# Audio properties of perceived boundaries in music

Jordan B. L. Smith, *Student member, IEEE,* Ching-Hua Chuan, Elaine Chew

***Abstract*—Data mining tasks such as music indexing, information retrieval, and similarity search, require an understanding of how listeners process music internally. Many algorithms for automatically analyzing the structure of recorded music assume that a large change in one or another musical feature suggests a section boundary. However, this assumption has not been tested: while our understanding of how listeners segment melodies has advanced greatly in the past decades, little is known about how this process works with more complex, full-textured pieces of music, or how stable this process is across genres. Knowing how these factors affect how boundaries are perceived will help researchers to judge the viability of algorithmic approaches with different corpora of music.**

**We present a statistical analysis of a large corpus of recordings whose formal structure was annotated by expert listeners. We find that the acoustic properties of boundaries in these recordings corroborate findings of previous perceptual experiments. Nearly all boundaries correspond to peaks in novelty functions, which measure the rate of change of a musical feature at a particular time scale. Moreover, most of these boundaries match peaks in novelty for several features at several time scales. We observe that the boundary-novelty relationship can vary with listener, time scale, genre, and musical feature. Finally, we show that a boundary profile derived from a collection of novelty functions correlates with the estimated salience of boundaries indicated by listeners.**

***Index Terms*—boundaries, corpus analysis, music information retrieval, music analysis.**

## I. Introduction

ALTHOUGH a piece of music reaches a listener's ear as a continuous, uninterrupted audio signal, it is perceived as a set of discrete events. Moreover, these events are grouped, and these groups are themselves grouped recursively. Locating the boundaries between groups is essential to the task of music structure analysis, a task that has gained increased attention from the music information retrieval (MIR) community in recent years. Information about the structure of a piece of music has numerous applications: for instance, it can be used to generate summaries of songs [1], or to identify recordings of the same pieces in a corpus [2].

Many algorithms have been proposed for estimating musical structure directly from audio recordings (for a review, see [3]). These algorithms are mostly based on a few simple principles: that boundaries in a piece of music correspond to great changes in musical features (a premise introduced by [4]); that sections tend to be homogeneous with respect to features; and that sections often consist of repeated sequences. These principles are all reasonable, and the best analysis algorithms perform very well in evaluations. However, none of these algorithms are based on cognitive models of the perception of structure. As a result, they may not be adaptable to new data sets. Evidence for this is seen in the contortions required to account for specific musical situations, such as variations in tempo [5], or a final chorus that modulates upward [6]. Thus we may expect that an algorithm that performs well on a given dataset could perform poorly on a new dataset that includes new genres or styles. For this reason we look toward cognitive models of musical structure.

Lerdahl and Jackendoff's Generative Theory of Tonal Music (GTTM) [7], one of the most important models of the perception of structure, works from the ground up. The model begins its analysis at the atomic level, and provides rules to specify how music's indivisible units—read notes—are grouped in the listener's mind into sequences. The same general principles, with some adjustments, can often explain how these sequences can in turn be grouped together to form longer sequences.

Experimental evidence suggests that GTTM, and other comparable models such as [8], implemented as the Melisma system[1], and the Local Boundary Detection Model [9], make accurate predictions about when listeners perceive boundaries [10], [11]. Algorithmic implementations of all of these models exist [8], [9], [12], [13].

Unfortunately, these approaches have significant drawbacks. First, the above implementations all work only for very simple musical contexts: namely, monophonic melodies. This is

Manuscript received November 19, 2012. This research was supported by the Social Sciences and Humanities Research Council of Canada, by a USC Provost's PhD fellowship, and by a QMUL EPSRC Doctoral Training Account studentship. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect those of the funding agencies.

J. B. L. Smith is with the School of Electrical Engineering and Computer Science, Queen Mary University of London, E1 4NS, United Kingdom. Phone: +44 20 7882 6247; e-mail: jblsmith@eecs.qmul.ac.uk.

C.-H. Chuan is with the College of Computing, Engineering and Construction, University of North Florida, Jacksonville, FL 32224 USA. E-mail: c.chuan@unf.edu.

E. Chew is with the School of Electrical Engineering and Computer Science, Queen Mary University of London, E1 4NS, United Kingdom. E-mail: elaine.chew@eecs.qmul.ac.uk.

[1] <http://www.link.cs.cmu.edu/music-analysis/>, accessed 10 November 2012.

because the underlying models are harder to apply as the music grows in complexity from monophony to polyphony, or from monotimbral to multitimbral music. Second, while the models can all claim some degree of generality, the focus on melodic segmentation hints that they mainly apply to Western tonal music. Finally, because each model approaches segmentation in a ground-up way, they can be difficult to implement at longer time scales where it is more likely that the rules governing the analysis may conflict, or that the predictions will be muddied by factors that are hard to model, such as parallelism.

Hence there is a need for general models of the perception of structure in full-textured, polyphonic music. We propose beginning to develop such a model by taking advantage of resources that the MIR community has produced for the large-scale evaluation of algorithms: ground truth collections of structural annotations. A ground truth annotation is a description provided by a listener that is assumed to be the sole correct formal analysis. Of course, no such absolute truth exists. Evidence that different listeners perceive musical structure differently is found in any perceptual study that shows listeners marking boundaries at different times (e.g., [10], [14]). However, in these same examples, it was mainly observed that despite this variation, listeners tend to mark boundaries at similar times. Similarly, while [15] found that listeners' ability to detect repetitions changed across exposures, they changed in consistent ways. Thus, we may hope that ground truth provided by one listener represents how many listeners might hear a piece of music. Accordingly, while continuing to research best practices for collecting and labeling ground truth, the MIR community has produced many corpora of annotations of musical structure.

Collections of annotations are mainly used to evaluate the effectiveness of algorithms. But what if we treated these annotations not as ground truth, but as objects of study in themselves? Since each annotation reflects a listener's perception of a piece of music, we can analyze the annotation to test basic assumptions about how music is heard. One drawback of this approach is that existing collections of annotations, to save resources, tend to include just one listener's analysis per piece, in contrast to studies such as [14], which compared the responses of 21 listeners. This drawback is offset by the opportunity to study far more pieces: compared to the six songs studied in [14], the corpus studied in this article has two listeners per piece, but 746 pieces (over 100 times more than [14]'s 6).

Thinking of annotations as objects of study rather than tools for studying algorithms, we may actually derive some interesting conclusions about music cognition from the existing MIR literature. For example, [16] used machine learning to classify points in a recording as either boundaries or non-boundaries, and found that of over 800 feature dimensions considered, all three time scales and all four feature classes (harmony, melody, timbre and rhythm) were represented among the most informative 20. This suggests that listeners are likely to integrate information from many musical parameters at many time scales when judging the location of boundaries. Paulus and Klapuri [17] found that, when searching for similar sequences in music, it was optimal to calculate audio features over short time windows, but when searching for similar homogenous sections, a longer window was preferable. This may be evidence that when listeners judge two sections to be similar based on repeated sequences, the sequences they attend to are relatively short, whereas when listeners judge two sections based on their having an overall similar sound, this has been determined over a longer time scale.

An important question about the perception of boundaries is why listeners make the boundary indications they do. Both [10] and [14] collected free responses from participants about what cues they were attending to when they indicated a boundary. In both cases, listeners mostly indicated that a change in a particular parameter, such as timbre, rhythm, melody, register, articulation or harmony, motivated the response, while some indications were also attributed to parallelism or to a pause or break. Deliège [18], in testing the applicability of GTTM's grouping rules to perception, found that the salience of the rules varied with regards to implying boundaries. Sanden, Befus & Zhang [19] asked listeners to indicate boundaries while paying attention only to a single musical feature, such as timbre or harmony, and found that the resulting segmentations differed in how well they related to the overall perceived structure of the songs. Studies such as these can be limited either by their reliance on hand-picked, often very short stimuli that present exactly the musical contrasts being investigated, or by the infeasibility of collecting listener's impressions of large numbers of long stimuli.

Using large corpora of structural annotations that the MIR community has collected for the purpose of evaluation, we may efficiently investigate a large amount of music. By comparing the annotations to the recordings, we may investigate the relationship between features of the annotations—a record of how the structure was perceived by a listener—and features of the recordings, a record of what they heard. In this article, we report on our investigation of the acoustic features of those points designated by listeners as structural boundaries.

### *A. Proposed experiment*

We conducted a statistical analysis of how the acoustic properties of recordings relate to the boundary indications of listeners. We first test the hypothesis that boundaries correspond to moments in the recording at which relevant musical features change greatly. Secondly, we investigate how the answer to this question depends on the listener, the genre of the piece, and the musical features considered. Finally, we examine how the rate of agreement among musical features varies with the rate of agreement among listeners. Our approach is similar to [19], in which listeners were asked to segment eight pieces while paying attention to a single musical feature, and their responses were correlated to the perceived

structure of the pieces; in our case, we relate the actual acoustic properties of the signal to the perceived structure.

The present work differs from previous research in several important respects. First, our musical stimuli were complete, full-textured recordings, rather than short excerpts or simplified stimuli such as melodies or MIDI renditions. Second, our study does not focus on a narrow genre of music; since our investigation spans a wide range of genres, our observations may be more generalizable. Both of these differences lend our analysis an ecological validity that can be difficult to achieve in an experiment using few or artificial stimuli. Finally, our methodology is notable since, rather than collect data from an experiment, we are seeking insights into music perception by mining information from a large dataset developed for other applications. As will be seen, while we study a sizeable dataset, we have only exploited a fraction of the data available in this domain, and shown the beginnings of the discoveries possible.

## II. Materials and Methods: the SALAMI dataset

The data analyzed were originally created for the Structural Analysis of Large Amounts of Music Information (SALAMI) project.[2] The SALAMI project's goal is to use automatic structural analysis algorithms to analyze several hundred thousand musical recordings, which would allow musicologists interested in form to pursue research on a scale that was previously impossible. The project funded the creation of the largest ever corpus of human-generated structural annotations in order to demonstrate the effectiveness of these algorithms [20]. This corpus contains descriptions of nearly 1400 recordings, nearly 1000 of which were each analyzed by two independent listeners. Annotations for half of the total collection have been released to the public domain; the private half, which was not used in this study, will be released after serving for a few years as a benchmark dataset for evaluations such as the Music Information Retrieval Evaluation Exchange. The SALAMI data are described briefly in this section; a complete account of its design and its properties can be found in [20], and the "Annotator's Guide" used as a reference by the participants is available on the SALAMI website.[2]

### A. Participants and apparatus

The nine annotators (four men, five women) hired to provide annotations were all in their 20s and pursuing an advanced degree (Master's or PhD) in either Music Theory or Composition. They were trained to use Sonic Visualiser, a powerful software package that allows quick data entry and navigation of the recording, and they could use any means to listen to the music.

### B. Stimuli

The SALAMI collection contains roughly one quarter each of popular, jazz, classical, and world music. An additional portion was drawn from the Live Music Archive[3] (LMA), consisting mostly of popular and jazz recordings. Of the public half of SALAMI, 761 recordings were considered: 498 were annotated by two listeners and 263 by one listener. A breakdown of the number of annotations within each genre is given in Table 1.

All recordings were mp3s with 44.1 kHz sampling rates. The sound quality varied somewhat between files—while most mp3s had a bit rate between 128 and 192 kbps, some had variable bit rates and others had bit rates as low as 96 kbps—but none of these differences were expected to affect listeners' perceptions of structure, and this is not investigated here. Indeed, the poor sound quality of the recordings themselves was often a greater concern: the LMA includes some audience recordings of live concerts, which may include background noise or clipping.

SALAMI's annotations do not record the listeners' familiarity with the music. It is unlikely that any annotator had heard much of the corpus before given the extreme breadth of the corpus, but it is also unlikely that the occasional hits in the collection, such as Michael Jackson's "Thriller," were unknown to the annotators.

### C. Procedure

The annotators' descriptions were multi-dimensional in that three kinds of information were indicated separately: musical similarity (which was annotated at short and long time scales), formal function (e.g., "chorus" or "transition" labels), and lead instrumentation. Only the long time scale of the musical similarity layer was considered in the present research. In this layer, annotators indicated boundaries and provided uppercase letter labels ("A", "B", etc.) to indicate which sections were similar or shared the same fundamental musical idea. Annotators decided for themselves whether the unifying idea was primarily harmonic or melodic, or due to some other musical attribute. Labels could be inflected with a prime symbol to indicate substantial variation. Annotators were encouraged to indicate on average five distinct uppercase letters per song, and to align their analyses with the metrical grid of the piece, if applicable, so a section beginning with a pickup would be annotated as beginning on the down beat. An example pair of annotations is shown in Figure 1.

Annotators used Sonic Visualiser[4] according to the following workflow: first, listen through the full piece and indicate section boundaries in real time by pressing a key. On a second listening, pause to correct or adjust the position of boundaries as necessary. Next, provide labels for each section in each of the three layers—similarity, function, lead instrument. Finally, after skipping around to make corrections or resolve ambiguities as necessary, listen through the song a final time to confirm. The number of times each recording was fully heard is not known, but was requested to be at least three.

[2] <http://salami.music.mcgill.ca/>, <http://www.music.mcgill.ca/~jordan/salami/SALAMI-Annotator-Guide.pdf>, accessed 1 October 2012.

[3] <http://archive.org/details/etree>, accessed 1 October 2012.

[4] <http://www.sonicvisualiser.org/>, accessed 1 October 2012.

## III. Data Analysis

Structural analysis algorithms are commonly evaluated by executing the algorithm on a recording and grading the result against a ground truth annotation. This grade is difficult to interpret in isolation, so to fairly assess the significance of the result, a baseline approach, such as an algorithm that outputs random analyses or that makes predictions according to some naïve approach, should be executed on the same corpus.

In contrast to a typical evaluation, our goal is to study the annotations themselves, and not the effectiveness of an algorithm. Thus our analysis proceeds in an inverted manner: instead of comparing how well a given algorithm and a naïve baseline approach can predict the boundaries in an annotation, we compare how well the annotated boundaries and a random baseline set of points (non-boundaries) can predict the output of an analysis algorithm. In our case, this algorithm is based only on the rate of change of selected musical features. Our approach will effectively measure the amount of information that the annotations contain regarding these changes.

This section describes the audio features used to characterize the music, and the steps used to estimate the points of greatest musical change. None of these features is alleged to represent how the listener processes these musical attributes; the listener certainly perceives the music more holistically, basing their analysis on the properties not of frequency bands but of notes and other discrete events. The novelty-seeking approach tested here could be applied to more abstract representations using automatic beat tracking, transcription and source separation. However, these remain areas of active research: we lack robust tools with known error rates for these tasks that have been tested on a corpus as varied as the SALAMI data used here. Rather than employ intermediate and imperfect transcription efforts, we choose to estimate features directly from the audio. Assuming that changes in musical parameters are reflected by changes in our audio features, our study will test how these musical changes relate to the perception of boundaries.

### A. *Audio processing*

Five audio features were used to encapsulate information from the following musical parameters: timbre, harmony, key, rhythm, and tempo. The object was to select features that would differ when these musical parameters differed, and be stable when the parameters did not differ. None of the audio features chosen are totally independent of each other, but each has been designed to efficiently encapsulate information about a particular parameter while minimizing input from other information.

For timbre we chose Mel-frequency cepstral coefficients (MFCCs), widely regarded as a suitable representation of the timbre of a short audio snippet [21]. The values in an MFCC vector indicate the strength of different periodicities in the Mel-scaled spectrum and hence characterize the shape of the spectrum with a minimum of harmonic information. MFCCs were calculated using windows of 0.19 seconds and a hop size of half that. The lowest coefficient was discarded, since it relates specifically to overall loudness, and the next 12 coefficients were used.

For harmony we used the chromagram, which gives the strength in the signal of each pitch class from A to G#. The method used takes the constant-Q transform of the signal, which scales the spectrum so that each bin corresponds to a single pitch, and then sums the contributions of each pitch class. Our window size was 0.1 seconds with a hop size of half that. Both MFCCs and chromagrams were calculated using Queen Mary's Vamp Plugin set [22].

The center of effect (CE), which refers to a music segment's estimated tonal center within Chew's Spiral Array model of tonality [23], was used to provide information on the key. While the center of effect generator (CEG) algorithm finds the key itself, we use only the CE as a proxy for the key, which facilitates rate of change computations. The CE was calculated using the audio key-finding system from [24], which uses a fuzzy analysis scheme to extract the pitches sounded from the spectrum, maps the pitches to their letter names, then calculates the CE, i.e. the geometric mean of their representations in the Spiral Array. Window size was 0.37 seconds with a hop size of one quarter of that.

The remaining two features are derived from a sonogram, which applies a model of the ear to estimate the perceived loudness in each of the twenty Bark scale frequency bands. A fluctuation pattern (FP), also called a rhythmogram, measures the strength of loudness fluctuations between 0 and 10 Hz in each frequency band [25]. A 1200-element vector, giving the strength of 60 modulation frequencies in each of the 20 Bark scale frequency bands used, describes each window of the FP. The periodicity histogram gives the estimated strength of periodicities over the tempo range of 40 to 240 bpm (0.6 to 4 Hz) in a version of the signal that has been filtered to emphasize sharp attacks [26]. The strength of a period is the number of times its amplitude (estimated using a comb filter approach) exceeds a given threshold over a short series of windows. FPs and periodicity histograms were calculated with the MA Toolbox [27] using a window size of 3 seconds and a hop size of 0.37 seconds.

### B. *Generating novelty functions and picking peaks*

From each feature, we calculate a novelty function. Novelty functions were first proposed for segmenting audio by Foote [4], who estimated the amount of novelty at a point as the sum of the self-similarity of the passages that preceded and followed that point, and the dissimilarity between the two. Our novelty function ignores the internal similarity of the windows and focuses on the dissimilarity distance: we calculated at each point the Euclidean distance between the average feature vector before and after that point. It is essentially the same as the function used by the Argus algorithm for segmentation by tonal center [28], and can be seen as a continuous-time version of the difference features successfully employed by [16].

Varying the window size over which to take this average allows one to look at how the musical parameters evolve at different time scales; we used values starting at 0 (i.e., the first

derivative of the feature vectors) and up to 30 seconds at 5 second intervals, meaning 7 different time scales altogether. Given that listeners have indicated that they usually perceive boundaries in response to a changing musical feature, difference features are a natural physical measure to use. In both [28] and [16], using difference functions at multiple time scales has been shown to be effective means for predicting boundaries.

Peaks in the novelty function are hypothesized to indicate likely positions for boundaries. Of course, if the novelty function is sufficiently noisy, then there will be peaks throughout, and all boundaries and non-boundaries will be found to lie near peaks. We thus want to select only the tallest peaks. Our chosen peak-picking method first applies a smoothing filter to the novelty function that averages each value with the 10 previous and subsequent values; then we pick the top 10 peaks with the following heuristic: once a peak was added, any other peaks within 6.5 seconds were made ineligible. These choices broadly match the properties of our collection of annotations: the median number of segments per recording was 10, and the smallest average segment length for a recording was greater than 6.5 seconds.

### C. Random baseline

To properly assess the audio properties of the boundaries, it is necessary to compare them to a set of non-boundaries. We selected random non-boundaries with the following constraints: first, for each recording, there should be an equal number of non-boundaries and boundaries. This ensures that the mean segment lengths are identical. Second, the boundaries should lie a minimum distance from all true boundaries. We set a buffer of 1.5 seconds, ensuring that even in the annotation with the shortest mean segment length, non-boundaries could be drawn from at least half of the recording. (Note that the mean segment length across the entire corpus was over 25 seconds, so this problem was rare.) With these two constraints, non-boundaries were drawn with uniform probability over the eligible portions of the recording.

### D. Analysis metric

The chosen peaks in the many novelty functions now constitute our "ground truth," and we have two sets of points, one the annotated set of boundaries, the other a random set of non-boundaries. We can now calculate how well each set of points predicts the peaks, and compare them. Although two annotations were available for some recordings, we evaluated each separately.

The evaluation metrics we use are precision, recall, and *f*-measure. If we designate the set of annotated boundaries as A and the set of novelty function peaks as P, then the set of boundaries in A that 'hit' or are nearer to some peak in P than some given threshold (we use values of 3.0 and 0.5 seconds) is expressed as $A \cap P$. We can then express precision as the fraction of attempts that are successful ($|A \cap P| / |A|$) and recall as the fraction of peaks that are found ($|A \cap P|/|P|$). We are most interested in the *f*-measure, their harmonic mean.

Finally, we did not include in our evaluation any trivial boundaries, such as those that indicate the start or end of the recording, or any boundaries occurring in the first or final 1.5 seconds of the piece.

## IV. Results & Statistical Analysis

### A. Are boundaries points of novelty?

We first ask: are boundaries points of novelty? For each of the 761 recordings, we calculated 35 novelty functions, for each combination of five features and seven time scales, and extracted sets of peaks as described in section III.B. We calculated the average *f*-measure between these novelty functions and the boundaries and non-boundaries for each of the 1,253 annotations, resulting in 1,253 paired trials. The median *f*-measure for boundaries (0.328) was nearly twice that for non-boundaries (0.178) using a boundary match threshold of 3 seconds (see Figure 2). A paired Wilcoxon Signed Rank test[5] confirmed that the difference in medians was significant ($U = 771{,}373.5$, p-value $< 10^{-15}$), with a large effect size ($r = 0.59$). This indicates that boundaries are a better indicator of novelty peaks than non-boundaries. Indeed, the mean *f*-measure for boundaries over each set of 35 novelty functions was larger than that of non-boundaries in 93.9% of the annotations.

When the boundary match threshold is reduced to 0.5 seconds, the chance that a random point will be near a boundary also shrinks, and the contrast between the two groups grows: the median *f*-measure for boundaries (0.078) was more than twice that for non-boundaries (0.028). A Wilcoxon test again confirmed that the distributions have a different median ($U = 744{,}216.5$, $p < 10^{-15}$). The mean *f*-measure was greater for boundaries for 90.3% of the annotations. Despite the poorer overall performance, the effect size ($r = 0.58$) still indicates a large practical significance.

Since the boundaries surpassed the non-boundaries at predicting points of novelty, we can conclude that boundaries indeed tend to be more novel than other points in a piece. But what do the numbers mean qualitatively? The maximum *f*-measure possible is 1, indicating perfect recall and precision, but in practice, even two similar listeners are unlikely to replicate each other's analyses with such accuracy. Since we would not expect any algorithm to predict boundaries as well as another listener, we can use inter-annotator agreement as a performance ceiling. Using the subset of 492 pieces in our corpus that were annotated twice, and a threshold of 3.0 seconds, the median *f*-measure of inter-annotator agreement was 0.769. This is more than twice the median agreement

[5] Since the *f*-measures collected were not normally distributed, we use non-parametric tests. The Wilcoxon Signed Rank test is a non-parametric alternative to the paired Student's t-test, and gives the probability that two distributions of paired samples have the same median. The Wilcoxon Rank Sum test does the same for independent samples. The Kruskal-Wallis test is a non-parametric version of one-way ANOVA and tests whether all the medians of a set of independent distributions are the same. The Friedman test does the same, but accounts for a blocking factor.

between the novelty functions and the boundaries, which was 0.326 for this subset. This large difference was of course significant according to a Wilcoxon test ($U = 18.9$, $p < 10^{-15}$), and the effect size ($r = 0.60$) reflects that the factor by which points of novelty predict boundaries better than non-boundaries is almost the same as the factor by which boundaries are better predicted by another listener's annotated boundaries than by points of novelty.

We may also calculate the inter-annotator "disagreement," or the agreement between the boundaries of one annotation and the non-boundaries of the other, as a performance floor. The median of this measure was 0.118, which differed from the above medians with approximately the same significance and effect size. Using boundaries instead of non-boundaries to predict points of novelty led *f*-measure to increase from 0.178 to 0.326; a listener attempting to identify instead of to avoid the boundaries indicated by another listener led *f*-measure to increase from 0.118 to 0.769. The larger increase in the latter case suggests that although the boundaries relate more to novelty than do the non-boundaries, qualitatively, this is less significant than the perceptual difference between boundaries and non-boundaries.

If we were to compare our novelty functions to state-of-the-art structural analysis systems, we would likely find that they surpass our performance. At the 2012 MIREX evaluation[6], using a corpus of annotations comparable to ours, the mean *f*-measure achieved by nine algorithms varied between 0.42 and 0.49 using a 3.0-second threshold, and between 0.16 and 0.29 with a 0.5-second threshold. While all of these means far exceed the mean *f*-measures achieved in this study, this comparison is not meaningful: the algorithms submitted to MIREX use far more information than novelty (e.g., sequential repetitions, multimodal feature distributions), to estimate structure, and so it is expected that they would fare better. The purpose of this experiment is to investigate how well measures of novelty explain the information contained in the annotations; hence the relevant comparison is between the annotated boundaries and the random sets of non-boundaries.

However the results are parsed, we have observed that boundaries annotated by listeners are more likely than chance to be associated with a peak in novelty, suggesting that annotators do attend to novelty in the signal—and that the annotations, in turn, contain information about acoustic novelty. Does the size of this effect vary according to the listener, to the genre, or to the type of novelty function calculated? In the next four subsections, we address these questions by examining the effect of these factors on *f*-measure contrast, which we define as the amount by which the boundary *f*-measure exceeds the non-boundary *f*-measure for each novelty function, using a threshold of 3 seconds.

*Differences among listeners*

Among 1,253 annotations, a Kruskal-Wallis test indicated a significant effect of annotator ($\chi^2 = 15.577$, $df = 8$, $p = 0.049$) on the *f*-measure contrast, suggesting that the annotator's responses correlated with boundaries to varying degrees. However, a multiple comparison test (using a Bonferroni correction) found no pairs of annotators for which *f*-measure contrast differed significantly. The distributions shown in Figure 3a show that differences between the annotators are minimal, suggesting that altogether the annotators were similar in the way their annotations reflected musical changes.

*Differences among genres*

The effect of genre (see Figure 3b) was also significant according to a Kruskal-Wallis test ($\chi^2 = 63.631$, $df = 4$, $p < 10^{-12}$). A multiple comparison test found a difference in the *f*-measure contrast between five of the ten pairs of genres: four of these indicated that *f*-measure contrast was smaller in classical than in other genres, with a small to moderate effect size ($0.19 \leq r \leq 0.33$); the fifth indicated a small difference between popular and jazz ($r = 0.17$). This could indicate that when annotating classical music, listeners paid more attention to criteria other than novelty, such as parallelism; or, that the transitions between sections in a classical piece tend to be less sudden—that is, there are more elided boundaries than in other kinds of music.

*Differences among time scales*

To evaluate the effect of window size, we averaged the *f*-measure contrast across features for each of the seven window sizes and for each annotation. A Friedman test found a significant effect of window size ($\chi^2 = 844.94$, $df = 6$, $p < 10^{-15}$), and many pairs of time scales differed. All comparisons between the 0-second window size and another showed a small to moderate effect size ($0.20 \leq r \leq 0.32$). As seen in Figure 3c, the immediate derivative (time scale 0) did not improve very much on the baseline at all, suggesting that novelty at this time scale was of little relevance to the annotators. Additional comparisons yielded a small difference between the 30-second window size and window sizes between 5 and 20 seconds ($0.11 \leq r \leq 0.19$), and between the 25-second window size and window sizes between 5 and 15 seconds ($0.10 \leq r \leq 0.16$). This suggests that these longer time scales are also less relevant in terms of acoustic novelty. The 10-second time scale improved the *f*-measure the most, suggesting that it was the most perceptually relevant time scale for establishing section boundaries. It is interesting that although the mean segment length across all pieces was roughly 25 seconds, the 25-second window offered less contrast to the baseline than the 10-second window. This could simply be explained by the fact that the boundaries of short sections risk being obscured by a large window, but a section larger than a shorter window size is less likely to be obscured.

*Differences among features*

A Friedman test found differences in *f*-measure contrast among features averaged across time scales to be significant ($\chi^2 = 529.71$, $df = 4$, $p < 10^{-15}$). A multiple comparison test

[6] <http://nema.lis.illinois.edu/nema_out/mirex2012/results/struct/sal/summary.html>. Accessed 13 November 2012.

followed by calculation of effect size yielded small differences between timbre and key ($r = 0.27$), timbre and tempo ($r = 0.21$), as well as rhythm and key ($r = 0.25$) and rhythm and tempo ($r = 0.21$), suggesting that timbre and rhythm were both more reliable indicators of boundaries than tempo or key (see Figure 3d). The effectiveness of harmony lay somewhere in between: it was found to differ from timbre, tempo and rhythm with a small effect size ($0.10 \leq r \leq 0.12$) and to differ from key with a slightly larger effect size ($r = 0.19$).

That tempo should be a less reliable predictor of boundaries is a reasonable result, since in most popular and jazz music, which comprise at least half the data studied, tempo does not commonly vary across sections. However, it is a surprise for key. The features for key (center of effect) and harmony (chroma) provide similar information, but while chroma merely provide the raw pitch content, center of effect condenses this information into a single estimate of the tonal center. Our results suggest that for the purpose of locating boundaries, this process filters out more signal than noise.

### *B. Do any boundaries not match a novelty peak at all?*

The mean *f*-measure indicates how well the annotated boundaries predict the set of peaks given by a particular novelty function. But we would not expect every boundary to be suggested by every musical feature at every time scale. A further question to ask is if there are any boundaries that do not match any peak at all; this would indicate the minimum extent to which boundaries are not associated with changes in musical parameters.

To answer this question, we produce a histogram showing the number of novelty function peaks associated with each boundary, using a threshold of 3 seconds (Figure 4). The comparable histogram for non-boundaries is given below the *x*-axis. It shows that 7.1% of annotated boundaries do not match a peak in any novelty function, meaning 92.9% match at least one—and most match many more. The median number of novelty functions matched is eleven; since there are five features and seven time scales, the median indicates that half of the boundaries matched at least two distinct features at three distinct time scales, showing boundary perception to be a function of multiple features at multiple time scales. The non-boundary histogram is more heavily skewed to low values than the boundary histogram, and they are about equal when the number of novelty peaks matched is nine. Hence, if exactly nine novelty peaks match a particular point, then that point is about equally likely to be perceived as a boundary as not; the odds of the point being a boundary steadily increase as more novelty peaks match that point.

The light gray regions in Figure 4 indicate the subset of boundaries that are "symmetric," i.e., those where the labels of the sections before and after the boundary are the same (prime symbols attached to segment labels were disregarded here, so the labels 'A' and 'A′' were treated as equal). Symmetric boundaries are hypothesized to indicate less novelty than non-symmetric boundaries, and this is borne out modestly by the data. Of the boundaries that match no novelty function, 34.3% are symmetric, whereas only 26.7% of all boundaries were symmetric. The median number of matching novelty functions for non-symmetric boundaries is eleven; for symmetric boundaries, it is nine. A Wilcox rank-sum test showed that this was a significant effect ($U = 11{,}406{,}306$, $p < 10^{-15}$), with a small effect size ($r = 0.10$). The effect here is slight, but the measure of "symmetry" used is very rough, and does not take into account the annotated changes in lead instrumentation. In many of the jazz pieces, for instance, nearly every section is given the same label, and the most salient structural information lies with the changing soloists. Still, this result provides some support for the hypothesis that the perception of symmetric boundaries owes less to novelty and perhaps owes more to factors such as parallelism.

### *C. Can boundary salience be estimated by annotation concurrence?*

We have observed that boundaries vary in the number of novelty functions they match: nearly all boundaries match a few novel points, and a minority match several. This is curiously analogous to the finding in [14] that, in each piece studied, a few boundaries stood out as salient to all listeners, while the majority of boundaries were indicated by only a handful of listeners. They further found that the perceptual salience of a boundary correlated strongly with the number of people who indicated that boundary. Bruderer et al. [14] assembled the boundary indications of many listeners to produce a continuous boundary profile, indicating at each moment the potential salience of a boundary in that position. We conjecture that we could obtain a similar result by collecting information from a set of automated listeners (i.e., novelty functions), each indicating boundaries according to the parameter (i.e., a given musical feature at a given time scale) to which they are attending.

We do not have the boundary salience data to test this claim, but we may approximate salience by combining the annotations of two listeners and giving more weight to non-symmetric boundaries. We combined annotations with the technique proposed by [14]: all the boundary indications were collected (non-symmetric ones counted twice), and the result was convolved with a Gaussian function (we used a full-width half-maximum of 1.5 seconds instead of 1.25 seconds given by [14]).

Figure 5a shows the result of applying this procedure to the two annotations for the song "I Close My Eyes" by the band Shivaree. The dashed line gives the boundary function as estimated from the two annotations; the solid line gives the boundary function estimated from the 35 novelty functions. There is very close agreement with the largest peaks in the novelty functions, and less agreement among the less significant peaks. The Pearson correlation between the two time series is 0.60, a close overall fit. When we performed this procedure on all 492 pieces for which two annotations were available, we found the mean Pearson correlation to be 0.33 (sd = 0.18), suggesting a moderate relationship throughout the corpus. An example of a pair of boundaries that matched the

novelty functions poorly is given in Figure 5b. These are the annotation- and novelty-derived boundary functions for Precious Bryant's "Morning Train," and the Pearson correlation between them is -0.03. Even so, the fit is qualitatively good for the second half of the song.

This result shows that the simple measure of novelty defined in this article, versions of which are already used regularly in the MIR community, actually does seem to converge on the same information contained in the annotations. Moreover, this information, when collected from a variety of features at different time scales, can be combined into an overall novelty function that seems to reflect the same patterns of salience that listeners display.

## V. Conclusion

We have investigated a large corpus of recordings and annotations to show that acoustic novelty, as estimated by features reflecting timbre, harmony, key, rhythm and tempo, relates strongly to the position of boundaries indicated by listeners. The strength of this relationship was shown to be moderately affected by the feature and the time scale used to estimate the novelty—specifically, the novelty of tempo and key and the novelty at the shortest time scales were found to be less informative than the rest—as well as by genre, with the result that boundaries in classical music were less consistently novel according to our features than in other genres. Finally, we saw that a boundary profile derived from novelty functions correlated modestly with a boundary profile estimated from the annotations. Since [14] found that the fraction of listeners who indicated a boundary correlated with the judged salience of that boundary, our findings may help extend this result to suggest that the salience of a boundary is correlated to its acoustic novelty.

At the same time, our results show the limitations of solely analyzing points of novelty for the purpose of boundary estimation: although nearly all boundaries corresponded to a peak in novelty, not all peaks in novelty indicated a boundary (see Figure 4). This indicates that as a predictor of boundaries, acoustic novelty has high recall but low precision. Thus, while novelty is important to listeners, it is not the final word; listeners reject many novel points as false positives, perhaps using information relating to metrical structure, parallelism, or other factors to perceive seemingly novel moments as moments of continuation. The success of state-of-the-art structural analysis algorithms suggests this is indeed the case (Section IV.A).

Bruderer and McKinney [11] demonstrated the perceptual validity of segmentation models that used score-based representations. The present study may help develop comparable audio-based models that could be applied to any recorded music, whether or not a score exists—or whether the music even could be transcribed using Western music notation, as much electronic music cannot.

The annotations deserve further study, as there were many interesting interactions between the features and time scales used, and the genre of the piece, that could not be fully explored here. For example, the usefulness of the tempo feature was higher at longer time scales; the timbre feature was less useful on the LMA database, perhaps because many of these recordings were noisier; and the best time scale on the classical music was 25 seconds (even though this was among the worst time scales for the other genres), perhaps indicating that boundaries in classical music tend to reflect long-term changes, or that the most significant short-term changes are often misleading with respect to finding boundaries in classical music.

An important caveat to our findings is that there is no proof of causality: boundaries do tend to occur at novel moments, but this novelty is not necessarily what motivates the listener to perceive a boundary. An alternative explanation would be that listeners identify repeated sequences and infer boundaries between them, and that the novelty of the boundaries arises from the fact that these sequences tend to differ acoustically. This alternative is perhaps supported by the observation that symmetric boundaries (those between repetitions) are less well explained by novelty than the other boundaries. While this experiment cannot settle the question of causation, the studies conducted in [10] and [14] confirm that listeners often find the changes that occur at boundaries their most salient aspect. Still, as illustrated in Figure 4, many boundaries remain unexplained by any kind of acoustic novelty. Further studies should test how well these boundaries are explained by parallelism, pauses, and changes in other musical parameters not tested here.

## Acknowledgment

We thank Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and Stephen J. Downie, who with the first author assembled the SALAMI data set. The SALAMI project was supported by the Social Sciences and Humanities Research Council of Canada, the National Science Foundation of the United States and JISC of the United Kingdom.

## References

[1] G. Peeters, "Deriving musical structures from signal analysis for music audio summary generation: 'Sequence' and 'State' approach," *Computer Music Modeling and Retrieval*, Springer Berlin / Heidelberg, 2004, pp. 169–185.

[2] P. Grosche, J. Serrà, M. Müller, and J. L. Arcos, "Structure-based audio fingerprinting for music retrieval," in *Proc. ISMIR*, Porto, Portugal, 2012, pp. 55–60.

[3] J. Paulus, M. Müller, and A. Klapuri, "Audio-based music structure analysis," in *Proc. ISMIR*, Utrecht, The Netherlands, 2010, pp. 625–636.

[4] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. IEEE Int. Conf. on Multimedia & Expo*, 2000, pp. 452–455.

[5] M. Müller and F. Kurth, "Towards structural analysis of audio recordings in the presence of musical variations," *EURASIP J. Appl. Signal Process*, 2007.

[6] M. Goto, "A chorus section detection method for musical audio signals and its application to a music listening station," in *IEEE Audio, Speech, Language Process.*, Vol. 14 (5), 2006, pp. 1783–1794.

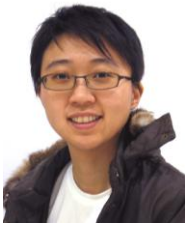
[7] F. Lerdahl, and R. S. Jackendoff, *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press, 1983.

[8] D. Temperley, *The Cognition of Basic Musical Structure*. Cambridge, MA: MIT Press, 2001.
[9] E. Cambouropoulos, "The local boundary detection model (LBDM) and its application in the study of expressive timing," in *Proc. of the ICMC*, Havana, Cuba, 2001.
[10] E. F. Clarke, and C. L. Krumhansl, "Perceiving musical time," *Music Perception*, 7 (3), 1990, pp. 213–251.
[11] M. Bruderer, and M. McKinney, "Perceptual evaluation of models for music segmentation," in *Proc. Conf. Interdisciplinary Musicology*, Thessaloniki, Greece, 2008.
[12] B. Frankland, and A. Cohen, "Parsing of melody: Quantification and testing of the Local Grouping Rules of Lerdahl and Jackendoff's A Generative Theory of Tonal Music," *Music Perception*, 21 (4), 2004, pp. 499–543.
[13] M. Hamanaka, K. Hirata and S. Tojo, "Implementing a Generative Theory of Tonal Music," *Journal of New Music Research*, 35 (4), 2006 pp. 249–277.
[14] M. Bruderer, M. McKinney, and A. Kohlrausch. "The perception of structural boundaries in melody lines of Western popular music," *Musicæ Scientæ*, 13 (2), 2009, pp.273–313.
[15] E. Margulis, "Musical repetition detection across multiple exposures," *Music Perception*, 29 (4), 2012, pp. 377–385.
[16] D. Turnbull, G. Lanckriet, E. Pampalk, and M. Goto, "A supervised approach for detecting boundaries in music using difference features and boosting," in *Proc. ISMIR*, Vienna, Austria, 2007, pp. 51–54.
[17] J. Paulus, and A. Klapuri, "Acoustic features for music piece structure analysis," in *Proc. DaFX*, Espoo, Finland, 2008, pp. 309–312.
[18] I. Deliège, "Grouping conditions in listening to music: An approach to Lerdahl and Jackendoff's Grouping Preference Rules," *Music Perception*, 4 (4), 1987, pp. 325–359.
[19] C. Sanden, C. R. Befus, and J. Z. Zhang, "A perceptual study on music segmentation and genre classification," *Journal of New Music Research*, 41 (3), 2012, 277–93.
[20] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and S. J. Downie, "Design and creation of a large-scale database of structural annotations," in *Proc. ISMIR*, Miami, FL, 2011, pp. 555–560.
[21] J.-J. Aucouturier, F. Pachet, and M. Sandler, "'The Way It Sounds': Timbre models for analysis and retrieval of music signals," *IEEE Trans. Multimedia*, 7 (6), pp. 1028–1035.
[22] C. Landone, M. Gasser, C. Cannam, C. Harte, M. Davies, K. Noland, T. Wilmering, W. Xue, and R. Zhou, QM Vamp Plugins, 2011. Available: <http://isophonics.net/QMVampPlugins>, accessed 1 October 2012.
[23] E. Chew, "Towards a Mathematical Model of Tonality," Ph.D. dissertation, Operations Research Center, MIT, Cambridge, MA, 2000.
[24] C.-H. Chuan and E. Chew, "Audio key finding: Considerations in system design and case studies on Chopin's 24 Preludes," *EURASIP Journal on Advances in Signal Processing*, 2007.
[25] E. Pampalk, A. Rauber, and D. Merkl, "Content-based organization and visualization of music archives," in *Proc. ACM Multimedia*, Juan les Pins, France, 2002, pp. 570–579.
[26] E. Pampalk, S. Dixon, and G. Widmer, "Exploring music collections by browsing different views," *Computer Music Journal*, 28 (2), 2004, 49–62.
[27] E. Pampalk, "A Matlab toolbox to compute similarity from audio," in *Proc. ISMIR*, Barcelona, Spain, 2004, pp. 254–257.
[28] E. Chew, "Regards on two regards by Messiaen: Post-tonal music segmentation using pitch context distances in the Spiral Array," *Journal of New Music Research*, 34 (4), 2005, 341–354.

**Jordan B. L. Smith** is currently pursuing a Ph.D. in the Centre for Digital Music at Queen Mary, University of London. He received a M.Sc. in operations research engineering in 2012 at University of Southern California (Los Angeles, CA, USA), a M.A. in music technology in 2010 at McGill University (Montreal, QC, Canada), and in 2006 an A.B. in music and physics at Harvard College (Cambridge, MA, USA).

As a research assistant at McGill, he planned and implemented the collection of ground truth for the Structural Analysis of Large Amounts of Music Information project [20]. His current research focuses on differences among listeners in the perception of musical structure.

Smith was awarded a doctoral research award from the Social Sciences and Humanities Research Council doctoral grants in 2012. He was awarded a Provost's Ph.D. fellowship from the University of Southern California in 2010.



**Ching-Hua Chuan** is an Assistant Professor at the School of Computing in the University of North Florida College of Computing, Engineering and Construction. She received her Ph.D. in computer science from the University of Southern California (Los Angeles, CA, USA) in 2008. She received B.S. and M.S. degrees in electrical engineering from National Taiwan University in 1999 and 2001, respectively. She has published refereed articles in journals and at conferences on audio content analysis and style-specific music generation.

She served as the Publicity Chair of the 12th International Conference on Music Information Retrieval (ISMIR), and Program Committee in the ACM Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies. She was the recipient of the best new investigator paper award at the Grace Hopper Celebration of Women in Computing in 2010. She is also the founder of Women in Music Information Retrieval (WiMIR).

**Elaine Chew** (M'05) received the B.A.S. degree in mathematical and computational sciences with honors, and in music with distinction, from Stanford University (Stanford, CA, USA) in 1992, and the S.M. and Ph.D. degrees in operations research from the Massachusetts Institute of Technology (Cambridge, MA, USA) in 1998 and 2000, respectively.

She joined Queen Mary, University of London, in the United Kingdom, as Professor of Digital Media in fall 2011, where she serves as Director of Music Initiatives in the Centre for Digital Music. Previously based in the US, she was a faculty member at the University of Southern California, visiting at Harvard University and Lehigh University, and recipient of a National Science Foundation Faculty Early Career Award, Presidential Early Career Award in Science and Engineering, and Edward, Frances, and Shirley B. Daniels Fellowship at the Radcliffe Institute for Advanced Study. Her research interests center on mathematical and computational modeling of music prosody and structure, on which she has authored over 80 refereed journal and conference articles. The goal of this research is to explain what it is that musicians do when they interpret and perform music, how they do it, and why. Applications of this work include automated music analysis and visualization, expressive performance analysis and synthesis, and ensemble interaction.

Prof. Chew serves on the editorial boards of the Journal of Mathematics and Music, Journal of New Music Research, ACM Computers in Entertainment, and Journal of Music and Meaning, and on the editors' panel of Computing in Musicology, and has served as program co-chair and edited proceedings of the International Conference on Mathematics and Computation in Music, and the annual conference of the International Society for Music Information Retrieval. She is also a member of the Institute for Operations Research and Management Science, and the Association for Computing Machinery.

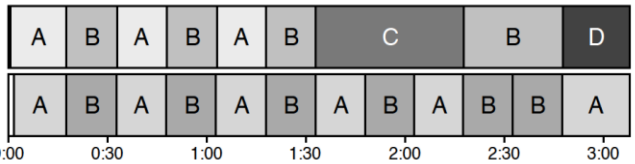



Figure 1. Two annotations for the song "Ain't Too Proud To Beg" by The Lost (SALAMI ID 1420). The shading of the segments emphasizes the labels within each annotation separately.

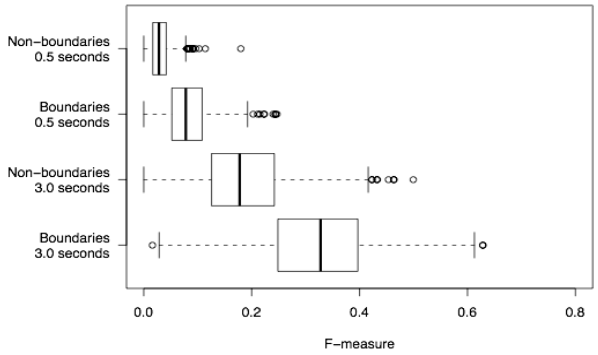



Figure 2. Distribution of *f*-measure scores for boundaries and for random sets of non-boundaries, given a grading threshold of 3.0 or 0.5 seconds. Outliers in a modified boxplot are those that lie more than 1.5 times the interquartile range beyond the third quartile.

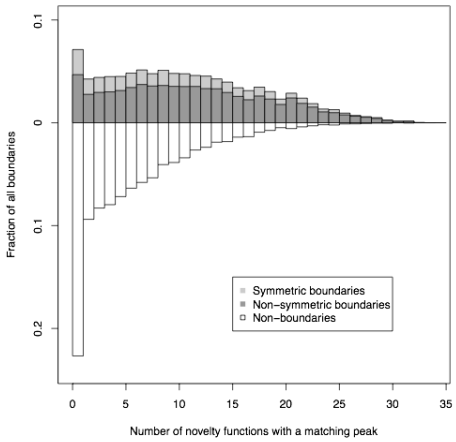



Figure 4. Comparison of histograms for boundaries (gray) and non-boundaries (white) according to number of novelty functions with a matching peak. Symmetric boundaries, i.e., those between sections with the same letter label, are distinguished from non-symmetric boundaries.

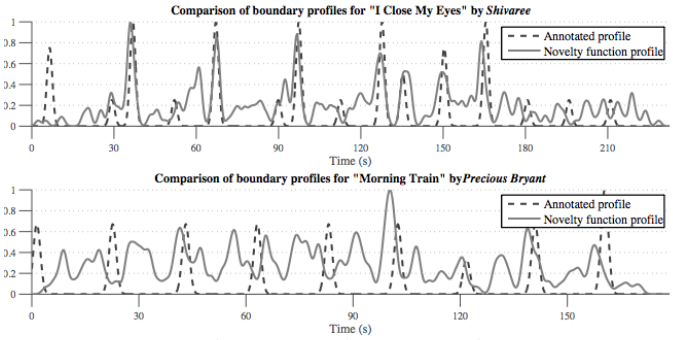
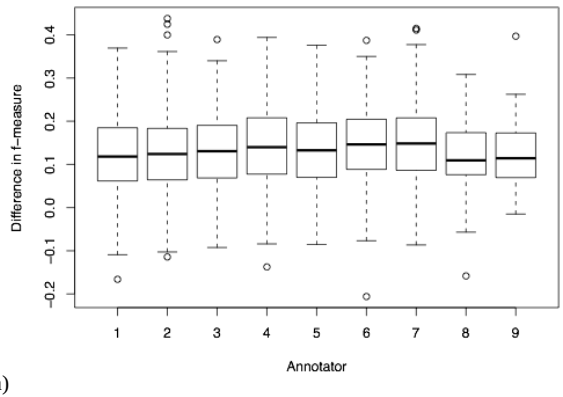


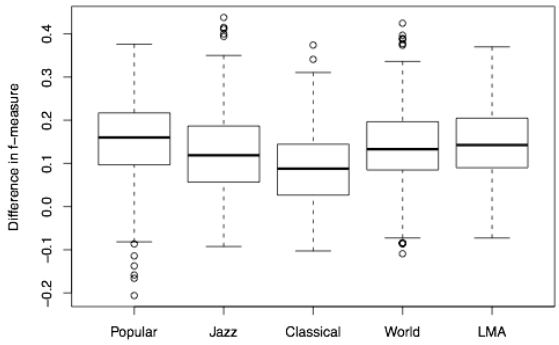

Figure 5. Comparison of boundary profiles estimated from annotations (solid line) and from novelty functions (dashed line) for (a) "I Close My Eyes" by Shivaree (SALAMI ID 4), and (b) "Morning Train" by Precious Bryant (SALAMI ID 36).

| Genre | One annotator | Two annotators |
|---|---|---|
| Popular | 51 | 101 |
| Jazz | 10 | 112 |
| Classical | 44 | 65 |
| World | 30 | 78 |
| Live Music Archive | 113 | 142 |

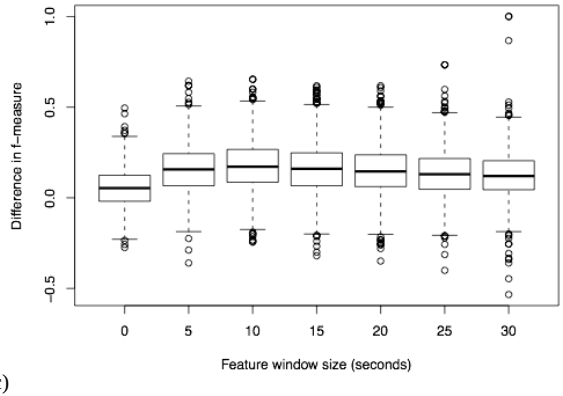
Table 1. Number of recordings analyzed according to genre and number of annotators.



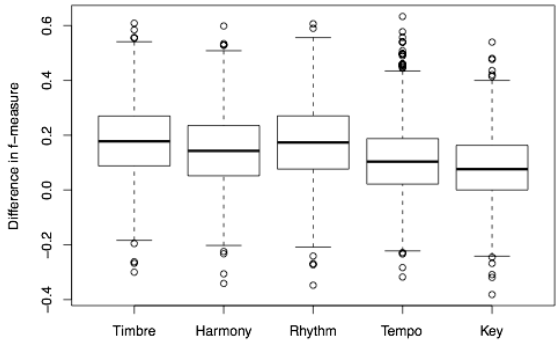

(a)




(b)




(c)




(d)

Figure 3. Distribution of *f*-measure contrast (the absolute improvement in *f*-measure achieved by sets of boundaries over non-boundaries) among (a) different annotators, (b) different genres, (c) different time scales, and (d) features. All results found using a grading threshold of 3.0 seconds.