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

User-Driven Music Thumbnailing

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. Whereas previous studies have looked into the music itself to define music thumbnails, we propose a human-centred approach based on listeners' perception.

User Study

The listeners perception was measured with a user study in which participants were asked to evaluate 60 pop songs by

1. listening to fragments of the each song
2. ranking these fragments on their representativeness
3. stating whether they are familiar of the song

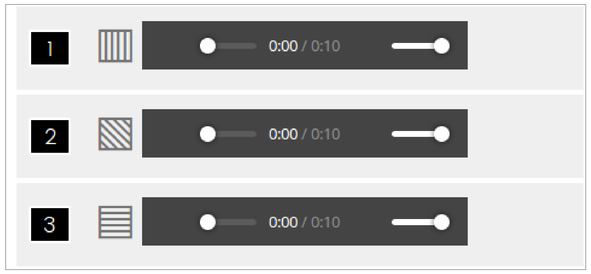
Six fragments per song were derived for evaluation and were segmented by choosing a random starting point, starting at 1-minute in, or by using a segmentation method implemented in the Python package MSAF. A replication study with 32 additional pop songs was conducted to strengthen the results.

Audio-Derived Features

Audio-derived features were computed with the CATCHY toolbox to measure:

- psychoacoustic features (e.g. loudness)
- common MIR features (e.g. MFFCs)
- three higher-dimensional melodic and harmonic features

Additionally, the toolbox computes *first-order* and *second-order* features. First-order features describe the intrinsic content of the audio, while second-order features describe the *commonality* or *recurrence* of a fragment.



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.

Dimensionality Reduction

To increase interpretability, an *Exploratory Factor Analysis* was conducted as a means of dimensionality reduction. This resulted in five factors which were:

Harmonic and Melodic Entropy describes the unpredictability or lack of motivic repetition in the harmony and melody.

Harmonic Conventionality indicates repetition within a song itself, as well as tonal language that does not stray too far from our corpus norm.

Raw Intensity scores fragments high when they sound noticeably more "aggressive" than those that do not.

Melodic Conventionality primarily describes the commonality and recurrence of the dispersion of melodic bigrams and the melody aligning with the harmony.

Conventionality of Intensity comprises corpus- and song-based second-order features for the most important components of Raw Intensity.

Regression

Lastly, we modelled a segment's representativeness with a log-linear regression with as independent variables the audio-derived features, the familiarity score, and the segmentation method.

Log-Linear Model

	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

Model 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.

Results

- A linear model based solely on the data of the original user study showed that familiarity had no significant impact on how participants ranked the segments ($b = -0.08$, $SE = 0.07$, $p = .27$).
- It was found that *Raw Intensity*, *Melodic Conventionality*, and *Conventionality of Intensity* are the most important factors to approximate a segment's worth, each having a positive effect.
- A positive effect of segmentation method on representativeness has not been found.

Proposed Thumbnailing Method

1. Obtain a reasonable amount of fragments from the song.
2. Compute the CATCHY features for each of these features.
3. Compute an approximation of the factors
4. Multiply these approximations with the estimates of the log-linear model.
5. Choose the fragments with the highest score as the music thumbnail.

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.

Contact

mail: anvannieuwenhuijsen@gmail.com