UDC: 004.93 https://doi.org/10.15407/jai2023.01.060

USING THUMBNAIL LENGTH BOUNDS TO IMPROVE AUDIO THUMBNAILING FOR BEATLES SONGS

D. Zasukha
International Research and Training Center for Information Technologies and Systems, Ukraine
40, Academician Hlushkov Ave., Kyiv, 03680
dmytro.zasukha@irtc.org.ua
https://orcid.org/0000-0003-1737-5522

Annotation. Optimising the parameters of the audio thumbnailing procedure can improve the final results. Previously, experiments with the thumbnail length parameter have shown strong potential to enhance thumbnail boundaries detection for Beatles songs. However, usage of the thumbnail length parameter has been limited to only changing the thumbnail length lower bound. The purpose is to use the thumbnail length upper bound in combination with the lower bound to improve thumbnail boundaries' detection for Beatles songs. I experiment with the thumbnail length upper bound while fixing the lower bound, then analyse the F-measure results based on segment boundaries. I use a thumbnail procedure with a repetition-based fitness measure as the foundation. The results demonstrate that the thumbnail length upper bound can increase an estimated thumbnail boundaries' accuracy for Beatles songs. I select a pair of lower and upper bounds that slightly improves the F-measure based on segment boundaries, unlike using only the lower bound. In conclusion, this study optimises the thumbnail length bounds to improve the audio thumbnailing procedure with a repetition-based fitness measure for Beatles songs. It is demonstrated that the upper bound can improve the F-measure if chosen correctly. Unexpectedly, the upper bound can be omitted without losing much in the accuracy of thumbnail boundaries' detection. Additionally, I indicate further directions to optimise thumbnail length bounds for popular music and its genres (like pop, rock). Also, I describe other supplemental tasks for future work.

Keywords: audio signal, music processing, music structure analysis, thumbnailing, thumbnail length, popular music.

Introduction

Music structure analysis is a multifaceted and often ill-defined problem that depends on many different factors. First, the problem's complexity depends on the analysed kind of music representation. For example, while detecting certain structures such as repeating melodies in sheet music is comparatively easy, it is often much harder to automatically recognise such structures in audio representations. Second, segmentation may be based on various principles including homogeneity, repetition, and novelty. Third, one must also account for different musical dimensions, such as melody, harmony, rhythm, or timbre. Finally, the segmentation and structure largely rely on the musical context and the considered temporal hierarchy [1].

In this study, I work on a well-known subproblem of music structure analysis known as audio thumbnailing. Given a music recording, the objective is to automatically choose the most representative section, which may serve as a kind of "preview", offering a listener a first impression of the song or piece of music. Among such previews, the user should be able to quickly determine if they would like to listen to the song or move on to the next one. Hence, audio thumbnails are vital browsing and navigation aid for finding interesting elements in large music collections [1]. In [2] it's stated that "thumbnail creation" is only one of two obvious commercial applications of music structure analysis. Audio thumbnailing was addressed in such works as [3-7].

In [3], as the main technical contribution, a new fitness measure was introduced, which assigns the value of this measure to each segment. It expresses to what extent and how the segment "explains" the repeating structure of the entire record. The thumbnail is then defined as the segment with the highest value of the fitness measure. The authors demonstrate various experiments based on different audio collections that include popular music (Beatles as a representative of popular music), classical music, and folk song recordings. For popular music, the F-measure based on segment boundaries is 0.76.

In the scientific article [4], continuing the work [3], the authors try to solve the problem of finding two candidates as a thumbnail for popular music, where both verse and chorus can be good candidates. As the main technical contribution, two approaches to computing double thumbnails are proposed, both of which extend the iteration-based thumbnailing procedure introduced in [3]. The MIREX evaluation measure for the extended approach is 0.74 with a theoretical maximum of 0.97 (with oracle information when identifying both verse, chorus and their repetitions perfectly). It was stated that the existing approaches approach the maximum result that can be obtained in practice.

Also, there are deep learning approaches for audio thumbnailing, some of them ([5-7]) are specialised in chorus detection of popular music.

However, tuning the thumbnail length parameter to improve the thumbnailing result has received little attention. Even though, it has shown huge potential to increase the result of F-measure based on segment boundaries in [3] experiments. Also, thumbnail length can be generalised according to pop music (pop music as a subset of popular music) industry standards for song segment duration [8].

Previous work's [3] usage of the thumbnail length parameter has been limited to investigating the thumbnail length role, and interdependencies with other parameters and to indicating conceptual benefits of the fitness measure. Therefore, specific parameter settings weren't advocated in the study. Also, the paper only investigates the role of thumbnail length's lower bound, omitting experiments with the upper bound.

The main focus of this study is optimising the expected thumbnail length bounds for Beatles music (which is representative of popular music). The thumbnail length parameter is one of the parameters that are a music signal property and directly translates to parameter settings.

To optimise the result of the F-measure, I will experiment with different values for the upper bound with a fixed lower bound of the expected audio thumbnail length, analyse the results and find optimal lower and upper bounds for the Beatles dataset.

The paper is organised as follows. First, I discuss the methods in the "Methods" section. Second, I provide results in the "Results" part. Finally, I conclude and talk about future steps in the "Conclusions and future work" segment.

Methods

The aim of this study is to optimise the expected thumbnail length bounds for Beatles music (a representative of popular music).

To better understand the directions of the research, it's important to explain the difference in terminology between popular music and pop music/genre/songs. As stated in [9]: "Although popular music sometimes is known as "pop music", the two terms are not interchangeable. Popular music is a generic term for a wide variety of genres of music that appeal to the tastes of a large segment of the population, whereas pop music usually refers to a specific musical genre within popular music".

I use the variation of the repetition-based audio thumbnailing approach using a custom fitness measure [3] and experiment with this approach. Specifically, in our procedure, I add the upper bound parameter θ.

I use a repetition-based approach [3] since it is intuitive, shows high results, and is easier to include the thumbnail length parameter to analyse and compare results.

In this paper as an evaluation measure, I will use the F-measure based on segment boundaries. This F-measure is one of two measures used in [3], which has a "soft nature". It expresses to what extent the estimated thumbnail agrees with one of the ground truth thumbnails contained in the ground truth thumbnail family (various possible ground truth thumbnails).

It's important to note that I couldn't receive as high F-measure values as in the original research due to issues with reproducibility, which are listed below.

To retrieve the next results, I use the libfmp python package [10], which provides implementations of well-established model-based algorithms for various MIR tasks. Those important for our task are methods for computing SSM ("compute_sm_from_filename" function), SSM normalization ("normalization_ properties_ssm"), fitness scale plot computation ("compute_fitness_scape_plot") and finding segment with maximal value in SP ("seg_max_sp"). For our task "seg_max_sp" was modified also to include lower and upper bounds. The following parameters were constant except mentioned further: threshold (thresh) parameter p = 0.2, lower bound (lower_bound) θ = 15 as in [3] work. Other parameters' values weren't mentioned in [3], so default parameters were used as in 1.2.3 libfmp version: length of smoothing filter (L) = 21, downsampling factor (H) = 16, length of filter (L_smooth) = 16, set of relative tempo values (tempo_rel_set) = list of 1, set of shift indices (shift_set) = list of 0, thresholding strategy (strategy) = "relative", whether to scale positive values (scale) = True, which specified value set for values below threshold (penalty) = 0.0, whether to binarize matrix (binarize) = False.

As a dataset, I use the Beatles dataset (as a representative of popular music) by Isophonics [11] of version 1.2 which is quite a coherent dataset. Also, there are a lot of numbers in the community for this dataset. Mirdata package was used to retrieve data and make work more convenient and standardised [12]. Some original annotations have incorrect end intervals which are less than the start interval for the last segment ("silence" label), so I replaced end intervals with start intervals in these cases. Those songs are: "Wild honey pie", "Sun king", "With a little help from my friends", "Glass onion, "Back in the USSR", and "Cry baby cry".

The authors of [3] mention that they removed 5 songs without clear repetitions from the dataset of 180 Beatles songs, but they didn't mention the exact songs. I found them as those that don't have any segment met for 2 or more times in music structure annotation. Those are "Happiness is a warm gun", "Revolution 9", "You never give me your money", "The end", and "Her majesty".

Additionally, to have a more general music structure and combine repetitive segments easier, I replaced some segment variations with a parent segment like replacing "verse_(instrumental)", "versea", "verseb", "verseguitar" etc with "verse". To support this decision, probably authors of [3] used some previous version of the Isophonics dataset (earlier than the 1.2 version) because the example of the Beatles' song "Birthday" mentioned in their work has labelling different from this song's annotation in Isophonics dataset 1.2 version ("V1", "V2", "V3" in the example vs "verse", "verse" "verse_(instrumental)" in the dataset).

Most audio files are different from the original ones (from which annotations were retrieved) due to complications in finding them. I mention all available text information for each Beatles' album I used: "Please please me" (The Parlaphone, 2014 Calderstone Productions Limited, 0602537825707 article, PMC1202, 5099963379815), "With the Beatles" (The Parlaphone, 2014 Calderstone Productions Limited, 0602537825714 article, PMC1206 5099963379914), "A Hard Day's Night" (The Parlaphone, 2014 Calderstone Productions Limited, 0602537825721 article, PMC1230, 5099963380019), "Beatles for Sale" (UK Tube Cut), "Help! (The Parlaphone, 2014 Calderstone Productions Limited, 0602537825745 article, PMC1255, 5099963380217), "Rubber Soul" (original), "Revolver" (UK Tube Cut), "Sgt. Pepper's Lonely Hearts Club Band" (The Beatles / Sgt. Pepper's Lonely Hearts Club Band [Disc 4: Mono Album and Bonus Tracks]), "Magical mystery tour" (cover is like Parlophone, PCTC 255, 1976, but not sure that it's this one), "The Beatles" (UK PBTHAL Vinyl Rip), "Abbey Road" (Parlophone, 2012, 5099963380910 article), "Let it be" (2012 E.M.I. Records, 094638247210 article). The files are 44.1 kHz FLAC.

There are many parameters which can be adjusted in the audio thumbnailing procedure [1, 11]:

1. Music dimension (chromogram, MFCC, tempogram) and its parameters like window size and hop length.
2. Feature smoothing parameters: smoothing length, downsampling, median vs average filtering, adaptive windowing.
3. Path enhancement parameters: filter length, set of tempo differences, backward

and/or forward smoothing, morphological operation, median vs average filtering, 2D convolution kernels vs resamble-based approach.

4. Thresholding parameters: thresholding strategy, value, scaling, penalty, binarization.

5. Transposition invariance. To use it or not.

6. Scale plot computation.

7. Evaluation measure choice i.e. MIREX music structure F-measure, F-measure based on segment boundaries.

8. Thumbnail detection approach i.e. repetition based.

9. Other parameters like minimum length of segments, and duration of music recording.

Among all parameters, I choose the length of the thumbnail as one of the parameters that are a music signal property and directly translate to parameter settings, so easier to apply and interpret. Also, experiments with thumbnail length in [3] have shown the potential to increase accuracy.

Except only experimenting with lower bound θ of thumbnail in seconds done by [3] (see Fig. 1), I also take upper bound θ.

To do it, I fix the lower bound θ to 15 as in [3] experiments. It's the value that maximizes the F-measure independent of a specific choice of threshold parameter. Then, I only change the upper bound θ from 15 (equal to lower bound θ) to 70 (long enough to show tendency).

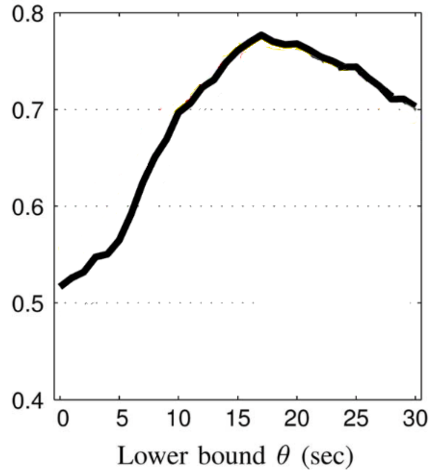

Fig. 1. Thumbnail F-measure values in dependency of different parameters. given in seconds). Taken from [3], only results with threshold p = 0.2 are left.The horizontal axis specifies the lower bound parameter θ (

Results

We can see (Fig. 2) that F-measure stops changing significantly after the upper bound θ equals 23 and stays quite flat after 28 seconds. From the upper bound θ equals 64 to 70, the values stop changing and are equal to the result without using the upper bound θ which equals 0.724. The max F-measure is slightly greater and equals 0.729 when upper_bound θ equals 48. It may be interesting to note that the longest ground truth family segment has a length of 67.9 seconds.

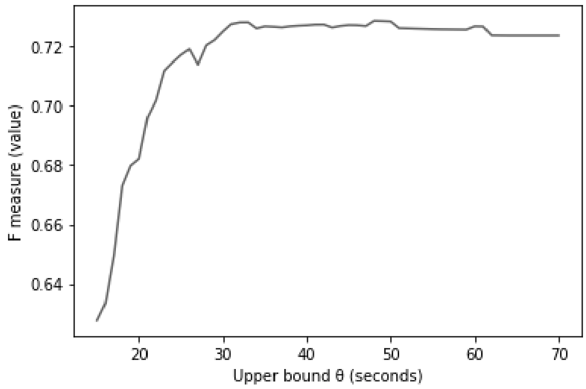

Fig. 2. Thumbnail F-measure values in dependency of the upper bound θ (given in seconds). Lower bound θ is fixed as 15 and threshold p = 0.2.

Finally, without both the lower bound θ and upper bound θ the result is significantly less and equals 0.65. In this case, threshold parameter p is libfmp "compute_sm_from_filename" function' default value and equals 0.15, which results in a higher F-measure than with threshold parameter p equals 0.2.

The important note is that exact values (optimal bounds, exact points of change) can differ for original audio files used in annotations because of duration differences. Nevertheless, it won't affect analysed tendencies.

Conclusions and future work

The results show that thumbnail length upper bound θ is important to capture so to say thumbnails, but after some value, it doesn't make much change. It makes sense since the average length of chorus and verse in pop songs (pop songs as a subset of popular music) can differ, but not significantly according to music industry standards (pop song duration, structure and ratio of segments' durations) [8]. Also, since the fitness measure used in [3, 4] is balanced for both coverage, and score and slightly favours shorter segments, there is a lower risk of under-segmentation. As a result, we don't see issues with high upper bounds. Without a lower bound θ, the result is much lower. This is explained in [3] when they state that using the lower bound θ allowed them to disregard path families that consist of many short spurious path fragments.

I showed that the expected length of the thumbnail parameter is important to improve the repetition-based audio thumbnailing procedure evaluated with the F-measure based on segment boundaries. To prove it I used the Beatles dataset (representative of popular music) and experimented with thumbnail length lower and upper bound parameters. I demonstrated that both upper and lower bounds affect F-measure, but the upper bound can be omitted without losing much in the accuracy of thumbnail boundaries' detection.

With results achieved, there is a chance to improve the used F-measure based on segment boundaries for all popular music and each genre of popular music (pop, rock etc.). My future work is to find suitable lower bound θ and upper bound θ for popular music and its genres. To achieve it, I need statistics on chorus or verse duration. As [4] emphasizes, for popular music, both verse and chorus sections may serve as suitable thumbnail candidates. Popular music songs can be taken as a subset of the One Million Song dataset [13] to get a wide variety of popular music songs with genre annotations available. To derive chorus/verse statistics, I will use Spotify API [14], where I can get the duration of each song section. Spotify has a vast library of over 80 million tracks. However, there is no label in the data. That's why I can't tell whether a section is a chorus or verse. To cope with it, I will assume that the second section is a verse. This is motivated in [3] by the fact that many songs start with an intro and then continue with a verse corresponding to the thumbnail. After retrieving durations, I will get appropriate min and max durations which can be set as lower bound θ and upper bound θ. To test annotated data with these bounds, I can use the same Beatles dataset [11] and others like Billboard (McGill), RWC popular, QMUL: Michael Jackson etc. [15].

After this improvement, I plan to test the modification of the end-to-end pipeline solution with an automatic genre retrieval [16] and/or move on to other genres where I can also try to apply the expected bound of thumbnail length and/or other parameters.

Additionally, I plan to try other evaluation measures, like the "harder" evaluation measure described in [3], the MIREX evaluation measure used in [4] or other evaluation measures like the ones discussed in [17]. Besides, I can implement the enhancement suggested in [4]. Also, I can try other approaches to music structure analysis like spectral clustering homogeneity-based approach [18]. Plus, thumbnail length may be used to improve other approaches to audio thumbnailing like [5-7]. Not to mention that other parameters can be tuned to improve the result.

Acknowledgements**:** I would like to thank Prof. Dr Meinard Müller for his research and educational (FMP Notebooks [1], libfmp [10] which are co-authored with Frank Zalkow) efforts in this field. Also, for giving comments and inspiring me to move in this direction.

References

1. Müller, M., & Zalkow, F. (2019). FMP Notebooks: Educational Material for Teaching and Learning Fundamentals Of Music Processing. In *ISMIR Conference* (pp. 573–580). Retrieved December 21, 2022, from https://www.audiolabs-erlangen.de/resources/MIR/FMP/data/C0/2019_MuellerZalkow_FMP_ISMIR.pdf.
2. Nieto, O., Mysore, G. J., Wang, C.-. i ., Smith, J. B. L., Schlüter, J., Grill, T., & McFee, B. (2020). Audio-Based Music Structure Analysis: Current Trends, Open Challenges, and Applications. Transactions of the International Society for Music Information Retrieval, 3(1), 246–263. http://doi.org/10.5334/tismir.54.
3. Muller, M., Jiang, N., & Grosche, P. (2013). A robust fitness measure for capturing repetitions in music recordings with applications to audio thumbnailing. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(3), 531–543. https://doi.org/10.1109/tasl.2012.2227732.
4. Jiang, N., & Muller, M. (2015). Estimating double thumbnails for Music Recordings. *2015 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP). https://doi.org/10.1109/icassp.2015.7177949.

5. He, Q., Sun, X., Yu, Y., & Li, W. (2022). Deepchorus: A hybrid model of multi-scale convolution and self-attention for chorus detection. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 411–415. https://doi.org/10.1109/icassp43922.2022.9746919.

6. Wang, J.-C., Smith, J. B. L., Chen, J., Song, X., & Wang, Y. (2021). Supervised chorus detection for popular music using convolutional neural network and multi-task learning. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 566–570. https://doi.org/10.1109/icassp39728.2021.9413773.

7. Huang, Y.-S., Chou, S.-Y., & Yang, Y.-H. (2017). Music thumbnailing via neural attention modeling of Music Emotion. *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 347–350. https://doi.org/10.1109/apsipa.2017.8282049.

8. Atlanta Institute of Music and Media. (2019, March 25). How to Structure a Pop Song [web log]. Retrieved December 22, 2022, from https://www.aimm.edu/blog/how-to-structure-a-pop-song.

9. Wikimedia Foundation. (2022, December 5). *Popular music*. Wikipedia. Retrieved December 22, 2022, from https://en.wikipedia.org/wiki/Popular_music.

10. Müller, M., & Zalkow, F. (2021). Libfmp: A python package for fundamentals of Music Processing. *Journal of Open Source Software*, *6*(63). https://doi.org/10.21105/joss.03326.

11. Mauch, M., Cannam, C., Davies, M. E. P., Dixon, S., Harte, C., Kolozali, S., Tidhar, D., & Sandler, M. (2009). OMRAS2 metadata project 2009. In *12th International Society for Music Information Retrieval Conference*. Retrieved December 21, 2022, from https://www.eecs.qmul.ac.uk/~simond/pub/2009/late-breaking-C4DM.pdf.

12. Magdalena Fuentes, Rachel Bittner, Marius Miron, Genís Plaja, Pedro Ramoneda, Vincent Lostanlen, David Rubinstein, Andreas Jansson, Thor Kell, Keunwoo Choi, Tom Xi, Kyungyun Lee, & Xavier Serra. (2021). mirdata v.0.3.0 (0.3.0). Zenodo. https://doi.org/10.5281/zenodo.4355859.

13. Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., & Lamere, P. (2011). The Million Song Dataset. In *International Society for Music Information Retrieval Conference (ISMIR 2011)* (pp. 591–596). Retrieved January 10, 2023, from https://doi.org/10.7916/D8NZ8J07.

14. Spotify. (2018, May 31). *Get Track's Audio Analysis*. Spotify for Developers. Retrieved January 10, 2023, from https://developer.spotify.com/web-api/get-audio-analysis.

15. Lerch, A., Balke, S., Sarmento, P., Rosenzweig, S., Humphrey, E. J., Porter, A., Ramires, A., Bogdanov, D., McLeod, A., Hawthorne, C. F., Baker, D. J., Miron, M., Stöter, F. R., Giraud, M., & Seetharaman, P. (2019, October 29). *ISMIR Datasets*. ISMIR. Retrieved January 10, 2023, from https://www.ismir.net/resources/datasets/.

16. Zasukha, D. (2021). Development of a method for increasing the accuracy of the basic formation algorithm of an informative, concise sound image for the description of musical works. In *Information technologies and automation* (pp. 308–309). Retrieved December 21, 2022, from https://card-file.ontu.edu.ua/ handle/123456789/18645.

17. Lukashevich, H. (2008). Towards quantitative 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008 (pp. 375–380). measures of evaluating song segmentation. In

18. Tralie, C. J., & McFee, B. (2019). Enhanced hierarchical music structure annotations via feature level similarity fusion. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 201–205. https://doi.org/10.1109/icassp.2019.8683492.

The article has been sent to the editors 21.10.22.
After processing 10.11.22.