with loadings above 0.4 for one of the factors. The factors group features together that explain the same variance. 1. Harmonic and Melodic Entropy; 2. Harmonic Conventionality; 3. Raw Intensity; 4. Melodic Conventionality; 5. Conventionality of Intensity.

3.2.2 Harmonic Conventionality

The second factor also consists of second-order features for harmony and melody. For this factor, however, the loadings prefer higher values for the commonality and recurrence of these features rather than entropy calculations. Note that the commonality of HIC and HI is of high importance both with within songs and across the entire corpus, whereas the melody-based MIB loads only against the full corpus. There is, of course, a high correlation between melody and harmony, and so we call this factor Harmonic Conventionality, while acknowledging that it may also have some melodic aspects. This factor can indicate repetition within a song itself, as well as tonal language that does not stray too far from our corpus norm.

3.2.3 Raw Intensity

The third factor mostly relies on high positive values for loudness (mean and standard deviation) and roughness. It also prefers a lower MFCC variance, which means a fragment is more consistent, a high MFCC mean in comparison to the complete corpus, which could be caused due to

	Overall Result			Original – Replication		
Feature	*b*	*SE*	*p*	*b*	*SE*	*p*
Intercept	0.15	0.06	0.006	−0.22	0.11	0.044
Audio Factors						
Harmonic and Melodic Entropy	0.03	0.04	0.448	−0.32	0.09	<0.001
Harmonic Conventionality	0.11	0.05	0.014	0.01	0.09	0.912
Raw Intensity	0.20	0.04	<0.001	−0.12	0.09	0.149
Melodic Conventionality	0.19	0.04	<0.001	0.19	0.09	0.036
Conventionality of Intensity	0.27	0.05	<0.001	0.16	0.09	0.082
Segmentation Strategies						
MSAF	−0.21	0.06	<0.001	0.14	0.13	0.249
Random	0.13	0.11	0.238	0.06	0.15	0.585
1-minute	0.10	0.08	0.205	−0.43	0.16	0.009

Table 2. GLM results showing how features contribute to perceived representativeness of thumbnails ($R^2 = 0.09$). The left-most part shows estimates using the data from both the original and replication study. The right-most results shows the differences in estimates between the two studies. For each of these results, the estimate or coefficient (b), the standard error (SE), and p-value are given.

MFCCs also measuring loudness, and a high melodic pitch height. We call this factor the Raw Intensity of a fragment, as fragments that score high on this factor sound noticeably more "aggressive" than those that do not.

3.2.4 Melodic Conventionality

The fourth factor is heavily based on corpus as well as song-based second-order features for the MIB and HI entropy. This means that the values for this factor rise when the dispersion of MIB and HI is typical for a song or the corpus. Remember also that HI is a measure that explicitly incorporates melodic information. Thus, this factor primarily describes the commonality and recurrence of the dispersion of melodic bigrams and the melody aligning with the harmony. We call it Melodic Conventionality, although it is somewhat less directly linked to conventionality than the Harmonic Conventionality factor.

3.2.5 Conventionality of Intensity

The last factor comprises corpus- and song-based second-order features for the most important components of Raw Intensity. It is easy to understand but hard to name; we call it Conventionality of Intensity.

3.3 Log-Linear Model

A GLM based solely on the data of the original user study showed that familiarity had no significant impact on how participants ranked the segments ($b = -0.08$, $SE = 0.07$, $p = .27$). As mentioned above, we excluded familiarity from the replication study in order to lessen the burden on our participants. We also exclude it from further analysis.

Table 2 showcases the results of a larger GLM with both the original and replication studies combined ($R^2 = .09$). The left part shows how each variable contributes to the representative worth of a fragment overall. The right side shows the differences in these parameter estimates between the original and replication studies. The results indicate that the Raw Intensity, Melodic Conventionality, and Conventionality of Intensity are the most important factors to approximate a segment's worth, each having a positive effect. Although there is a statistically significant difference between the two studies with respect to the *size* of the effect of Melodic Conventionality, the *direction* of the effect is the same in both studies. The role of Harmonic and Melodic Entropy is less clear: its effect on worth goes in opposite directions between the original study and the replication. Harmonic Conventionality has a small positive effect in each study. The effects of segmentation are less consistent (there is a significant Segmentation × Experiment interaction, $\chi^2(2) = 6.80$, $p = .03$) but with one surprising finding: in both studies, choosing thumbnails that line up with (estimated) structural boundaries seems to make users' opinions *worse*.

4. DISCUSSION

The results show that the most significant features that could contribute to the representative worth of a fragment are the Raw Intensity, Melodic Conventionality, and Conventionality of Intensity. Conventionality of Intensity has the highest impact on the representative worth: users prefer typical levels of intensity, neither too "hard" nor too "soft", for thumbnails. In addition to Conventionality of Intensity, higher-intensity thumbnails are preferred, as well as thumbnails with typical, familiar melodic patterns. The effect of Harmonic Conventionality is statistically significant, but its effect size is quite small; if anything, it may have a small positive effect on the perceived quality of a thumbnail.

Our results also show that the effect of Harmonic and Melodic Entropy seems to differ between the original and replication study. As both data sets had the same data format and were used to create the factors, the difference is most likely caused by the songs themselves. The replica-

tion study contained primarily pop-rock songs, whereas the original study also contained a broader of popular styles, e.g., rap and trance. Harmonic and melodic entropy are fundamental and sometimes genre-defining musical characteristics, and as such, it is not surprising that the effect of this factor would differ. This possibly genre-dependent aspect of thumbnails could be an interesting area for future work.

The impact of the segmentation method on the representative worth shows that in contrast to our hypothesis, segments chosen by a segmentation method do not outperform the base cases: in fact, boundary-aligned thumbnails seem to perform worse. While a thumbnail may benefit from containing the most memorable and recognisable part of a song, it does not necessarily need to start at that point. In practice, an algorithm for selecting thumbnails is going to be more successful if it simply has many candidate thumbnails to choose from, without worrying about where they start.

Altogether, users most prefer music thumbnails with high intensity and conventional, frequently recurring intensities and melodic patterns. This aligns with previous automatic thumbnailing studies, which have mostly focused on detecting the most repeated section or chorus [1,3–5,9,11]. Moreover, previous research shows that the chorus is generally louder, has a higher and more salient pitch, and has less dynamic diversity [28], which overlaps with the factor for Raw Intensity in this study.

A similarity can also be found with research on catchiness, which shows that the most memorable parts of a song have a more typical sound, more conventional melodies, more recurrence in the timbral aspects, as well as a prominent vocal line [14]. Earworms, which are related to catchiness, also seem to appear more in often recurring fragments with a faster tempo and a common melodic contour [22, 29]. In short, our findings about listeners' thumbnail preferences are consistent with previous studies on thumbnails, choruses, and catchiness.

4.1 Proposed Thumbnailing Method

Based on these results, we propose a new method for music thumbnailing. First, several fragments should be obtained from the song. The results of this study show that there is no preferred segmentation method and therefore that any method that results in a reasonable amount of fragments suffices. Then, the CATCHY features for each of these fragments need to be computed. An approximation of the factors in this study can be computed by multiplying standardised feature values by the highest factor loadings for the Raw Intensity, Melodic Conventionality, and Intensity Conventionality. Thereafter, these approximations are multiplied by the estimates of the GLM of the combined results to gain a representative score. The fragment of a song with the highest score can be selected as the music thumbnail.

4.2 Limitations

Like any user study, our research has some limitations. First, this study only focuses on pop music; the results cannot necessarily be transferred to other musical genres [11]. Apart from the musical genre, the choice of a linear model might also have been too simplistic to grasp fully how audio features are related to perceived representativeness. More insights might be gained by also considering non-linear models that could pick up more intricate relationships. While this study does consider features for psychoacoustics and harmony, rhythm is not considered. Further research might look into the effects of rhythm features on representativeness. Lastly, the segmentation method used here had a negative impact on the representativeness score; perhaps a different algorithm might have yielded better results. Nonetheless, it is clear from our findings that simple heuristics like starting at a fixed time point or even a fully random starting point can also yield effective thumbnails.

5. CONCLUSION

This study aimed to create a user-driven music thumbnailing method based on easily computable audio features and an easy-to-implement scoring strategy. Segments of well-known pop songs were obtained and audio features of these segments were derived with the CATCHY toolbox. Thereafter, the segments were presented in two user studies where participants could rank segments on their representativeness. Using the data from the user studies, we used a log-linear model to understand how audio features might explain the perceived worth of a potential thumbnail. The results were significant: representativeness seems to be positively influenced by a higher intensity, and a higher commonality and recurrence of intensity and melodic dispersion. Based on these findings, we propose a new, easy-to-apply method for music thumbnailing.

6. ACKNOWLEDGEMENT

We would like to thank Muziekweb for the idea to look into music thumbnailing, providing the audio files of the music, and for their help to reach out to possible participants. We also would like to thank the participants of the user studies; without their help the results could never have been obtained.

7. REFERENCES

[1] W. Chai and B. Vercoe, "Music thumbnailing via structural analysis," in *Proceedings of the 11th Association for Computing Machinery (ACM) International Conference on Multimedia*, 2003, pp. 223–226, https://doi.org/10.1145/957013.957057.

[2] M. L. Cooper and J. Foote, "Automatic music summarization via similarity analysis," in *Proceedings of the 3rd International Society of Music Information Retrieval Conference (ISMIR)*, Paris, France, 2002.

[3] Y.-S. Huang, S.-Y. Chou, and Y.-H. Yang, "Music thumbnailing via neural attention modeling of music emotion," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 347–350, https://doi.org/10.1109/apsipa.2017.8282049.

[4] M. Müller, N. Jiang, and P. Grosche, "A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 531–543, 2013, https://doi.org/10.1109/tasl.2012.2227732.

[5] M. Levy and M. Sandler, "Application of segmentation and thumbnailing to music browsing and searching," in *Proceedings of the Audio Engineering Society Convention 120*, 2006.

[6] D. F. Silva, F. V. Falcao, and N. Andrade, "Summarizing and comparing music data and its application on cover song identification," in *Proceedings of the 19th International Society of Music Information Retrieval Conference (ISMIR)*, Paris, France, 2018.

[7] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *Journal of VLSI Signal Processing Systems for Signal, Image and Video Technology*, vol. 41, no. 3, pp. 271–284, 2005.

[8] J. Van Balen, F. Wiering, and R. Veltkamp, "Audio bigrams as a unifying model of pitch-based song description," in *Proceedings of the 11th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, Plymouth, United Kingdom, 2015.

[9] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005, https://doi.org/10.1109/tmm.2004.840597.

[10] H. Nawata, N. Kamado, H. Saruwatari, and K. Shikano, "Automatic musical thumbnailing based on audio object localization and its evaluation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, https://doi.org/10.1109/icassp.2011.5946323.

[11] B. Schuller, F. Dibiasi, F. Eyben, and G. Rigoll, "Music thumbnailing incorporating harmony- and rhythm structure," in *International Workshop on Adaptive Multimedia Retrieval*, 2008, pp. 78–88.

[12] J. A. Burgoyne, D. Bountouridis, J. Van Balen, and H. Honing, "Hooked: A game for discovering what makes music catchy," in *Proceedings of the 14th International Society of Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013, pp. 245–250.

[13] H. J. Honing, "Lure(d) into listening: The potential of cognition-based music information retrieval," *Empirical Musicology Review*, vol. 5, no. 4, pp. 146–151, 2010, https://doi.org/10.18061/1811/48549.

[14] J. Van Balen, J. A. Burgoyne, D. Bountouridis, D. Müllensiefen, and R. C. Veltkamp, "Corpus analysis tools for computational hook discovery," in *Proceedings of the 16th International Society on Music Information Retrieval (ISMIR)*, Malaga, Spain, 2015, pp. 227–233.

[15] D. Müllensiefen and A. R. Halpern, "The role of features and context in recognition of novel melodies," *Music Perception: An Interdisciplinary Journal*, vol. 31, no. 5, pp. 418–435, 2014, https://doi.org/10.1525/mp.2014.31.5.418.

[16] J. Serra, M. Müller, P. Grosche, and J. Ll. Arcos, "Unsupervised music structure annotation by time series structure features and segment similarity," *IEEE Transactions on Multimedia*, vol. 16, no. 5, pp. 1229–1240, 2014, https://doi.org/10.1109/tmm.2014.2310701.

[17] O. Nieto and J. P. Bello, "MSAF: Music structure analysis framework," in *Proceedings of the 16th International Society on Music Information Retrieval (ISMIR)*, 10 2015.

[18] G. Burns, "A typology of 'hooks' in popular records," *Popular music*, vol. 6, no. 1, pp. 1–20, 1987, https://doi.org/10.1017/s0261143000006577.

[19] C. Kronengold, "Accidents, hooks and theory," *Popular Music*, vol. 24, no. 3, pp. 381–397, 2005.

[20] S. Beggs, S. Cardell, and J. Hausman, "Assessing the potential demand for electric cars," *Journal of Econometrics*, vol. 17, no. 1, pp. 1–19, 1981, https://doi.org/10.1016/0304-4076(81)90056-7.

[21] H. L. Turner, J. van Etten, D. Firth, and I. Kosmidis, "Modelling rankings in R: The PlackettLuce package," *arXiv preprint arXiv:1810.12068*, 2018.

[22] K. Jakubowski, S. Finkel, L. Stewart, and D. Müllensiefen, "Dissecting an earworm: Melodic features and song popularity predict involuntary musical imagery," *Psychology of Aesthetics, Creativity, and the Arts*, vol. 11, no. 2, pp. 122–135, 2017, https://doi.org/10.1037/aca0000090.

[23] J. W. Osborne, A. B. Costello, and J. T. Kellow, "Best practices in exploratory factor analysis," *Best Practices in Quantitative Methods*, pp. 86–99, 2008, https://doi.org/10.4135/9781412995627.d8.

[24] H. E. Tinsley and D. J. Tinsley, "Uses of factor analysis in counseling psychology research," *Journal of Counseling Psychology*, vol. 34, no. 4, pp. 414–424, 1987, https://doi.org/10.1037/0022-0167.34.4.414.

[25] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[26] J. A. Nelder and R. W. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society: Series A (General)*, vol. 135, no. 3, pp. 370–384, 1972.

[27] T. A. Brown, *Confirmatory Factor Analysis for Applied Research*, 2nd ed. New York: Guilford, 2015.

[28] J. Van Balen, J. A. Burgoyne, F. Wiering, and R. C. Veltkamp, "An analysis of chorus features in popular song," in *Proceedings of the 14th International Society of Music Information Retrieval Conference (ISMIR)*, Curitiba, Brazil, 2013.

[29] V. J. Williamson, S. R. Jilka, J. Fry, S. Finkel, D. Müllensiefen, and L. Stewart, "How do 'earworms' start? classifying the everyday circumstances of involuntary musical imagery," *Psychology of Music*, vol. 40, no. 3, pp. 259–284, 2012, https://doi.org/10.1177/0305735611418553.