Journal of the Audio Engineering Society
Vol. 65, No. 6, June 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2017.0011

Music Thumbnailing for Radio Podcasts: A Listener Evaluation

ADIB MEHRABI[1], CHRIS HARTE,[1] *AES Member*,
CHRIS BAUME,[2] *AES Associate Member*, AND SIMON DIXON[1]

[1]*School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK*
[2]*BBC R&D, London, UK*

When radio podcasts are produced from previously broadcast material, 30-second "thumbnails" of songs that featured in the original program are often included. Such thumbnails are made up of continuous or concatenated sections from a song and provide the audience with a summary of the music content. However, editing full-length songs down to representative thumbnails is a labor intensive process, particularly when concatenating multiple song sections. This presents an ideal application for automatic music editing tools and raises the question of how a piece of music is best summarized for this task. To gain insight into this problem we asked 120 listeners to rate the quality of thumbnails generated by eight methods (five automatic and three manual). When asked to judge overall editing quality (on a five point Likert scale) listeners gave higher ratings to methods where the edit points were quantized to bar positions, although we found no preference for structural content such as the chorus. Ratings for two automatic editing methods (one containing the chorus, one containing only the intro and outro) were not significantly different to their manual counterparts. This result suggests that the automatic editing methods applied here can be used to create production quality thumbnails.

1 INTRODUCTION

When a BBC radio show is broadcast, both audio and metadata are generated. This includes a stereo recording of the transmission (post mix processing, pre-broadcast processing) and metadata of start and end times for music items and presenter voiceovers. When a downloadable podcast is made from such a show, licensing agreements typically restrict the duration of each piece of commercial music to be less than 30 seconds, excluding sections of the music that contain presenter voiceover. It would therefore be useful to have a system that automatically performs this transfer of material from radio show to podcast, ensuring that any music included in the podcast-ready version is edited in line with the relevant licensing requirements. Fig. 1 shows an example of the workflow for a such a system.

This use case presents an interesting challenge in deciding which section (or sections) of a piece of music should be included in a 30-second clip. Previous work on music thumbnailing typically focuses on either finding a continuous representative section of music [1–7] or by concatenating multiple song segments together [8–11]. However, in radio shows it is very common for the presenter to talk over the start and end of a song. In this case it is only the section of "clean music" (without voiceover) that must be less than 30 seconds long. The thumbnail must also contain the sections of music immediately following and preceding the voiceover. These constraints present a unique use case for music thumbnailing, which differs from the typical music summary application. They also raise two interesting questions: First, do listeners have a preference for how a song is edited into a 30-second thumbnail for a podcast? Second, how does the quality of automatically made edits compare to like-for-like manual ones?

2 BACKGROUND

A number of approaches have been suggested for summarizing the material in music recordings, many of which use song structure to determine the most representative part of a piece of music (often taken as being the most repeated section). Foote [12] first proposed the use of similarity matrices for music structure analysis. This technique was extended to music summarization using Mel-frequency cepstral coefficients (MFCCs) [3], to create song summaries with lengths of 10, 20, and 30 seconds. Bartsch and Wakefield [2] later presented a similarity matrix based method for chorus

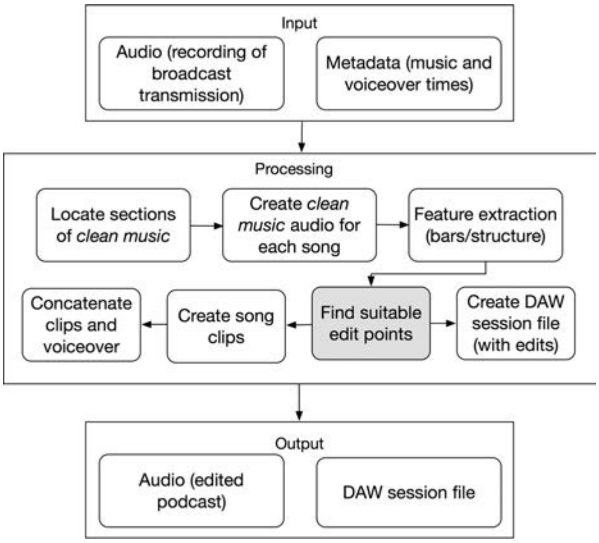
Input
Audio (recording of broadcast transmission)
Metadata (music and voiceover times)
Processing
Locate sections of *clean music*
Create *clean music* audio for each song
Feature extraction (bars/structure)
Concatenate clips and voiceover
Create song clips
Find suitable edit points
Create DAW session file (with edits)
Output
Audio (edited podcast)
DAW session file

Fig. 1. Processing workflow for an automatic podcast editor. Note: "clean music" is the section of a song between any sections of presenter voiceover. DAW = digital audio workstation. Here we are concerned with how to find suitable edit points (the grey processing block).

detection using chroma features. They compared chroma features to MFCCs and found chroma to give better results for the task (using 93 songs with hand-labelled annotations). Goto [5] further demonstrated the suitability of this method by using chroma features and similarity matrices to achieve 80% accuracy in a chorus detection task using a set of 100 test songs. Kelly et al. [11] used chroma based similarity matrices but also applied beat-synchronous segment boundaries to compile multi-section thumbnails from Irish folk songs. This task is somewhat similar to ours, where sections of a song need to be edited together into a summary. Using this method on a set of 30 songs they automatically generated thumbnails that included 83% of the "ideal" sections based on hand-labelled annotations. Several alternatives to similarity matrix based approaches including state-based methods using hidden Markov models (HMM) and support vector machines have also been applied to segmentation [1, 6, 8, 13].

User-centric evaluation is essential to establish the suitability of any music thumbnailing, or indeed any music information retrieval method that involves humans as listeners [14]. Various approaches have been used to evaluate listener preference for music thumbnailing techniques. In [1] listeners were asked to rate different automatic thumbnailing methods by giving a score of one, two or three for "poor," "average," or good" respectively and were also asked to provide subjective criteria for a "good" thumbnail. The responses indicated that it is desirable to have the song title in the summary, a vocal portion is better than an instrumental one, and it is preferable to start an edit at the beginning of a phrase. Xu et al. [6] asked listeners to rate automatically generated thumbnails using a five-point scale (low to high) for three criteria: clarity, conciseness, and coherence. Two automatic thumbnailing methods were compared and the average ratings across the criteria were then used to give a rating to each method. This evaluation method was also used in [7], although the authors only took ratings for one thumbnailing approach. We considered adopting this evaluation method for the present study, however the evaluation criteria are quite abstract and do not provide explicit information about the quality of the editing in terms of cut positions. This was not an issue for the above mentioned studies because they were only concerned with continuous thumbnails, which is not an option given our application.

Table 1. Summary of the editing methods used to create the clips.

Method	Type	Song Parts
A	auto	Intro, outro
B	auto	Intro, outro
C	auto	Intro, chorus, outro
D	auto	Intro, repeated segments, outro
E	auto	First and last 2 bars, every 4th bar
F	manual	Intro, outro
G	manual	Intro, chorus, outro
H	manual	BBC producer edits

To evaluate non-continuous thumbnails, Meintanis and Shipman asked listeners to choose one of four summaries that best represented a song [10]. There were three multi-phrase summaries and one made using only the song introduction. Ratings for the introduction summary were significantly lower than for the multi-phrase summaries, although there was no significant difference between the multi-phrase variants. Participants were also asked which parts of a song were important for becoming familiar with and recalling music, where the refrain and introduction scored highest for familiarity and recall respectively.

In summary, it is clear that similarity matrix based methods using chroma features show promising performance for the task of structural segmentation, and that the addition of beat/bar tracking can improve the rhythmic coherence of an edit point when concatenating sections of a song. Methods used to evaluate thumbnail quality vary and are determined by the intended use of the thumbnail: most of the work reviewed here is concerned with a thumbnail as a music summary. In such studies, listeners are typically asked to provide ratings for different aspects of the thumbnail. To the authors' knowledge no studies have specifically looked at listener preference for music thumbnails in a radio podcast.

3 METHODOLOGY

3.1 Editing Methods

In this study we asked participants to rate eight different editing methods, labelled A–H (see Table 1 for a summary of each method). Methods A–E are automatic and F–H are manual. Of the five automatic methods, four use bar position information as determined by a bar tracking (BT) algorithm [15]. The algorithm first computes the difference between all band limited (>1.4 kHz) beat-synchronous spectral frames in a song. Bar positions are then selected

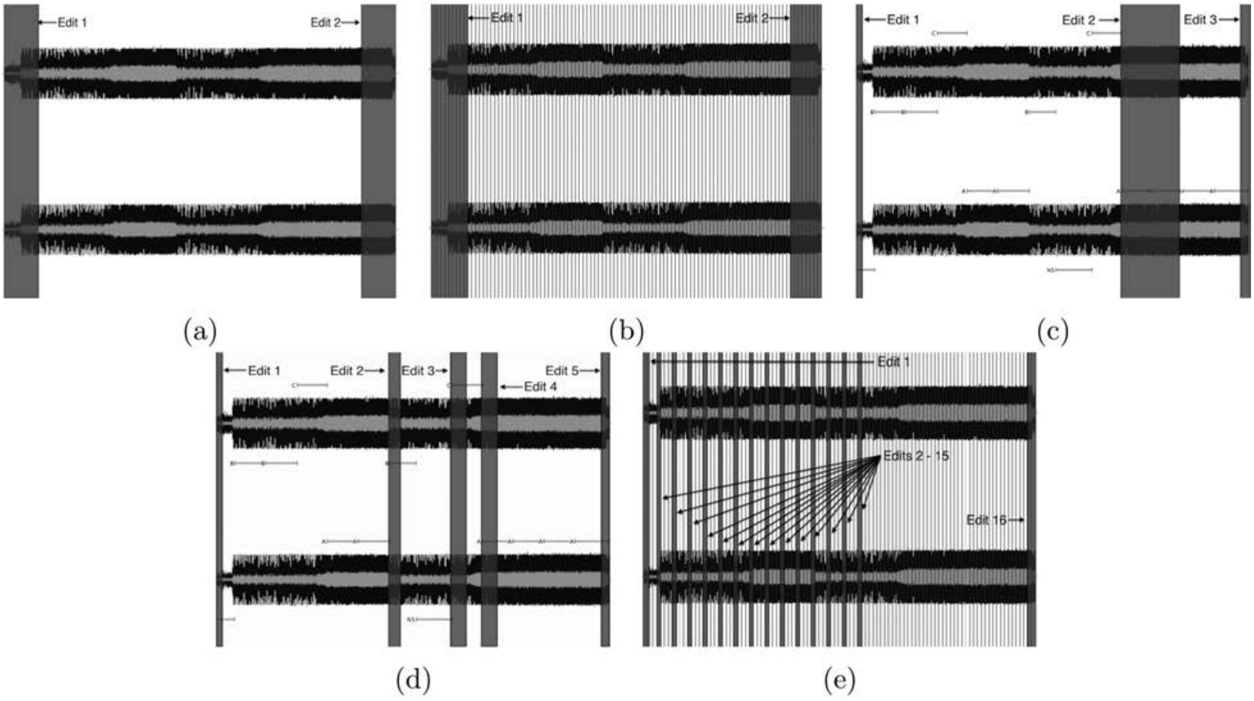

Fig. 2. Example of editing methods A–E (Figures a–e respectively) applied to the same song. The shaded sections are the edits that are concatenated to make the clip. The light grey vertical lines in (b) and (e) denote bar positions.

as the beat transitions that give the highest spectral change, assuming a given meter.

The structural segmentation used for methods C and D is performed using an algorithm taken from [16]. This was selected as the state of the art in music structural segmentation (it performed best in three of the four datasets in the 2015 MIREX [17] challenge for structure segmentation). To extract segments, the Pearson correlation coefficient (r) is computed in a pairwise manner between chroma vectors for every beat, creating a beat-wise similarity matrix. Candidate segments are then found by searching the diagonal of the matrix for repeated sequences of beats between 28–128 beats long, where $r > 0.65$ for the first beat in the set. Repeated sets are selected based on a threshold for the amount of repetition (as repeated segments will not be 100% identical for certain music styles). Both the bar tracking and segmentation algorithms are implemented as *Vamp* plugins for the *Sonic Annotator* host [18].

Manual methods F and G were created by one of the authors (an experienced audio editor), with the specification of being manual counterparts to methods B and C respectively. Method H was performed by a member of the production team at BBC Radio who regularly edits podcasts from radio show content.

3.1.1 Automatic Methods

In the five automatic methods a thumbnail is created by reducing a song length to 30 seconds or less, by removing sections from the middle of the song. The first and last two bars of each song are kept in the final clip to ensure a smooth transition from the presenter voiceover into the song, therefore the approaches only differ in the content that is included between these sections. The crossfades between all the edits are 0.5 s long.

Method A: This simple approach is used as baseline to which the other automatic methods may be compared. The musical content is not considered and the aim is to optimize the clip length to 30 seconds. The first edit is made 15 seconds into the "clean music" and a second edit is made 15 seconds from the end. An example of the edits used to make the clip can be seen in Fig. 2a. The middle section of music between the edits is removed and the two sections are concatenated.

Method B: This is similar to method A in terms of the parts of the song that make up the clip, however the edit positions are quantized to bar positions as determined by the BT algorithm [15]. The purpose of the quantization is to maintain rhythmic continuity between the sections. Fig. 2b shows an example of this method applied to a song.

Method C: In this method three edits are made: two bars into the song; at the start and end of the chorus; and two bars from the song end. These are concatenated to make the clip. The BT algorithm from method B is used to locate bar positions and the segmentation algorithm [16] is used to determine the song segments. After the segments have been identified, we apply chorus detection (CD) by selecting the single highest energy (RMS) repeated segment as the most likely candidate for the chorus. If the clip is too long then the latter section of the chorus is truncated on a bar position to make the clip fit within the 30 second restriction. This can be seen in Fig. 2c.

Method D: This method provides an overview of the segments in the song. When song segments are extracted as per method C, there are often repeated segments with the same segment label (typically representing verses and choruses). In this case we select one version of each repeated segment, based on the highest RMS values. These are used to create the clip. Each segment is truncated so the clip fits within the 30-second limit, so they will typically only be a few bars long. This can be seen in Fig. 2d.

Method E: This provides an overview of the song by creating a clip made up of one-bar segments from every four bars (assuming a 4/4 meter). As many bars as possible are included up to the 30-second limit. An example of this applied to a song can be seen in Fig. 2e.

3.1.2 Manual Methods

Method F: This is the manual counterpart to methods A and B. The editor was instructed to create a musically coherent 30 second clip by making two edits (thus creating a clip with only intro and outro). The position of the edits were left at the discretion of the editor, and the proportion of intro to outro was not defined.

Method G: This is the manual counterpart to method C. The editor was instructed to create a musically coherent clip that included the first and last bars of the song and as much of the chorus as possible.

Method H: The radio producer was asked to create a 30-second clip for each song, ensuring the first and last two bars were included; the content between these sections was left at the discretion of the producer. The clips were edited as part of the producers' normal working day, therefore the editing was subjected to typical real-world time constraints (approximately one hour was spent editing 10 songs).

3.2 Test Songs

BBC Radio provided test data of four radio shows from two of the target radio stations for downloadable podcasts (Radio 2 and Radio 6 Music) that contained 88 commercial songs. Editing methods B–E rely on the accuracy of the BT and CD algorithms, so we conducted a preliminary study to establish the success of these algorithms when applied to each of the songs. This analysis was performed manually by one of the authors. The BT algorithm was only considered successful if the position of every bar in the song was accurately detected, and the CD algorithm was considered successful if the chorus occurred within the detected segment.

A representative sample of 10 songs was then selected for the study (shown in Table 2). These included a balanced number from each radio station (five from each) and a range of pop music sub-genres, which is of specific interest for our application. These songs represent the proportion of the 88 songs where the BT and CD algorithms were successful (approximately 60% for BT, 60% for CD, and 40% where both algorithms worked). The inclusion of test songs where the BT and CD algorithms fail allows us to establish if the success of these algorithms has any effect on the perceived editing quality for the automatic methods.

Table 2. Songs used for the user study with indicators for whether the bar tracking (BT) and chorus detection (CD) algorithms were successful.

Artist	Song	BT	CD
George Ezra	Budapest		
Laura Marling	Where can I go?		√
R.E.M.	Nightswimming		
Ed Sheeran	Sing		√
Wah!	Story of the Blues	√	√
Earth, Wind and Fire	September	√	√
Mapei	Don't wait	√	
Friendly Fires	Paris	√	√
Mansun	Take it easy chicken	√	
Bob Marley	Could you be loved?	√	√

The number of songs used in this study is comparable to previous listener-based thumbnail evaluation studies. In [6] 10 pop/rock songs were used, and in [7] 10 songs were used ranging in genre from pop to classical. It is also a small enough number of songs to allow for the study to be completed in a reasonable time (approximately 30 minutes), which enabled us to recruit a substantial number (128) of participants. Sections of voiceover/music overlap were removed from the songs, so the start and ends times are from when the presenter stopped talking to when they started again. The editing methods detailed in Sec. 3.1 were applied to the 10 test songs, resulting in 80 audio clips with a duration ≤ 30 seconds.

3.3 Listening Study

The listening study was conducted online. The participants provided basic demographic information regarding age, gender, and listening habits. They were then presented with a single web page containing eight randomized clips from a single randomly selected song to evaluate and were instructed to listen to all the clips before providing any responses.

Listening tests for music evaluation are typically concerned with comparing the quality of two or more versions of related material, which can be auditioned by switching between the versions (e.g., the MUSHRA[1] standard for evaluating audio quality). However, this type of evaluation method is not suitable for the present study because the summary clips are 30 seconds in duration and the edits need to be heard in full before listeners can make a judgment on quality. For this reason, we adopted a similar method to that in [6, 7, 10], where participants listened to all the summaries in full before providing responses. We asked the participants to provide ratings in response to the three prompts below, on a Likert scale from one (low) to five (high) for each clip.

1. Please rate the quality of the editing in this clip regarding the transitions between parts.
2. Please rate the quality of the editing in this clip regarding the selection of song parts (chorus, verse, etc.).

[1] https://www.itu.int/rec/R-REC-BS.1534/en

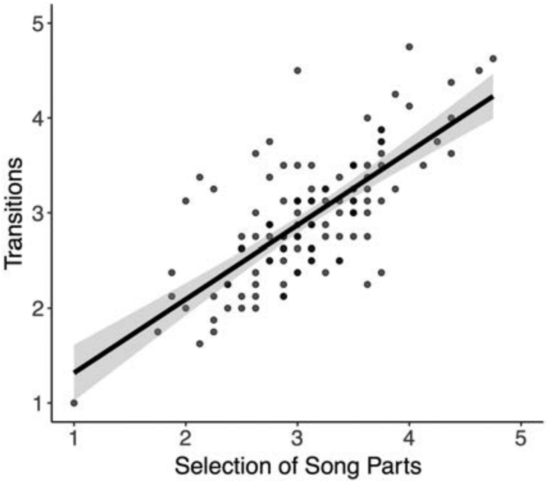

Fig. 3. Relationship between quality ratings with respect to transitions and the selection of song parts. Units on both axis are the raw ratings from 1–5.

3. Please rate the quality of the editing in this clip regarding the overall quality (considering the two criteria above)

3.4 Participants

In total 128 participants completed the study. The 8 most recent submissions were removed to balance the responses across the songs, leaving 12 participants per song. The participants were recruited by convenience sampling, mostly through academic and industry networks. Of the 120 participants used in the analysis 70% were male, 28.3% were female, and 1.6% chose not to report their sex. 26.7% never listen to podcasts, 50% listen to podcasts sometimes, and 23.3% frequently listen to podcasts. In terms of age, 64.2% were 18–35, 34.2% were 36–55, and 1.6% were over the age of 56.

4 RESULTS AND DISCUSSION

4.1 What Makes a Good Edit?

We observed a strong correlation (Spearman's rho) between the responses for each of the three rating criteria: overall quality and transition quality ($r_s(118) = .82$, $p < 0.01$); overall quality and selection of song parts ($r_s(118) = .84$, $p < 0.01$); transition quality and the selection of song parts ($r_s(118) = .66$, $p < 0.01$). The correlation between ratings for transition quality and the selection of song parts (Fig. 3) is smaller than between each of these factors and overall quality, however this is to be expected because we asked the participants to rate overall quality in terms of both of these factors. This shows that listeners did discriminate between the rating criteria to some extent, however in general both transition quality and the selection of song parts appear to similarly contribute to overall quality, therefore this can be taken as a measure that incorporates both factors of interest.

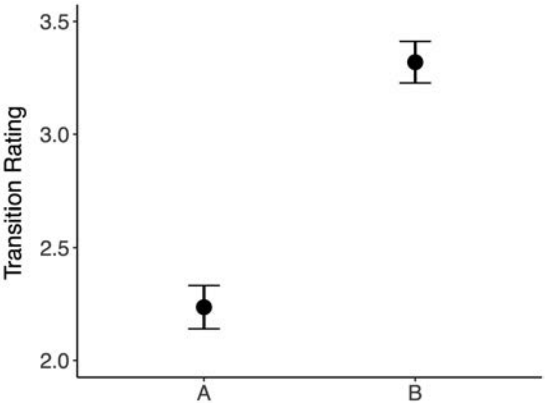

Fig. 4. Transition ratings for methods A and B (without and with bar quantization respectively). Error bars indicate standard error.

4.2 Does Bar Quantization Matter?

To determine whether bar quantization is an important factor we compared the transition quality ratings for methods A and B. The resultant clips contain the same parts of the song and it is only the transitions that differ. This comparison only considers like-for-like edits with and without bar quantization, therefore the four songs where BT did not work were excluded from the analysis. The transition ratings for these two methods are presented in Fig. 4. A Wilcoxon signed rank test with the Pratt method was used to determine the significance of the differences between the ratings.

The effect size for this is large and the difference is statistically significant ($r = 0.58$, $p < 0.01$). This indicates that bar quantization has a positive effect on the perceived quality of transitions. When running the same analysis on the songs where the BT algorithm did not work the difference is still significant ($r = 0.47$, $p < 0.01$), however the effect size is slightly smaller. This indicates that even when BT does not work it still performs better than the non-quantized approach. This is probably due to the nature of failures in the BT algorithm: if the detected bar positions are incorrect they still tend to fall on a valid beat or half measure.

4.3 Comparing the Overall Ratings for All Edits

We have shown in Sec. 4.1 that listeners tended to give similar ratings for all three criteria (overall quality, transition quality, and selection of song parts), so here we focus only on the overall ratings to make comparisons between all eight editing methods (Fig. 5). Clips generated using methods A and E were rated lowest for the automatic methods regardless of whether the bar tracking was successful. Method H was rated lowest among the manual methods for all songs and even lower still for songs where the automatic editing was successful. Methods B, C, F, and G performed best (with similar ratings) and the ratings for method D were slightly lower than these.

The overall ratings were submitted to a Friedman rank sum test — first for all songs and then for only the songs where the automatic editing was successful. The results show that the differences between both sets of ratings are

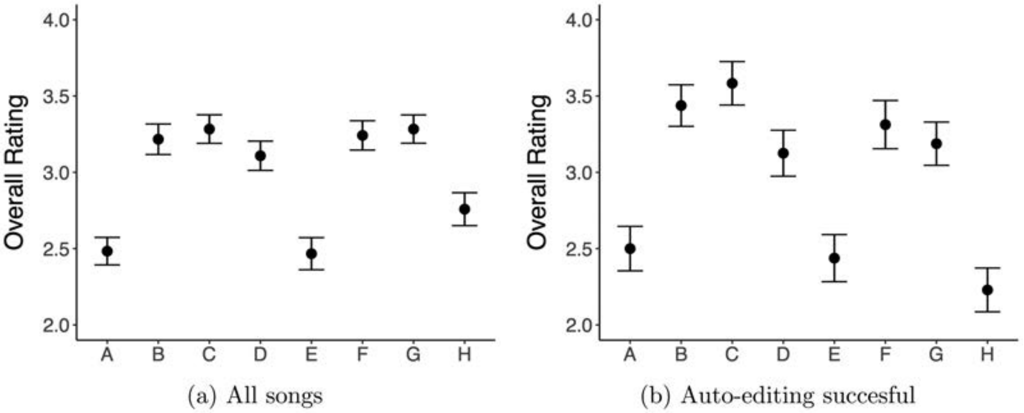

Fig. 5. Overall ratings for each editing method. Error bars indicate standard error.

significant: ($\chi^2(7, N = 120) = 106.06$, $p < 0.01$) for all songs and ($\chi^2(7, N = 48) = 68.83$, $p < 0.01$) for only the songs where the automatic editing was successful. A post-hoc analysis (Wilcoxon signed rank test) was performed on the ratings for both sets of songs, the results of which are presented in Fig. 6.

Considering the ratings for all the songs (Figs. 5a and 6a), the differences between the low scoring (A and E) and higher scoring (B, C, D, F, and G) methods are significant and driving the significant results from the Friedman test. The low ratings for methods A and E are likely due to the fact that edits do not tend to occur exactly at the start of phrases: this has previously been identified as a potential reason for not preferring music thumbnails [1]. The ratings for method H are also significantly lower than the three highest scoring methods (C, F, and G) but are not significantly different to B and D.

When considering only songs where the automatic editing methods were successful (Figs. 5b and 6b), the difference between the ratings for method H and methods B, C, D, F, and G is significant, due to method H receiving lower ratings and methods B and C being rated higher for this subset of songs. The higher ratings for methods B and C indicate a positive effect of the BT and CD algorithms working. The clips from method H do not tend to have edits that fall perfectly on beat or bar positions (this is not the case for clips made using the other methods except method A). This highlights the preference for edits to be made on bar or beat positions as discussed in Sec. 4.2.

There is also no preference for different multi-phrase summaries (B, C, D), as has been previously shown in [10], however there is clearly a limit to the number of phrases or parts that a clip can contain before is it disliked, as we can see from the ratings for method E. Interestingly, the similar ratings for methods B and C, and F and G indicate two main findings: that automatic edits, if performed well, are rated similarly to their manual counterparts, and that there is no significant preference for clips made up of intro and outro sections compared to clips that contain the chorus from a song.

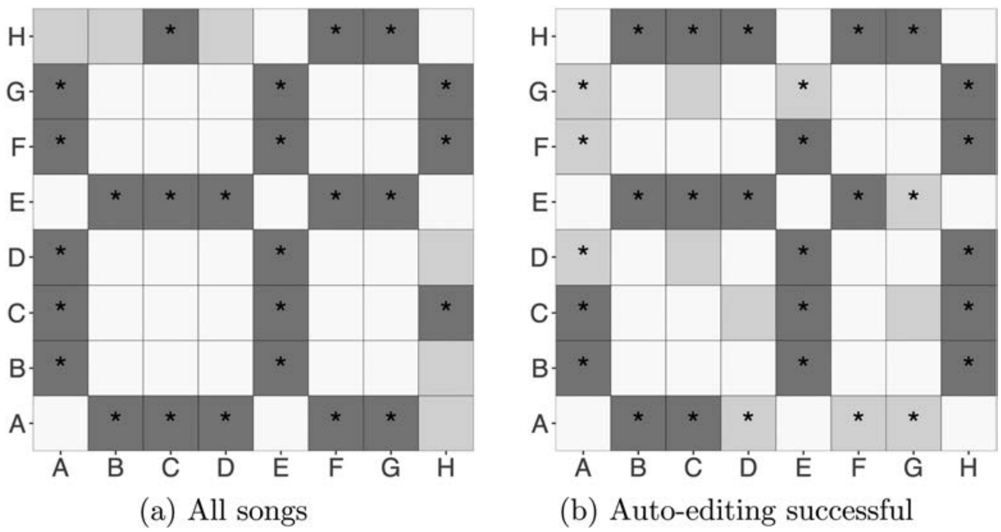

Fig. 6. Results of pairwise comparisons between overall ratings for all methods. Effect sizes as described by Cohen [19] are given by shading where dark = large, light = medium, white = small. Significant differences between methods are indicated by * where $p_{adj} <$ 0.05 (with Bonferroni correction).

5 CONCLUSION

We have conducted a listening study to compare the quality of eight music editing methods applied to radio podcasts. Participants were asked to rate the editing methods based on the song part selection and transition quality in the edited clips, as well as the perceived overall quality. The results suggest that listeners found both song part selection and transition quality to be similarly important when judging the overall quality of the edited clip.

When comparing like-for-like automatic methods with and without bar tracking (A and B) we found that participants rated the clips with bar tracking significantly higher. This suggests that rhythm-informed edit point selection is an important factor in perceived edit quality. The algorithms for methods B, C, and D rely upon successful bar tracking, and method C relies upon successful segmentation and chorus detection. Listeners gave slightly (albeit not significantly) higher quality ratings for these methods when applied to songs where both the bar tracking and chorus detection algorithms were successful. This indicates that by improving the generalizability of these algorithms we can expect to see a stronger preference for the automatic editing methods across a wider range of songs.

We found no significant difference in the overall ratings for the highest rated automatic editing methods (B, C, and D) and the highest rated manual ones (F and G). Given that the performance of these automated methods is comparable to that of a human editor, there is a good case for further research into improving these algorithms and deploying them within radio podcast production systems. It should be noted that the highest rated methods significantly outperformed manual method H, which is probably due to the radio producer working to a tight schedule (methods F and G were edited without such time pressure). Such time constraints are typical of a radio production environment, and this result only serves to reinforce the case for automated editing to be developed further.

There is also no significant difference between ratings for methods B and C: method B cuts straight from the intro to the outro on a bar position while C includes the chorus in the edit. Without clear user preference, it is difficult to choose a single algorithm to use for the proposed application. Future work may focus on listener preference between only these two methods, using a larger corpus of test songs. This could include the study of preferences for sub-groups of the population of listeners: of particular interest is the effect of listening to podcasts and time spent listening to music on the responses.

The results presented here are encouraging for future work within the automatic song editing/music thumbnailing community. Further research is required to determine which method to adopt in a “one size fits all” automatic podcast editing system. It may be that listeners’ opinion is split between wanting to hear the chorus or not in the 30-second time limit, or that the song has a large effect on this. Furthermore, the editing approaches may be so dependent on the song that some form of pre-processing analysis (for genre or mood) may be appropriate to dynamically select the editing method most appropriate to the song.

6 ACKNOWLEDGMENTS

The authors thank Frank Melchior, Zia Mehrabi, and the reviewers for their invaluable comments. This work was supported by the Media and Arts Technology programme, EPSRC Doctoral Training Centre EP/G03723X/1, and BBC Research and Development.

7 REFERENCES

[1] B. Logan and S. Chu, “Music Summarization Using Key Phrases,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. II749–II752 (IEEE, 2000). https://doi.org/10.1109/ICASSP.2000.859068

[2] M. A. Bartsch and G. H. Wakefield, “To Catch a Chorus: Using Chroma-Based Representations for Audio Thumbnailing,” *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (New Platz, New York, USA), pp. 15–18 (IEEE, 2001). https://doi.org/10.1109/ASPAA.2001.969531

[3] M. L. Cooper and J. Foote, “Automatic Music Summarization via Similarity Analysis,” *Proceedings of the 3rd International Conference on Music Information Retrieval* (Paris, France) (Ircam - Centre Pompidou, 2002).

[4] L. Lu and H.-J. Zhang, “Automated Extraction of Music Snippets,” *Proceedings of the Eleventh ACM International Conference on Multimedia* (Berkeley, CA, USA), pp. 140–147 (ACM, 2003). https://doi.org/10.1145/957013.957043

[5] M. Goto, “A Chorus-Section Detecting Method for Musical Audio Signals,” *Proccedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 5 (Hong Kong, China), pp. V–437 (IEEE, 2003). https://doi.org/10.1109/ICASSP.2003.1200000

[6] C. Xu, M. Maddage, and X. Shao, “Automatic Music Classification and Summarization,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 441–450 (2005).

[7] J. Głaczyński and E. Łukasik, “Automatic Music Summarization. A ‘Thumbnail’ Approach,” *Archives of Acoustics*, vol. 36, no. 2, pp. 297–309 (2011).

[8] G. Peeters, A. La Burthe, and X. Rodet, “Toward Automatic Music Audio Summary Generation from Signal Analysis.,” *Proceedings of the 3rd International Conference on Music Information Retrieval* (Paris, France), pp. 94–100 (Ircam - Centre Pompidou, 2002).

[9] W. Chai and B. Vercoe, “Music Thumbnailing via Structural Analysis,” *Proceedings of the Eleventh ACM International Conference on Multimedia*, pp. 223–226 (ACM, 2003). https://doi.org/10.1145/957013.957057

[10] K. A. Meintanis and F. M. Shipman III, “Creating and Evaluating Multi-Phrase Music Summaries,” *Proceedings of the 9th International Conference on Music Information Retrieval* (Philadelphia, USA), pp. 507–512 (Drexel University, 2008).

[11] C. Kelly, M. Gainza, D. Dorran, and E. Coyle, "Audio Thumbnail Generation of Irish Traditional Music," *Proceedings of the Irish Signals and Systems Conference* (Cork, Ireland) (IET, 2010). https://doi.org/10.1049/cp.2010.0504

[12] J. Foote, "Automatic Audio Segmentation Using a Measure of Audio Novelty," *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 1 (New York, New York, USA), pp. 452–455 (IEEE, 2000). https://doi.org/10.1109/ICME.2000.869637

[13] M. Levy, M. Sandler, and M. Casey, "Extraction of High-Level Musical Structure from Audio Data and its Application to Thumbnail Generation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing* (Toulouse, France), pp. 1433–1436 (2006). https://doi.org/10.1109/ICASSP.2006.1661200

[14] M. Schedl, A. Flexer, and J. Urbano, "The Neglected User in Music Information Retrieval Research," *J. Intel. Info. Sys.*, vol. 41, no. 3, pp. 523–539 (2013). https://doi.org/10.1007/s10844-013-0247-6

[15] M. Davies and M. Plumbley, "A Spectral Difference Approach to Extracting Downbeats in Musical Audio," *Proceedings of the 14th European Signal Processing Conference* (Florence, Italy) (EURASIP, 2006).

[16] M. Mauch, K. Noland, and S. Dixon, "Using Musical Structure to Enhance Automatic Chord Transcription," *Proceedings of the 10th International Conference on Music Information Retrieval* (Kobe, Japan), pp. 231–236 (2009).

[17] J. S. Downie, "The Music Information Retrieval Evaluation Exchange (2005–2007): A Window into Music Information Retrieval Research," *Acoust. Sci. & Tech.*, vol. 29, no. 4, pp. 247–255 (2008).

[18] C. Cannam, M. O. Jewell, C. Rhodes, M. Sandler, and M. d'Inverno, "Linked Data and You: Bringing Music Research Software into the Semantic Web," *J. New Music Res.*, vol. 39, no. 4, pp. 313–325 (2010). https://doi.org/10.1080/09298215.2010.522715

[19] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences* (Hillsdale, NJ, USA: Routledge Academic, 1998), 2nd ed..

THE AUTHORS

Adib Mehrabi

Chris Harte

Chris Baume

Simon Dixon

Adib Mehrabi is a Ph.D. candidate at the Centre for Digital Music at Queen Mary University of London, under the supervision of Dr. Simon Dixon. He received his B.Sc. degree in audio and music technology from the University of the West of England (UWE), England, in 2012, graduating with first class honors. His current research interests are in vocal imitation of musical sounds, query by vocalization, timbre perception, and modelling sound similarity.

•

Chris Harte graduated with M.Eng. first class honors in electronics and music technology from the University of York in 2001 and received a Ph.D. from Queen Mary University of London in 2010 for a dissertation on automatic audio harmony analysis. He worked as a lecturer at Queen Mary's Centre for Digital Music and then at York before moving to industry in 2016. He is now the Head of Research and Development at Scored Ltd. and also holds a visiting industrial lectureship at Queen Mary University of London.

•

Chris Baume is a Senior Research Engineer at BBC R&D in London and the BBC lead for the Orpheus project. His research interests span a number of areas including object-based broadcasting, semantic audio analysis, interaction design, and spatial audio. He is currently developing an object-based audio production system as part of Orpheus, and semantic audio production tools as part of his Ph.D. research at the University of Surrey. Chris is a Chartered Engineer and a member of the BBC's audio research group where he leads the production tools research.

•

Simon Dixon is a Reader (Assoc. Prof.), Director of Graduate Studies and Deputy Director of the Centre for Digital Music at Queen Mary University of London. He has a Ph.D. in computer science (Sydney) and L.Mus.A. diploma in classical guitar. His research interests include high-level music signal analysis, computational modelling of musical knowledge, and the study of musical performance. Particular areas of focus include automatic music transcription, beat tracking, audio alignment, and analysis of intonation and temperament. He was President (2014–15) of the International Society for Music Information Retrieval (ISMIR), is member of the Editorial Board of the *Journal of New Music Research* (since 2011), and has published over 150 refereed papers in the area of music informatics.