2013 LATE-BREAK SESSION ON MUSIC SEGMENTATION

Oriol Nieto
New York University
New York, NY, USA
oriol@nyu.edu

Jordan B. L. Smith
Queen Mary University of London
London, UK
j.smith@qmul.ac.uk

1. INTRODUCTION

Given the great number of publications on music segmentation at ISMIR 2013 (e.g. [1, 4–7, 10, 11, 14]), a special session on this topic was organized at the end of the conference with a strong participation. We discussed the correctness of the current evaluation metrics —both for the boundaries and the labels— and how they could potentially better capture the subjectivity of the task itself. We debated the importance of either having higher precision or higher recall when examining boundary results. We finally discussed how the ground truth data could be improved, by using a more standardized definition of music segments, and having as many annotations as possible.

2. EVALUATION METRICS

2.1 Boundaries

The boundaries of the segments have been typically assessed by using 0.5 and/or 3-second windows [16]. We discussed how a 3-second window can be perceptually too large in order to successfully quantify the quality of a given boundary; such deviations can miss the downbeat quite a lot, and annotations tend to be synchronized to downbeats. This, plus the fact that human annotations do not seem to exactly agree [3, 13], made us debate other possible evaluations.

A different method to evaluate the boundaries using a Gaussian window instead of a square window was discussed. It would be an interesting task to design an experiment to decide the amount of variance of the Gaussian window, and maybe the actual shape of the window, which might result in an asymmetric Gaussian window. Using a Gaussian window implies a probabilistic view of the existence and placement of boundaries. This view may be more in keeping with findings of perceptual studies such as [3], who have found disagreements among human annotators common.

The precision and recall values of the boundaries were also discussed. Nieto argued that in listener experiments, high precision tended to be more important to listeners than higher recall. However, it was also pointed out that the choice of metric must always be suited to the purpose of the evaluation: in some applications (say, an augmented music browsing application), recall may be more important. Following this example, and as opposed to Nieto's findings, Peeters found that recall is usually more significant. Further research in order to explain these inconsistencies should be performed in the future.

We also discussed the practice of including the beginning and ending of the annotation in the set of boundaries. Both of these points are trivial to retrieve, but many still use them in the evaluation, perhaps artificially inflating results over the baseline (e.g. precision of 1 can be trivially obtained by effectively making no guesses). We agreed that good evaluation practices should be codified in evaluation scripts and these made available publicly, a subject that was discussed in more detail at the separate late-breaking session on MIREX Evaluation.

2.2 Labels

The evaluation of the labels is usually performed using the pair-wise clustering [8] and the Entropy scores [9]. The latter scores were proposed as an improvement to the pairwise evaluation, since, as it is shown in [9], they seem more robust against imprecise boundaries, and strongly penalize randomly chosen labels.

McFee and Nieto included the F_1-measure between the undersegmentation (S_u) and oversegmentation (S_o) scores in their new publications, and argued whether this was a good practice. This might help comparing results, but in some cases it is important to also investigate the absolute difference between S_u and S_o, since we generally try to obtain results that are as uniform as possible in their values of the F_1-measure (i.e. we usually want to keep $|S_u - S_o|$ as small as possible). In any event, we suggested to also include this F_1-measure in the MIREX evaluation, along with S_o and S_u.

3. DATASETS

Another important discussion during the session was about the significance of the datasets for music segmentation, and how we can improve them. We agree that given the subjectivity of the task, we want as many human annotations as possible. We mentioned the SALAMI dataset [15], for being the first dataset on music segmentation to contain more than one human annotation per song. However, as far as the authors know, there have been no publications on how to aggregate multiple annotations yet. Bruderer showed how humans tend to agree on how salient a boundary is [2], so multiple annotations of a single track might help setting up a *salience score* for each boundary that could help evaluating an algorithm against a dataset.

Finally, it should be necessary to formally propose a general definition of *musical segment* in order to gather more consistent and normalized datasets. We should agree on whether all boundaries must fall on downbeats, or the numbers of layers in order to label the segments (like in SALAMI or in [12]). We believe that agreeing on various of these aspects would yield datasets that would facilitate the evaluation of music segmentation systems more efficiently.

4. CONCLUSIONS

Music segmentation is an important task in MIR that keeps challenging the research community. Given the complexity and the subjectivity of the problem, we might need to reconsider how we evaluate our algorithms and how we gather more consistent datasets. We discussed how some of the numbers in the evaluation might be misleading, and we proposed to use the F_1-measure between the undersegmentation and oversegmentation scores to be included in future MIREX editions. We also encourage researches to gather as many annotations per track in new datasets as possible, even though it is still unclear how to exploit this information in the right direction.

5. REFERENCES

[1] Jan Van Balen, John Ashley, Burgoyne Frans, and Wiering Remco. An Analysis of Chorus Features in Popular Song. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[2] Michael J Bruderer. *Perception and Modeling of Segment Boundaries in Popular Music*. PhD thesis, Universiteitsdrukkerij Technische Universiteit Eindhoven, 2008.

[3] Michael J Bruderer, Martin McKinney, and Armin Kohlrausch. Structural Boundary Perception in Popular Music. In *Proc of the International Society of Music Information Retrieval*, volume 4, pages 198–201, Victoria, BC, Canada, 2006.

[4] Pranay Dighe, Harish Karnick, and Bhiksha Raj. Swara Histogram Based Structural Analysis and Identification of Indian Classical Ragas. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[5] Harald Grohganz, Michael Clausen, Nanzhu Jiang, and Meinard Müller. Converting Path Structures into Block Structures using Eigenvalue Decomposition of Self-Similarity Matrices. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[6] Nanzhu Jiang and Meinard Müller. Automated Methods for Analyzing Music Recordings in Sonata Form. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[7] Florian Kaiser and Geoffroy Peeters. A Simple Fusion Method of State and Sequence Segmentation for Music Structure Discovery. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

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

[9] Hanna Lukashevich. Towards Quantitative Measures of Evaluating Song Segmentation. In *Proc. of the 10th International Society of Music Information Retrieval*, pages 375–380, Philadelphia, PA, USA, 2008.

[10] Oriol Nieto and Morwaread Farbood. MIREX 2013: Discovering Musical Patterns Using Audio Structural Segmentation Techniques. In *Music Information Retrieval Evaluation eXchange*, Curitiba, Brazil, 2013.

[11] Johan Pauwels, Florian Kaiser, and Geoffroy Peeters. Combining Harmony-based and Novelty-based Approaches for Structural Segmentation. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[12] Geoffroy Peeters and Emmanuel 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.

[13] Joan Serrà, Meinard Müller, Peter Grosche, and Josep Lluís Arcos. Unsupervised Detection of Music Boundaries by Time Series Structure Features. In *Proc. of the 26th AAAI Conference on Artificial Intelligence*, number 2009, pages 1613–1619, Toronto, Canada, 2012.

[14] Diego F. Silva, Gustavo E. Batista, and Daniel P. W. Ellis. A Video Compression-based Approach to Measure Music Structure Similarity. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.

[15] Jordan B. Smith, J. Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J. Stephen 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.

[16] Jordan B. L. Smith and Elaine Chew. A Meta-Analysis of the MIREX Structure Segmentation Task. In *Proc. of the 14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, 2013.