RE-VISITING THE MUSIC SEGMENTATION PROBLEM WITH CROWDSOURCING

Cheng-i Wang
UCSD
chw160@ucsd.edu

Gautham J. Mysore
Adobe Research
gmysore@adobe.com

Shlomo Dubnov
UCSD
sdubnov@ucsd.edu

ABSTRACT

Identifying boundaries in music structural segmentation is a well studied music information retrieval problem. The goal is to develop algorithms that automatically identify segmenting time points in music that closely matches human annotated data. The annotation itself is challenging due to its subjective nature, such as the degree of change that constitutes a boundary, the location of such boundaries, and whether a boundary should be assigned to a single time frame or a range of frames. Existing datasets have been annotated by small number of experts and the annotators tend to be constrained to specific definitions of segmentation boundaries. In this paper, we re-examine the annotation problem. We crowdsource the problem to a large number of annotators and present an analysis of the results. Our preliminary study suggests that although there is a correlation to existing datasets, this form of annotations reveals additional information such as stronger vs. weaker boundaries, gradual vs. sudden boundaries, and the difference in perception of boundaries between musicians and non-musicians. The study suggests that it could be worth re-defining certain aspects of the boundary identification in music structural segmentation problem with a broader definition.

1. INTRODUCTION

Music segmentation has been a fundamental task in automatic music content analysis. The task includes detecting boundaries between contiguous segments and labeling each detected segment within a music piece. In order to evaluate this task and train supervised machine learning algorithms, researchers have developed datasets that contain boundary timing and segment labeling annotations. In the majority of these datasets (such as *Beatles-TUT* [11], *CHARM Mazurka* [3] and *Beatles-ISO* [7], boundary timings are annotated by music experts and are defined as the time points that separate a music piece into non-overlapped contiguous sections representing meaningful song structures. These annotations provide clean-cut data for developing algorithms. During evaluation, the metrics are, in short, a measure of how close the automatically detected boundaries are to the ground truth annotations [6, 14].

Nevertheless, the instructions and rules for expert annotators to annotate these boundaries differs between datasets and the annotations inevitably conform to the subjective judgments of the annotators [13]. Moreover, the "one time point for one boundary" definition prevents the concept of short ambiguous/transitional/developing musical regions to be explored by researchers. To be more specific, it is established that different listeners will disagree on whether certain boundaries should exist or not in a music piece, and the saliences between boundaries might be different, while almost none of the existing datasets provide information about these intuitions [1, 13].

In [13], the issues of inter-annotator disagreement and the lack of multiple level annotations are discussed. The creation of the *SALAMI* dataset then attempts to solve such issues by having two versions of labels (by two annotators) and two levels (long and short time scale) of annotations in part of the dataset. The *SPAM* dataset also has five annotators with two levels of annotation for 50 songs [9]. In [8, 10], the issue of lacking support for hierarchical segmentation is discussed. Although the evaluation metrics for hierarchical segmentation problems are proposed in [8], only two datasets having two levels of hierarchy currently support this concept.

Apart from the inter-annotator agreement and between-boundaries saliency issues, the problem of not being able to model different types of boundaries is also an issue due to the current format of annotations. Sometimes the disagreements between annotators are not about whether one boundary should exist or not, but rather the timing of that boundary. The disagreement is likely because the exact change point or boundary between two larger segments is difficult to recognize if there are smaller transitional, pivotal, building, fading, or developing musical region connecting these two segments.

One of the major reasons for these limitations in existing datasets is the large amount of time and effort required by music experts to annotate songs. It tends to prevent datasets from having numerous (more than 5) annotators, and also tends to prevent detailed annotations for each song. One can alleviate the amount of effort required for annotating segmentation boundaries by crowdsourcing such task to the web. To the best of our knowledge, no methodology has been proposed utilizing crowdsourcing

Dataset	Song Name	Artist
Beatles-TUT	All You Need is Love Help! Here, There and Everywhere Strawberry Fields Forever Come Together(∗)	The Beatles
SALAMI	Smoke Machines You Done Me Wrong Out in the Cold Black or White We will Rock You(∗)	Atom Orr Cindy Woolf Carole King Michael Jackson Queen

Table 1. Song lists of the subsets from *Beatles-TUT* and *SALAMI*. Songs followed by asterisks are the hidden reference songs during the task.

to collect music segmentation boundaries. Crowdsourcing with Amazon's Mechanical Turks has been used to collect music similarity [4], music mood [5] data collection and audio sound quality evaluation [2].

In this paper, we present an preliminary study to support the above observations and address the above issues. We believe that this could lead to the creation of richer datasets with significantly less effort than previously required. In order to investigate the inter-annotator, between-boundary saliency problems and explore different types of segmentation boundaries, we used small subsets from existing datasets and annotate these via crowdsourcing. The results are a collection of annotations from at least 53 annotators (with at least 6 annotators annotated each song in full coverage) for each song for a total of 8 songs. The methodology of collecting annotations via crowdsourcing is described in section 2. The validation and analysis of the collected annotations are elaborated in section 3. Conclusions, and proposed future works are in section 4.

2. CROWDSOURCING

To perform the music segmentation boundary annotation collection task on the web, we use Amazon's Mechanical Turk (AMT). The proposed methodology is implemented as an extension of the **CAQE** (The Crowdsourced Audio Quality Evaluation Toolkit [1]) python package [2].

2.1 Data

Two subsets of songs from two music segmentation datasets were selected randomly to be annotated by the proposed methodology. The two datasets are the *Beatles-TUT* and *SALAMI* datasets. From each dataset, 5 songs were randomly selected. Among the 5 songs, 1 song is used as the hidden reference during the collection task to determine whether to accept or reject a person's annotations. Table 1 shows the 5 songs from both datasets.

2.2 Task Design

The typical off-line annotation process by music experts is to let them listen to a whole song and then annotate boundaries. This allows them to fine tune their annotations without time constraints. It also typically necessitates familiarity with an audio editing software package. Annotators on AMT, on the other hand, typically spend less time on such a task since their payment is fixed for a given task. They also typically have less (or no) experience with audio editing software packages. Therefore, the annotation process needs to be redesigned to accommodate it as an AMT based task. There are two goals of the redesign. The first goal is to simplify the annotation process so that AMT annotators could learn how to annotate quickly and repeat the process easily. The second goal is to maintain the quality of the annotations so that the results are informative and usable.

Since music structural segmentation is subjective in nature, we aim to not bias AMT annotators toward listening to specific musical cues. Therefore, the working definition used in the description of the task is kept as concise as possible and no musical terms are used. It is as follows:

> This listening task aims at collecting the **boundary timings** between parts of a song. During this task, you will be asked to listen for **when** a part of a song changes to another.

Rather than asking AMT annotators to listen to an entire song, we present them with short clips of music. For each clip, their task is to listen to the clip, determine if a boundary exists in the clip, and label the location of the boundary. We segment each song into clips of 20 seconds long with 10 seconds of overlap. The choice of 20 seconds is made according to the average length of segments defined by ground truth annotations in existing datasets as reported in [12]. The annotator is only given the option of labeling a single boundary and asked to choose the strongest boundary if they hear more than one. They also have the option of choosing no boundary.

The goal of the user interface design is to simplify the task and not bias the annotators to any clues apart from what is heard in the clip. To ensure that the annotator listened to the full clip and made an annotation decision before going to the next clip, all buttons and sliders are disabled except the Play/Pause button until the first full playback of the clip. The annotator uses the Play/Pause button and audio progress bar to listen to and navigate the clip. The annotation is done by clicking on the boundary selection slider. The darker green area surrounding the clicked location on the boundary selection slider indicates the playback region for the check selection button. The annotator has the option to simply not choose a boundary if they do not hear one. The next trial button is disabled until the annotator clicks on the change heard (submit current clicked location on the boundary selection slider) or no change heard button. A snapshot of the user interface is shown in Figure 1.

Each boundary collection task ("HIT" in AMT's terms) contains 10 clips, with 9 clips randomly selected from all non-hidden reference songs and 1 clip from the hidden reference song. The randomization of selected clips is designed such that annotators will not be presented with overlapping clips within one task (10 clips) and will cover the

[1] https://github.com/interactiveaudiolab/CAQE

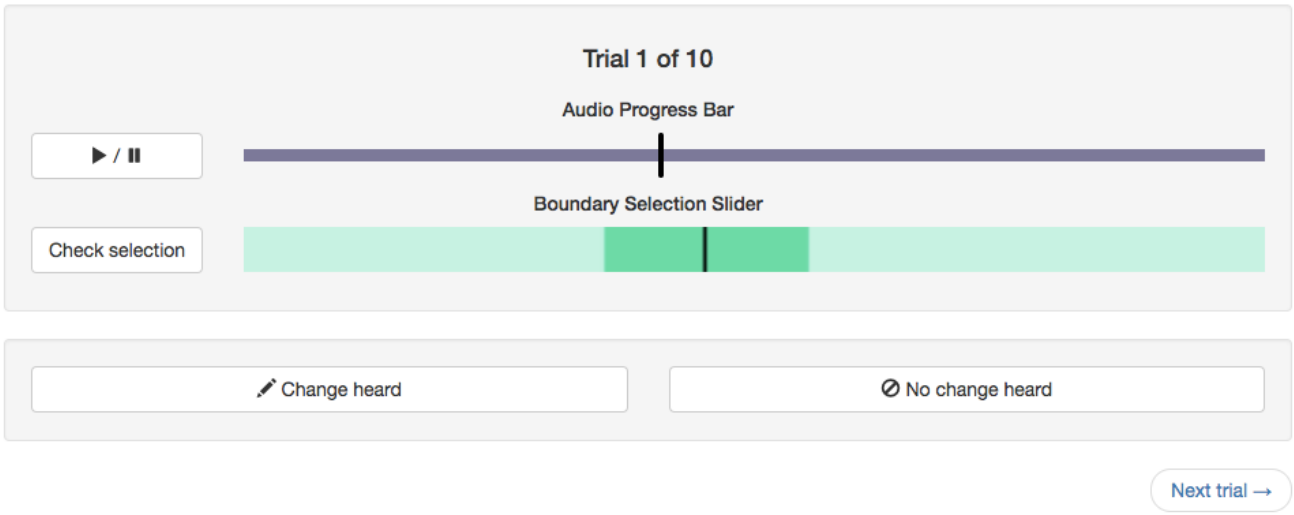


Figure 1. The user interface for the AMT annotating task. The annotator uses the Play/Pause button and audio progress bar to listen to and navigate the clip. The annotation is done by clicking on the boundary selection slider. The darker green area surrounding the clicked location on the boundary selection slider indicates the playback region for the check selection button.

full range of all songs once and once only if they finish all the tasks available to them. The order of the 10 clips within each task is also randomized. In order to further make sure that each song is covered with enough annotations, the annotation tasks are divided into two batches according to the two dataset subsets, meaning one batch for *Beatles-TUT* and one batch for *SALAMI*, and are collected separately. The batch for *Beatles-TUT* has 10 tasks (100 clips) and the batch for *SALAMI* has 8 tasks (80 clips) for each annotator.

It is mentioned in the previous section that out of the 5 songs, there is 1 song selected as a hidden reference acting as a quality check after collecting boundary annotations from the AMT annotators. To use this song as a quality check, a few clips from the song that have clear and obvious boundary regions are selected and manually annotated by the authors. The authors avoided using the ground truth annotations from the original datasets since the goal for the task is not to identify the "correct" boundaries defined by the original datasets, but rather simply annotating reasonable boundaries (that might differ from the ground truth annotations from the original datasets).

In order to take this concern into account, multiple boundary candidates for each hidden reference clip are allowed so the quality check accepts wider and reasonable results than just using ground truth annotations (only one boundary annotation for each clip) from the original datasets. After collecting boundary annotations from an AMT annotator, their boundary annotations on the hidden reference clips are compared against the author's annotations on the same clips. If the average distance between AMT annotator's boundary annotation to the closest annotation by the author for each clip is less than 3 seconds, all of the AMT annotator's annotations are accepted, otherwise they are rejected and not used for the analysis. The annotations by the authors for the hidden references could be found in the code repository [2] .

In order to determine if the annotator is following basic instructions and listening on a device (speakers/headphones) with a sufficient frequency response, we insert a hearing screening before the task begins, as done in [2]. After the completion of the task, the annotator is presented with a post task survey gathering demographical, musical background, and qualitative information. For the musical background, the AMT annotator is asked to answer if they consider themself to be a musician. We also ask for qualitative feedback on the task and what the annotators were listening for.

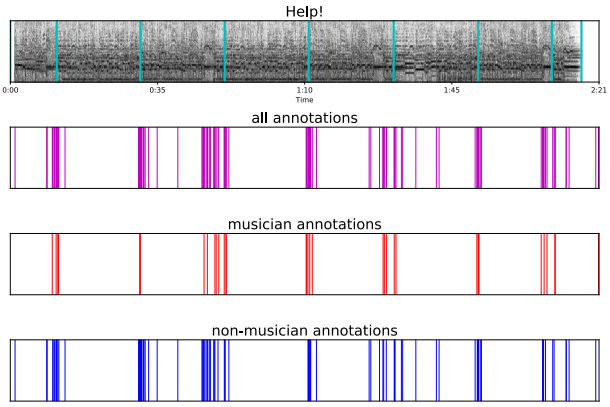
3. ANALYSIS

We consolidated the annotations from the individual clips so that we have all annotations of a given song on a common timeline. Since there is an overlap of 10 seconds between consecutive clips, there is a chance that the same boundary will be labeled twice by the same annotator (cases in which the time difference between the two annotations are less than 3 seconds). In such cases, we simply randomly discarded one of the annotations.

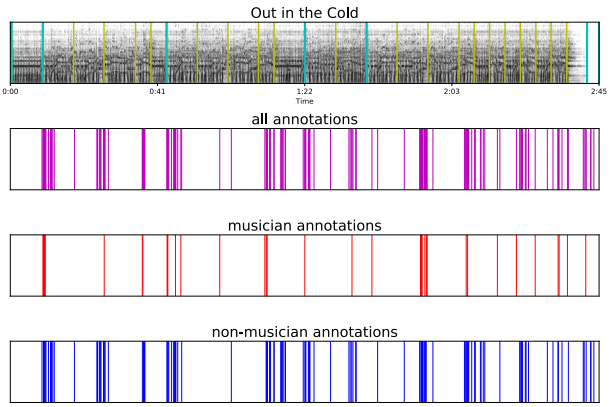
Two example songs showing the annotations from the AMT annotators along with a comparison to the ground truth annotations from the original datasets are shown in Figure 2. A correlation can be seen between the two.

In Table 2 and Table 3, overall statistics of the collected annotations and AMT annotators are shown. Though the task is only exercised on subsets of existing datasets, the number of AMT annotators and annotations easily outnumbered those of existing datasets. This observation is true even when considering only the statistics from AMT annotators that are self-identified musicians (numbers in

[2] https://github.com/wangsix/caqe_segmentation



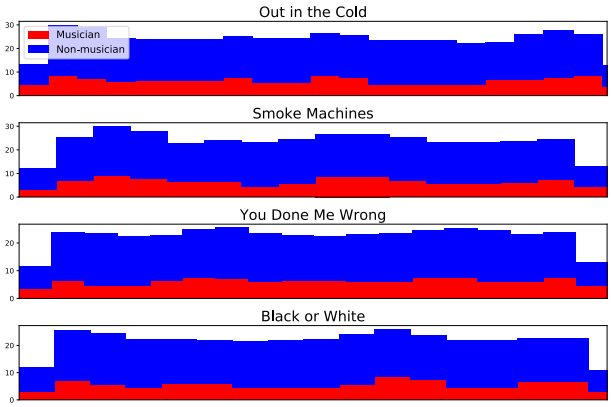
(a) Help! - Beatles



(b) Out in the Cold - Carole King

Figure 2. Two example songs with their annotations from selected subsets. The first row of each subplot is the CQT-spectrogram of the example song. The light blue lines in this row are the ground truth annotated by music experts. The yellow lines in "Out in the Cold" are the lower level ground truth by music experts. The vertical lines in the rest of the rows represent annotations from all AMT annotators, musician annotators and non-musician annotators respectively.

(a) Beatles subset



(b) Salami subset

Figure 3. Histograms of annotation count along the time line of each song. It shows a roughly evenly distributed coverage for each song being annotated.

parentheses in Table 2 and Table 3). It is also true if only complete annotations from single annotator are considered.

The seemingly low average coverage rate of each song by each annotator (right most column of Table 3) is a natural result of distributing clips randomly to AMT annotators throughout the AMT tasks. Even with the randomized task distribution, there are still at least 6 completed annotations for each song (2nd right column of Table 3).

The counts of annotations for each clip in each song are shown in Figure 3. From the histograms it could be observed that every song is fully covered by multiple annotators in a more or less evenly distributed manner. The beginning and ending of songs have less accumulated counts since they are not fully covered by overlaps.

3.1 Validation

To validate the collected annotations with the proposed methodology, the aggregated annotations for one song are treated as one segmentation boundary prediction from an arbitrary algorithm and compared to the ground truth boundary annotations from the original datasets.

The annotated timings of a song via crowdsourcing are first discretized into a binary vector with ones representing the presence of annotated boundaries. The discretization is done with a sampling rate of 22050Hz and 512 sample hop size. Then the binary vector of each song is normalized by its annotation count histogram (Figure 3) to account for different number of times each time region is annotated. After the normalization, a Gaussian window with 0.5 second standard deviation is convolved with the binary vectors obtaining a boundary detection function for each song. The boundary detection functions are renormalized to be between $[0, 1]$. A simple peak-picking strategy using the same principle as [12] with 0.5s window and 0.1 threshold is applied to select segmentation boundaries from the boundary detection function. These functions are shown in Figure 4 with the ground truth plotted against them.

The validation of such predictions are done using the standard MIREX 3 seconds structural music segmentation boundary accuracy evaluation metric. The validation results are shown in Table 4. From the F-measures and recall rate, it can be observed that the aggregated results from AMT based annotations in general agree with the music

	Annotators			Accepted Annotations	
Datasets	Accepted(musician)	Rejected	Invalid	Total	Per Annotator
Beatles-TUT	61(17)	11	89	1468	24.06
SALAMI	61(14)	41	65	1652	27.08

Table 2. Statistics of the AMT annotators and their annotations. Accepted annotators are the ones that passed the hidden reference quality check. Rejected annotators are the ones that failed the hidden reference quality check. Invalid annotators are the ones that did not pass the hearing screening or failed to submit their results.

Song	Count(musician)			Avg. Coverage
	Annotators	Annotations	Complete	Per Annotator(%)
All You Need is Love	61(17)	452(131)	6(3)	28.53
Help!	53(15)	275(80)	8(3)	30.69
Here, There and Everywhere	53(13)	264(75)	8(3)	29.93
Strawberry Fields Forever	59(16)	477(133)	7(3)	27.25
Smoke Machines	58(14)	376(103)	13(4)	25.31
You Done Me Wrong	61(14)	405(110)	12(4)	23.35
Black or White	61(14)	502(131)	14(5)	24.6
Out in the Cold	58(14)	369(92)	13(3)	23.7

Table 3. Song statistics from accepted annotations. The numbers in parentheses in columns 2 and 3 (from the left) are the numbers for self-identified musicians. The 4th column is the number of complete annotations by one annotator. The average coverage of a song per annotator (5th column) is calculated by dividing the average number of annotated clips per annotator for a given song by the total number of clips for that song.

experts annotating the original datasets. Also the self-identified musicians performed better than non-musicians in 7 cases out of 8. There might be two reasons for the higher recall rates. One reason is a potential bias due to the 20 seconds length clip. The other reason is that some of the peaks representing different levels of saliency or confidence (height of the boundary detection function), resulted in more peaks than the single level annotations by the music experts.

3.2 Inter-Annotator Analysis

In [9], the problem of subjectivity in music structural segmentation problem is studied by showing annotator effects with the two-way ANOVA factor analysis. The same analysis approach can not be applied here since the sets of annotators annotating each song are different (with overlapped annotators). In order to analyze the degree of agreement between AMT annotators, a simple measurement is devised. The agreement degree a_i of one annotation i of a clip to other annotations i' of the same clip is defined as

$$a_i = \frac{\sum_{i' \in I, i' \neq i} [1 \text{ if } i \text{ agrees with } i']}{\text{Number of items in } I}. \qquad (1)$$

where I is the set of annotations of a clip annotated by all annotators. The agreement of one annotation i to another i' is established if the annotated timing of i is within 3 seconds of i'' annotated timing, or if both i and i' has empty annotation. a_i is a value between $[0, 1]$ and could be thought of as the probability of one annotation agrees with other annotations in the same 20 seconds region. Since every annotator could only annotate a clip once, Equation 1 becomes a measurement of agreement between annotators. The agreement between annotators of one song could then be measured by calculating the average of all as over one song. The inter-annotator agreement of each song is shown in Table 5. From Table 5, one can observe that although the average agreement between annotators is above 70%, the standard deviations show that the agreement between annotators is not consistent throughout the song but varying a lot from time to time within a song. This observation supports the propositions made in [10] that multiple human annotations should be used during evaluation to take human subjectivity into account.

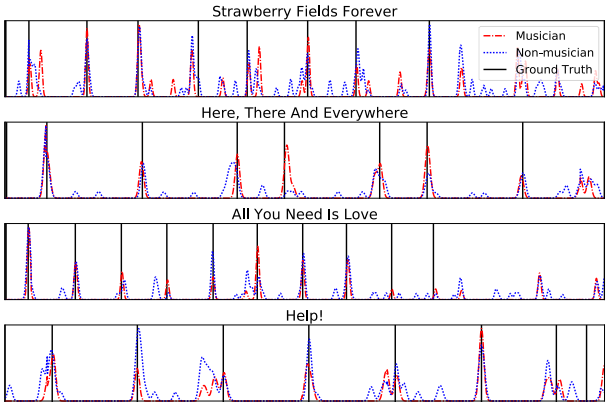
3.3 Boundaries Investigation

By qualitatively examining the AMT annotations and listening to the corresponding regions, it is evident that using a single time stamp representing the boundary between segmentations is inadequate. For example, the ground truth segmentation boundary around 0:50 in the song *Help!* by the Beatles is annotated at the beginning of the second verse when the lyrics start, but the AMT based annotations were spread across the region from 0:46 to 0:50 synchronizing with the sentence "Would you please, please, help me?". Musically, it could be correct to say that that region acts as the transition between two larger segments and the single boundary at 0:50 is inadequate to represent this musical property since the boundary is actually a prolonged region. This observation suggests that there exist different types of boundaries, with the easiest categorization being a clear-cut one versus a smooth/prolonged one. The AMT annotations of *Help!* are shown in Figure 2(a).

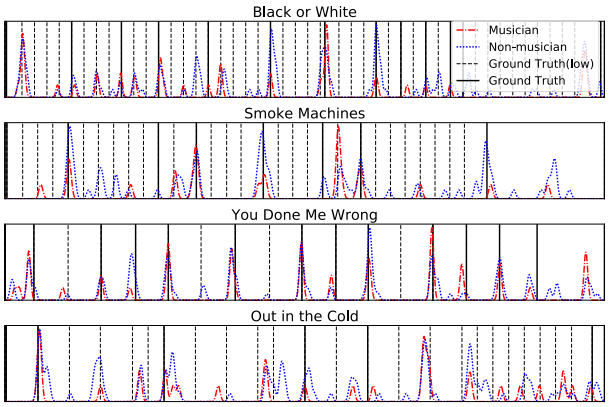
The other qualitative assessment is that the boundary detection function mentioned in section 3.1 shows that the hierarchical nature of structural segmentation boundaries exist and could be measured by the relative votes a boundary has compared to other boundaries in the same song. The boundary detection function also suggests that instead of having a discretized hierarchical representation of structural segmentation boundaries, a continuous version where

Song	Musician			Non-musician		
	Precision	Recall	F-measure	Precision	Recall	F-measure
All You Need is Love	0.73	0.85	0.79	0.5	0.69	0.58
Help!	0.5	0.78	0.61	0.33	0.89	0.48
Here, There and Everywhere	0.81	0.9	0.86	0.61	0.8	0.7
Strawberry Fields Forever	0.36	0.72	0.48	0.25	0.9	0.4
Smoke Machines	0.5	0.7	0.59	0.38	0.9	0.53
You Done Me Wrong	0.76	0.87	0.81	0.73	0.93	0.82
Black or White	0.67	0.8	0.72	0.33	0.73	0.46
Out in the Cold	0.39	0.88	0.54	0.2	0.75	0.32

Table 4. Standard 3-seconds precision, recall and F-measure evaluation metrics on the AMT annotator's annotation against ground truth from original datasets.



(a) Beatles subset



(b) Salami subset

Figure 4. The boundary detection functions obtained from AMT annotations against ground truth by music experts.

the saliency or confidence of boundaries is represented by a continuous curve might be another intuitive choice in terms of evaluating boundary detection algorithms.

4. FUTURE WORK AND CONCLUSION

In this paper, a methodology utilizing crowdsourcing for collecting alternative ground truth data for structural segmentation boundaries is proposed and validated. This methodology provides opportunities for researchers to create new segmentation boundary datasets in a fast and efficient way. To create a dataset with the proposed methodology, one has to make sure not to bias annotators toward

Song	Avg. agreement (Std.)
All You Need is Love	0.71 (0.29)
Help!	0.75 (0.26)
Here, There and Everywhere	0.65 (0.28)
Strawberry Fields Forever	0.72 (0.29)
Smoke Machines	0.72 (0.28)
You Done Me Wrong	0.72 (0.28)
Out in the Cold	0.76 (0.31)
Black or White	0.71 (0.31)

Table 5. The average and standard deviation of inter-annotator agreement of each song.

specific aural cues and manage the distribution of clips so that annotators work on evenly distributed clips between songs and songs get evenly distributed annotations from all annotators.

As suggested in section 3.3, different types of boundaries exist and could be investigated given the kind of data collected by the methodology proposed in this work. These boundary types could be categorized. Also the different types of musical cues that lead to the perception of music segmentation boundaries could be investigated by another round of crowdsourcing on the annotated data focusing on surveying the reasoning behind each annotations.

5. ACKNOWLEDGMENT

This work is supported by CREL (Center for Research in Entertainment and Learning at UCSD) and Adobe Research. Special thanks to Dr. Mark Cartwright (NYU) and Dr. Ian Wehrman (Adobe Research) for their kind help.

6. REFERENCES

[1] Michael J Bruderer, Martin F Mckinney, and Armin Kohlrausch. The perception of structural boundaries in melody lines of western popular music. *Musicae Scientiae*, 13(2):273–313, 2009.

[2] Mark Cartwright, Bryan Pardo, Gautham J Mysore, and Matt Hoffman. Fast and easy crowdsourced perceptual audio evaluation. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 619–623. IEEE, 2016.

[3] Nicholas Cook. Performance analysis and chopin's mazurkas. *Musicae scientiae*, 11(2):183–207, 2007.

[4] Jin Ha Lee. Crowdsourcing music similarity judgments using mechanical turk. In *ISMIR*, pages 183–188, 2010.

[5] Jin Ha Lee and Xiao Hu. Generating ground truth for music mood classification using mechanical turk. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 129–138. ACM, 2012.

[6] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE transactions on audio, speech, and language processing*, 16(2):318–326, 2008.

[7] Matthias Mauch, Chris Cannam, Matthew Davies, Simon Dixon, Christopher Harte, Sefki Kolozali, Dan Tidhar, and Mark Sandler. Omras2 metadata project 2009. In *Proc. of 10th International Conference on Music Information Retrieval*, page 1, 2009.

[8] Brian McFee, Oriol Nieto, and Juan Pablo Bello. Hierarchical evaluation of segment boundary detection. In *ISMIR*, pages 406–412, 2015.

[9] Oriol Nieto. *Discovering structure in music: Automatic approaches and perceptual evaluations*. PhD thesis, PhD thesis, New York University, 2015.

[10] Oriol Nieto. 4.7 approaching the ambiguity problem of computational structure segmentation. *Computational Music Structure Analysis*, page 177, 2016.

[11] Jouni Paulus and Anssi Klapuri. Music structure analysis by finding repeated parts. In *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 59–68. ACM, 2006.

[12] Jean Serra, Mathias Muller, Peter Grosche, and Josep Ll Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. *Multimedia, IEEE Transactions on*, 16(5):1229–1240, 2014.

[13] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. Design and creation of a large-scale database of structural annotations. In *ISMIR*, volume 11, pages 555–560, 2011.

[14] Douglas Turnbull, Gert RG Lanckriet, Elias Pampalk, and Masataka Goto. A supervised approach for detecting boundaries in music using difference features and boosting. In *ISMIR*, pages 51–54, 2007.