PERCEPTUAL ANALYSIS OF THE F-MEASURE FOR EVALUATING SECTION BOUNDARIES IN MUSIC

Oriol Nieto[1], **Morwaread M. Farbood**[1], **Tristan Jehan**[2], and **Juan Pablo Bello**[1]

[1]Music and Audio Research Lab, New York University, {oriol, mfarbood, jpbello}@nyu.edu
[2]The Echo Nest, tristan@echonest.com

ABSTRACT

In this paper, we aim to raise awareness of the limitations of the F-measure when evaluating the quality of the boundaries found in the automatic segmentation of music. We present and discuss the results of various experiments where subjects listened to different musical excerpts containing boundary indications and had to rate the quality of the boundaries. These boundaries were carefully generated from state-of-the-art segmentation algorithms as well as human-annotated data. The results show that humans tend to give more relevance to the *precision* component of the F-measure rather than the *recall* component, therefore making the classical F-measure not as perceptually informative as currently assumed. Based on the results of the experiments, we discuss the potential of an alternative evaluation based on the F-measure that emphasizes precision over recall, making the section boundary evaluation more expressive and reliable.

1. INTRODUCTION

Over the past decade, significant effort has been made toward developing methods that automatically extract large-scale structures in music. In this paper, we use the term *musical structure analysis* to refer to the task that identifies the different sections (or segments) of a piece. In Western popular music, these sections are commonly labeled as *verse*, *chorus*, *bridge*, etc. Given that we now have access to vast music collections, this type of automated analysis can be highly beneficial for organizing and exploring these collections.

Musical structure analysis is usually divided into two subtasks: the identification of section boundaries and the labeling of these sections based on their similarity. Here, we will only focus on the former. Section boundaries usually occur when salient changes in various musical qualities (such as harmony, timbre, rhythm, or tempo) take place. See [9] for a review of some of the state of the art in musical structure analysis.

Typically, researchers make use of various human-annotated datasets to measure the accuracy of their analysis algorithms. The standard methodology for evaluating the accuracy of estimated section boundaries is to compare those estimations with ground truth data by means of the F-measure (also referred to as the hit rate), which gives equal weight to the values of precision (proportion of the boundaries found that are correct) and recall (proportion of correct boundaries that are located). However, it is not entirely clear that humans perceive the type of errors those two metrics favor or the penalties they impose as equally important, calling into question the perceptual relevance of the F-measure for evaluating long-term segmentation. To the best of our knowledge, no empirical evidence or formal study exists that can address such a question in the context of section boundary identification. This work is an effort to redress that.

Our work is motivated by a preliminary study we ran on two subjects showing a preference for high precision results, thus making us reconsider the relevance of precision and recall for the evaluation of section boundary estimations. As a result, in this paper we present two additional experiments aimed at validating and expanding those preliminary findings including a larger subject population and more controlled conditions. In our experiments, we focus on the analysis of Western popular songs since this is the type of data most segmentation algorithms in the MIR literature operate on, and since previous studies have shown that most listeners can confidently identify structure in this type of music [1].

The rest of this paper is organized as follows. We present a review of the F-measure and a discussion of the preliminary study in section 2. We describe the design of two experiments along with discussions of their results in sections 3 and 4. We explore an alternative F-measure based on our experimental findings that could yield more expressive and perceptually relevant outcomes in section 5. Finally, we draw conclusions and discuss future work in section 6.

2. THE F-MEASURE FOR MUSIC BOUNDARIES

2.1 Review of the F-measure

In order to evaluate automatically computed music boundaries, we have to define how we accept or reject an estimated boundary given a set of annotated ones (i.e., find the intersection between these two sets). Traditionally, re-

searchers consider an estimated boundary *correct* as long as its maximum deviation to its closest annotated boundary is $\pm$ 3 seconds [8] (in MIReX, [1] inspired by [16], an evaluation that uses a shorter window of $\pm$ 0.5 seconds is also performed). Following this convention, we use a $\pm$ 3-second window in our evaluation.

Let us assume that we have a set of correctly estimated boundaries given the annotated ones (hits), a set of annotated boundaries that are not estimated (false negatives), and a set of estimated boundaries that are not in the annotated dataset (false positives). Precision is the ratio between hits and the total number of estimated elements (e.g., we could have 100% precision with an algorithm that only returns exactly one boundary and this boundary is correct). Recall is the ratio between hits and the total number of annotated elements (e.g. we could have a 100% recall with an algorithm that returns one boundary every 3 seconds, since all the annotated boundaries will be sufficiently close to an estimated one). Precision and recall are defined formally as

$$P = \frac{|\text{hits}|}{|\text{bounds}_e|}; \quad R = \frac{|\text{hits}|}{|\text{bounds}_a|} \tag{1}$$

where $|\cdot|$ represents the cardinality of the set $\cdot$, bounds_e is the set of estimated boundaries and bounds_a is the set of annotated ones. Finally, the F-measure is the harmonic mean between P and R, which weights these two values equally, penalizes small outliers, and mitigates the impact of large ones:

$$F = 2\frac{P \cdot R}{P + R} \tag{2}$$

When listening to the output of music segmentation algorithms, it is immediately apparent that false negatives and false positives are perceptually very different (an initial discussion about assessing a *synthetic* precision of 100% when evaluating boundaries can be found in [14]). Thus, in the process of developing novel methods for structure segmentation, we decided to informally assess the relative effect that different types of errors had on human evaluations of the accuracy of the algorithms' outputs. The following section describes the resulting preliminary study.

2.2 Preliminary Study

For this study we compared three algorithms, which we will term $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. $\mathcal{A}$ is an unpublished algorithm currently in development that relies on homogeneous repeated section blocks; $\mathcal{B}$ is an existing algorithm that uses novelty in audio features to identify boundaries; and $\mathcal{C}$ combines the previous two methods. All three methods were optimized to maximize their F-measure performance on the structure-annotated Levy dataset [5]. Table 1 shows each method's average F-measure, precision, and recall values across the entire set. Note how $\mathcal{C}$ maximizes the F-measure, mostly by increasing recall, while $\mathcal{A}$ shows maximum precision.

We asked two college music majors to rank the three algorithms for every track in the Levy set. The goal was not to compare the results of the algorithms to the annotated ground truth, but to compare the algorithms with each other and determine the best one from a perceptual point of view. The participants were asked to listen to each of the algorithm outputs for all the songs and rank the algorithms by the quality of their estimated section boundaries; no particular constraints were given on what to look for. We used Sonic Visualiser [3] to display the waveform and three section panels for each of the algorithms in parallel (see Figure 1). While playing the audio, listeners could both see the sections and hear the boundaries indicated by a distinctive percussive sound. The section panels were organized at random for each song so listeners could not easily tell which algorithm they were choosing.

Preliminary Study			
Algorithm	F	P	R
$\mathcal{A}$	49%	57%	47%
$\mathcal{B}$	44%	46%	46%
$\mathcal{C}$	51%	47%	64%

Table 1. Algorithms and their ratings used to generate the input for the preliminary study. These ratings are averaged across the 60 songs of the Levy dataset.

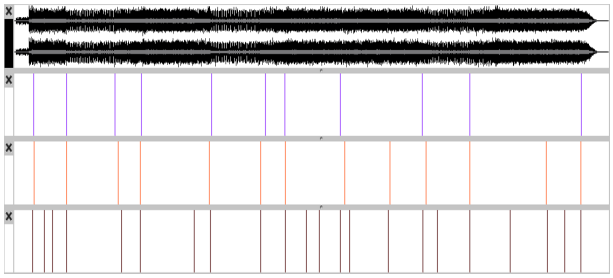

Figure 1. Screenshot of Sonic Visualiser used in the preliminary experiment. The song is "Smells Like Teen Spirit" by Nirvana. In this case, algorithms are ordered as $\mathcal{A}$, $\mathcal{B}$, and $\mathcal{C}$ from top to bottom.

Analysis of the results showed that 68.3% of the time, the two participants chose the same best algorithm. In 23.3% of the cases, they disagreed on the best, and in just 8.3% of the cases, they chose opposite rankings. When they actually agreed on the best algorithm, they chose $\mathcal{A}$ 58.5% of the time. $\mathcal{A}$ did not have the highest F-measure but the highest precision. Perhaps more surprising, they chose $\mathcal{C}$ only 14.6% of the time even though that algorithm had the highest F-measure.

These results raised the following questions: Is the F-measure informative enough to evaluate the accuracy of automatically estimated boundaries in a perceptually-meaningful way? Is precision more important than recall when assessing music boundaries? Would the observed trends remain when tested on a larger population of subjects? Can these results inform more meaningful evaluation measures? We decided to address these questions by running two more formal experiments in order to better understand this apparent problem and identify a feasible solution.

[1] http://www.music-ir.org/mirex/wiki/MIREX_HOME

3. EXPERIMENT 1: RATING BOUNDARIES

3.1 Motivation

The results of the preliminary study suggested that precision is more relevant than recall when perceiving boundaries. However, to fully explore this hypothesis, these two values had to be carefully manipulated. For this experiment, a set of boundaries was synthesized by setting specific values for precision and recall while maintaining a near-constant F-measure. Moreover, we wanted to ensure that the findings were robust across a larger pool of subjects. With these considerations in mind, the experiment was designed to be both shorter in time and available on line.

3.2 Methodology

We selected five track excerpts from the Levy catalog by finding the one-minute segments containing the highest number of boundaries across the 60 songs of the dataset. By having short excerpts instead of full songs, we could reduce the duration of the entire experiment with negligible effect on the results—past studies have shown that boundaries are usually perceived locally instead of globally [15]. We decided to use only five excerpts with the highest number of boundaries in order to maintain participants' attention as much as possible. For each track excerpt, we synthesized three different segmentations: ground truth boundaries (GT) with an F-measure of 100%; high precision (HP) boundaries with a precision of 100% and recall of around 65%; and high recall (HR) boundaries with a recall of 100% and precision of around 65%. The extra boundaries for the HR version were randomly distributed (using a normal distribution) across a 3 sec window between the largest regions between boundaries. For the HP version, the boundaries that were most closely spaced were removed. Table 2 presents F-measure, precision, and recall values for the five tracks along with the average values across excerpts. Note the closeness between F-measure values for HP and HR.

Experiment 1 Excerpt List						
Song Name (Artist)	HP			HR		
	F	P	R	F	P	R
Black & White (Michael Jackson)	.809	1	.68	.794	.658	1
Drive (R.E.M.)	.785	1	.647	.791	.654	1
Intergalactic (Beastie Boys)	.764	1	.619	.792	.656	1
Suds And Soda (Deus)	.782	1	.653	.8	.666	1
Tubthumping (Chumbawamba)	.744	1	.593	.794	.659	1
Average	.777	1	.636	.794	.659	1

Table 2. Excerpt list with their evaluations for experiment 1. The F-measure of GT is 100% (not shown in the table).

Subjects had to rate the "quality" of the boundaries for each version of the five tracks by choosing a discrete value between 1 and 5 (lowest and highest ratings respectively). Although this might arguably bias the subjects towards the existing boundaries only (reducing the influence of the missing ones), it is unclear how to design a similar experiment that would avoid this. Excerpts were presented in random order. Participants were asked to listen to all of the excerpts before submitting the results. As in the preliminary experiment, auditory cues for the section boundaries were added to the original audio signal in the form of a salient sharp sound. For this experiment, no visual feedback was provided because the excerpts were short enough for listeners to retain a general perception of the accuracy of the boundaries. The entire experiment lasted around 15 minutes (5 excerpts $\times$ 3 versions $\times$ one minute per excerpt) and was available on line [2] as a web survey in order to facilitate participation.

An announcement to various specialized mailing lists was sent in order to recruit participants. As such, most subjects had a professional interest in music, and some were even familiar with the topic of musical structure analysis. A total number of 48 participants took part in the experiment; subjects had an average of 3.1 $\pm$ 1.6 years of musical training and 3.7 $\pm$ 3.3 years of experience playing an instrument.

3.3 Results and Discussion

Box plots of accuracy ratings across versions can be seen in Figure 2. These experimental results show that higher accuracy ratings were assigned to GT followed by HP, and then HR.

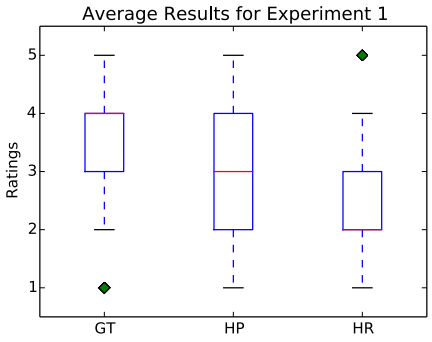

Figure 2. Average ratings across excerpts for Experiment 1; GT = ground truth; HP = high precition; HR = high recall.

A two-way, repeated-measures ANOVA was performed on the accuracy ratings with type (ground truth, high precision, high recall) and excerpt (the five songs) as factors. There were 48 data points in each Type $\times$ Excerpt category. The main effects of type, $F(2, 94) = 90.74$, $MSE = 1.10$, $p < .001$, and excerpt, $F(4, 188) = 59.84$, $MSE = 0.88$, $p < .001$, were significant. There was also an interaction effect, $F(6.17, 290.01) = 9.42$, $MSE = 0.74$, $p < .001$ (Greenhouse-Geisser corrected), indicating that rating profiles differed based on excerpt. Mean ratings by type and excerpt are shown in Figure 3.

Looking at the data for each excerpt, there was a clear pattern showing that subjects preferred segmentations with

[2] http://urinieto.com/NYU/BoundaryExperiment/

high precision over high recall (Figure 3). Post-hoc multiple comparisons indicated that differences between means of all three types were significant. The only excerpt where precision was not rated more highly than recall was in Excerpt 5 (Tubthumping), a difference that contributed primarily to the interaction. In this case, the excerpt contains a distinctive chorus where the lyrics "I get knocked down" keep repeating. This feature is likely the reason some subjects were led to interpret every instance of this refrain as a possible section beginning even though the harmony underneath follows a longer sectional pattern that is annotated in the ground truth. On the other hand, Excerpt 3 (Intergalactic) obtained similar ratings for ground truth and high precision, likely due to the high number of different sections and silences it contains. This can become problematic when extra boundaries are added (therefore obtaining poor ratings for the high-recall version). Nevertheless, given the subjectivity of this task [2] and the multi-layer organization of boundaries [10], it is not surprising that this type of variability appears in the results.

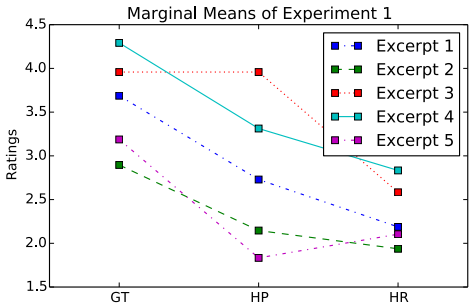

Figure 3. Means for excerpt and version of the results of Experiment 1.

The results of this experiment show that precision is more perceptually relevant than recall for the evaluation of boundaries, validating the preliminary findings (Section 2.2) in a controlled scenario and with a much larger population of subjects. Nevertheless, the number of tracks employed in this experiment was limited. As a follow-up, we explored these findings using a larger dataset in Experiment 2.

4. EXPERIMENT 2: CHOOSING BOUNDARIES

4.1 Motivation

The results of Experiment 1 show the relative importance of precision over recall for a reduced dataset of five tracks. However, it remains to be seen whether the F-measure, precision, and recall can predict a listener's preference when faced with a real-world evaluation scenario (i.e., boundaries not synthesized but estimated from algorithms). How this information can be used to redesign the metric to be more perceptually relevant is another question. In Experiment 2, we used excerpts sampled from a larger set of music, boundaries computed with state-of-the-art algorithms (thus recreating a real-world evaluation *à la* MIReX), and limited the evaluation to pairwise preferences.

4.2 Methodology

The analysis methods used to compute the boundaries included structural features (SF, [12]), convex non-negative matrix factorization (C-NMF, [7]), and shift-invariant probabilistic latent component analysis (SI-PLCA, [17]). These three algorithms yield ideal results for our experimental design since SF provides one of the best results reported so far on boundaries recognition (high precision and high recall) footnoteRecently challenged by Ordinal Linear Discriminant Analysis [6]. C-NMF tends to over segment (higher recall than precision), and SI-PLCA, depending on parameter choices, tends to under segment (higher precision than recall).

We ran these three algorithms on a database of 463 songs composed of the conjunction of the TUT Beatles dataset, [3] the Levy catalogue [5], and the freely available songs of the SALAMI dataset [13]. Once computed, we filtered the results based on the following criteria for each song: (1) at least two algorithm outputs have a similar F-measure (within a 5% threshold); (2) the F-measure of both algorithms must be at least 45%; (3) at least a 10% difference between the precision and recall values of the two selected algorithm outputs exists.

We found 41 out of 463 tracks that met the above criteria. We made a qualitative selection of these filtered tracks (there are many free tracks in the SALAMI dataset that are live recordings with poor audio quality or simply speech), resulting in a final set of 20 songs. The number of these carefully selected tracks is relatively low, but we except it to be representative enough to address our research questions. Given the two algorithmic outputs maximizing the difference between precision and recall, two differently segmented versions were created for each track: high precision (HP) and high recall (HR). Moreover, similar to Experiment 1, only one minute of audio from each track was utilized, starting 15 seconds into the song.

Table 3 shows average metrics across the 20 selected tracks. The F-measures are the same, while precision and recall vary.

Boundaries Version	F	P	R
HP	.65	.82	.56
HR	.65	.54	.83

Table 3. Average F-measure, precision, and recall values for the two versions of excerpts used in Experiment 2.

As in Experiment 1, the interface for Experiment 2 was on line [4] to facilitate participation. Each participant was presented with five random excerpts selected from the set of 20. Instead of assessing the accuracy on a scale, listeners had to choose the version they found more accurate. In order uniformly distribute excerpts across total trials, selection of excerpts was constrained by giving more priority to those excerpts with fewer collected responses. We obtained an average of 5.75 results per excerpt. The two versions were presented in random order, and subjects had

[3] http://www.cs.tut.fi/sgn/arg/paulus/beatles_sections_TUT.zip
[4] http://cognition.smusic.nyu.edu/boundaryExperiment2/

to listen to the audio at least once before submitting the results. Boundaries were marked with a salient sound like in the prior experiments.

A total 23 subjects, recruited from professional mailing lists, participated in the experiment. Participants had an average of 2.8 $\pm$ 1.4 years of musical training and 3.2 $\pm$ 2.9 years of experience playing an instrument.

4.3 Results and Discussion

We performed binary logistic regression analysis [11] on the results with the goal of understanding what specific values of the F-measure were actually useful in predicting subject preference (the binary values representing the versions picked by the listeners). Logistic regression enables us to compute the following probability:

$$P(Y|X_1,\ldots,X_n) = \frac{e^{k+\beta_1 X_1+\ldots+\beta_n X_n}}{1+e^{k+\beta_1 X_1+\ldots+\beta_n X_n}} \quad (3)$$

where Y is the dependent, binary variable, X_i are the predictors, β_i are the weights for these predictors, and k is a constant value. Parameters β_i and k are learned through the process of training the regressor. In our case, Y tells us whether a certain excerpt was chosen or not according to the following predictors: the F-measure (X_1), the signed difference between precision and recall (X_2), and the absolute difference between precision and recall (X_3).

Since 23 subjects took part in the experiment and there were five different tracks with two versions per excerpt, we had a total of $23 \times 5 \times 2 = 230$ observations as input to the regression with the parameters defined above. We ran the Hosmer & Lemeshow test [4] in order to understand the predictive ability of our input data. If this test is not statistically significant ($p > 0.05$), we know that logistic regression can indeed help us predict Y. In our case, we obtain a value of $p = .763$ ($\chi^2 = 4.946$, with 8 degrees of freedom) which tells us that the data for this type of analysis fits well, and that the regressor has predictive power.

The analysis of the results of the learned model is shown in Table 4. As expected, the F-measure is not able to predict the selected version ($p = .992$), providing clear evidence that the metric is inexpressive and perceptually irrelevant for the evaluation of segmentation algorithms. Furthermore, we can see that $P-R$ can predict the results in a statistically significant manner ($p = .000$), while the absolute difference $|P-R|$, though better than the F-measure, has low predictive power ($p = .482$). This clearly illustrates the asymmetrical relationship between P and R: it is not sufficient that P and R are different, but the sign matters: P has to be higher than R.

Based on this experiment we can claim that, for these set of tracks, (1) the F-measure does not sufficiently characterize the perception of boundaries, (2) precision is clearly more important than recall, and (3) there might be a better parameterization of the F-measure that encodes relative importance. We attempt to address this last point in the next section.

Logistic Regression Analysis of Experiment 2						
Predictor	β	S.E. β	Wald's χ^2	df	p	e^{β}
F-measure	-.012	1.155	.000	1	.992	.988
$P-R$	2.268	.471	23.226	1	.000	1.023
$\|P-R\|$	-.669	.951	.495	1	.482	.512
k	.190	.838	.051	1	.821	1.209

Table 4. Analysis of Experiment 2 data using logistic regression. According to these results, $P-R$ can predict the version of the excerpt that subjects will choose.

5. ENHANCING THE F-MEASURE

Based on our experiments, we have empirical evidence that high precision is perceptually more relevant than high recall for the evaluation of segmentation algorithms. We can then leverage these findings to obtain a more expressive and perceptually informative version of the F-measure for benchmarking estimated boundaries.

The F-measure is, in fact, a special case of the F_α-measure:

$$F_\alpha = (1+\alpha^2)\frac{P \cdot R}{\alpha^2 P + R} \quad (4)$$

where $\alpha = 1$, resulting in P and R having the same weight. However, it is clear from the equation that we should impose $\alpha < 1$ in order to give more importance to P and make the F-measure more perceptually relevant. Note that an algorithm that outputs fewer boundaries does not necessarily increase its F_α-measure, since the fewer predicted boundaries could still be incorrect. Regardless, the question remains: how is the value of α determined?

A possible method to answer this question is to sweep α from 0 to 1 using a step size of 0.05 and perform logistic regression analysis at each step using the F_α-measure as the only predictor (X_1=F_α, $n = 1$). The p-value of the F_α-measure predicting subject preference in Experiment 2 across all α is shown in Figure 4.

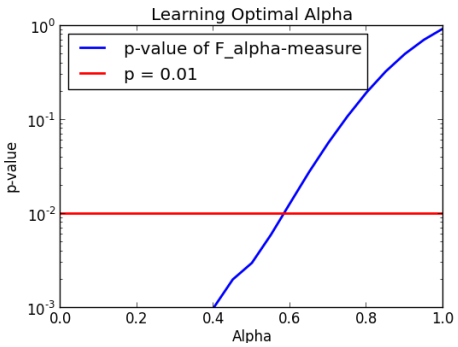

Figure 4. Statistical significance of the F_α-measure predicting the perceptual preference of a given evaluation for $\alpha \in [0,1]$

It is important to note that the data from Experiment 2 is limited as it does not include information at the limits of the difference between precision and recall. As a result, our model predicts that decreases of α always lead to highest predictive power. Naturally, this is undesirable since we will eventually remove all influence from recall in the measure and favor the trivial solutions discussed at

the beginning of this paper. At some point, as $P - R$ increases, we expect subject preference to decrease, as preserving a minimum amount of recall becomes more important. Therefore, we could choose the first value of α (0.58) for which F_α-based predictions of subject preference become accurate at the statistically significant level of 0.01.

We can re-run the evaluation of Experiments 1 and 2 using the $F_{0.58}$-measure (i.e. $\alpha = 0.58$) to illustrate that it behaves as expected. For Experiment 1, we obtain 83.3% for HP and 72.1% for HR (instead of 77.7% and 79.4% respectively). For Experiment 2, the values of HP and HR become 71.8% and 58.9% respectively, whereas they were both 65.0% originally. This shows how the new approximated measure is well coordinated with the preferences of the subjects from Experiments 1 and 2, therefore making this evaluation of section boundaries more expressive and perceptually relevant.

This specific α value is highly dependent on the empirical data, and we are aware of the limitations of using reduced data sets as compared to the real world—in other words, we are likely overfitting to our data. Nonetheless, based on our findings, there must be a value of $\alpha < 1$ that better represents the relative importance of precision and recall. Future work, utilizing larger datasets and a greater number of participants, should focus on understanding the upper limit of the difference between precision and recall in order to find the specific inflection point at which higher precision is not perceptually relevant anymore.

6. CONCLUSIONS

We presented a series of experiments concluding that precision is perceived as more relevant than recall when evaluating boundaries in music. The results of the two main experiments discussed here are available on line. [5] Moreover, we have noted the shortcomings of the current F-measure when evaluating results in a perceptually meaningful way. By using the general form of the F-measure, we can obtain more relevant results when precision is emphasized over recall ($\alpha < 1$). Further steps should be taken in order to determine a more specific and generalizable value of α.

7. ACKNOWLEDGMENTS

This work was partially funded by Fundación Caja Madrid and by the National Science Foundation, under grant IIS-0844654.

8. REFERENCES

[1] G. Boutard, S. Goldszmidt, and G. Peeters. Browsing Inside a Music Track, The Experimentation Case Study. In *Proc. of the Workshop of Learning the Semantics of Audio Signals*, pages 87–94, 2006.

[2] M. J. Bruderer, M. F. Mckinney, and A. Kohlrausch. The Perception of Structural Boundaries in Melody Lines of Western Popular Music. *MusicæScientiæ*, 13(2):273–313, 2009.

[3] C. Cannam, C. Landone, M. Sandler, and J. P. Bello. The Sonic Visualiser: A Visualisation Platform for Semantic Descriptors from Musical Signals. In *Proc. of the 7th International Conference on Music Information Retrieval*, pages 324–327, Victoria, BC, Canada, 2006.

[4] D. W. Hosmer and S. Lemeshow. *Applied Logistic Regression*. John Wiley & Sons, 2004.

[5] M. Levy and M. Sandler. Structural Segmentation of Musical Audio by Constrained Clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):318–326, Feb. 2008.

[6] B. McFee and D. P. W. Ellis. Learning to Segment Songs with Ordinal Linear Discriminant Analysis. In *Proc. of the 39th IEEE International Conference on Acoustics Speech and Signal Processing*, Florence, Italy, 2014.

[7] O. Nieto and T. Jehan. Convex Non-Negative Matrix Factorization For Automatic Music Structure Identification. In *Proc. of the 38th IEEE International Conference on Acoustics Speech and Signal Processing*, pages 236–240, Vancouver, Canada, 2013.

[8] B. S. Ong and P. Herrera. Semantic Segmentation of Music Audio Contents. In *Proc. 32nd of the International Computer Music Conference*, Barcelona, Spain, 2005.

[9] J. Paulus, M. Müller, and A. Klapuri. Audio-Based Music Structure Analysis. In *Proc of the 11th International Society of Music Information Retrieval*, pages 625–636, Utrecht, Netherlands, 2010.

[10] G. Peeters and E. Deruty. Is Music Structure Annotation Multi-Dimensional? A Proposal for Robust Local Music Annotation . In *Proc. of the 3rd International Worskhop on Learning Semantics of Audio Signals*, pages 75–90, Graz, Austria, 2009.

[11] C.-Y. J. Peng, K. L. Lee, and G. M. Ingersoll. An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, 96(1):3–14, Sept. 2002.

[12] J. Serrà, M. Müller, P. Grosche, and J. L. Arcos. Unsupervised Music Structure Annotation by Time Series Structure Features and Segment Similarity. *IEEE Transactions on Multimedia, Special Issue on Music Data Mining*, 2014.

[13] J. B. Smith, J. A. Burgoyne, I. Fujinaga, D. De Roure, and J. S. Downie. Design and Creation of a Large-Scale Database of Structural Annotations. In *Proc. of the 12th International Society of Music Information Retrieval*, pages 555–560, Miami, FL, USA, 2011.

[14] J. B. L. Smith. *A Comparison And Evaluation Of Approaches To The Automatic Formal Analysis Of Musical Audio*. Master's thesis, McGill University, 2010.

[15] B. Tillmann and E. Bigand. Global Context Effect in Normal and Scrambled Musical Sequences. *Journal of Experimental Psychology: Human Perception and Performance*, 27(5):1185–1196, 2001.

[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. of the 5th International Society of Music Information Retrieval*, pages 42–49, Vienna, Austria, 2007.

[17] R. Weiss and J. P. Bello. Unsupervised Discovery of Temporal Structure in Music. *IEEE Journal of Selected Topics in Signal Processing*, 5(6):1240–1251, 2011.

[5] http://www.urinieto.com/NYU/ISMIR14-BoundariesExperiment.zip