ESTIMATION OF THE RELIABILITY OF MULTIPLE RHYTHM FEATURES EXTRACTION FROM A SINGLE DESCRIPTOR

Elio Quinton, Mark Sandler, Simon Dixon
Center for Digital Music, Queen Mary University of London

ABSTRACT

The provision of a reliability or confidence measure can be critical for the usage of a given feature in complex systems and real-world applications. However, feature extraction systems often do not provide one. In the present study we investigate the relationship between the entropy of a rhythmogram and the reliability of the extraction of multiple high level rhythm related features. The results show that this single descriptor has potential for estimating the reliability of multiple rhythm features extraction.

1. INTRODUCTION

It has been extensively reported in the Music Information Retrieval (MIR) literature that it is difficult to reliably extract high-level rhythm related features from musical excerpt having properties such as soft onsets, heavy syncopation or expressive timing (e.g. rubato playing). There is relatively little effort in quantifying this, however. On the other hand, combining features or using one feature to inform the extraction of another (e.g. beat synchronous chromagram) has appeared to be a fruitful approach [1]. Indicators such as 'beat strength' [10] and 'pluse clarity' [6] have been proposed and directly evaluated against human judgment, but these studies did not investigate the impact of such an attribute on the extraction of related rhythm features. The estimation of the difficulty of feature extraction has received some attention in the particular case of beat-tracking [3–5]. In recent work, it has been demonstrated that the entropy of a cyclic tempogram can be used as an indicator of the tempo salience of a musical piece [8]. However, this feature is sensitive a number of properties of the musical signal, such as expressive timing [5] or strong syncopation [8], which have been reported to be problematic for high level rhythm features extraction such as beat tracking or tempo estimation. In this paper we show that the entropy of a rhythmogram has potential to be interpreted as a single estimate of the reliability of the automatic estimation of several high level rhythm features.

2. EXPERIMENT

In our experiment, we consider three feature extraction procedures, namely tempo, metrical structure estimation and beat tracking. Tempo estimation is performed using the Vamp plugin implementation [1] of the algorithm described in [2]. The metrical structure is extracted using the method described in [7]. Both are evaluated on the GTZAN dataset [9] using the corresponding annotations [2]. A tempo estimate is considered correct if it equals the corresponding annotation within a tolerance window of 8%, consistently with the standard adopted in the MIREX audio tempo evaluation task [3]. For each track, the quality of the metrical structure estimate is charaterised by an F-measure, we refer to [7] for a more detailed description. As per the beat tracking, the evaluation results are drawn from [5]. For each musical excerpt considered in this paper, we computed the average rhythmogram entropy and investigate its relationship with rhythm feature extraction performance.

3. RESULTS

The percentage of successful tempo estimation for each entropy class is given in Figure 1. The apparent trend in this data suggests that the tempo extraction accuracy decreases as the rhythmic entropy increases. The Pearson, Spearman and Kendall reveal a strong negative linear relationship between entropy and tempo estimation accuracy [4] with $p < 0.001$ in all cases. On the other hand, The Pearson, Spearman and Kendall coefficients do not reveal significant direct linear nor monotonic correlation between entropy and the metrical structure estimation performance. However, the performance distribution shown on Figure 2 suggests a tendency for the performance, characterised by the F-measure, to be relatively consistent up until the entropy reaches values around 0.8, when a clear decrease of both mean performance and performance consistency (characterised by the spread of the distribution) is observed. We ran a two sample Welch t-test on F-measures distributions belonging to adjacent entropy classes in order to assess the statistical significance of the drop in mean performance. The results confirm that the decrease of mean performance observed for entropy values higher than 0.8

Attribution: Elio Quinton, Mark Sandler, Simon Dixon. "Estimation of the reliability of multiple rhythm features extraction from a single descriptor", 16th International Society for Music Information Retrieval Conference, 2015.

[1] http://www.vamp-plugins.org/download.html

[2] For tempo: http://www.marsyas.info/tempo/genres_tempos.mf
For metrical structure: cf. [7]

[3] http://www.music-ir.org/mirex/wiki/2015:Audio_Tempo_Estimation

[4] Pearson = -0.947, Kendall = Spearman = 1.0

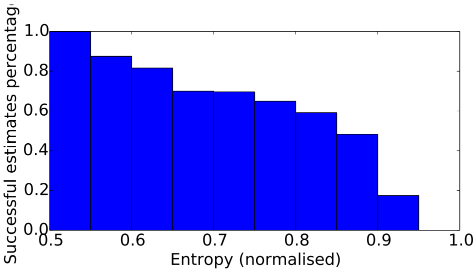

Figure 1. Mean tempo extraction accuracy (proportion of correct estimations) for different entropy classes.

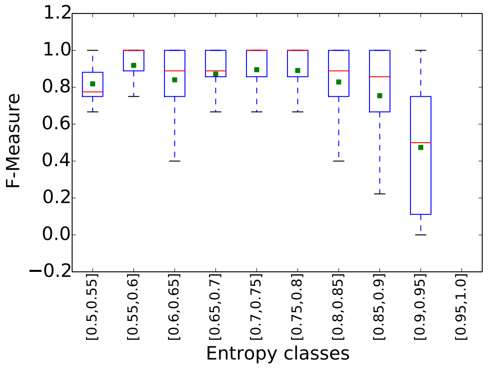

Figure 2. F-measure distribution for each entropy class. Mean is represented by a green dot, and median by a red line.

is statistically significant at the 0.001 level. As a product of the assessment of beat tracking difficulty by beat trackers disagreement proposed in [5], the authors composed a dataset of 'hard' tracks for beat tracking. Such tracks are characterised by their propensity to result in disagreement between beat trackers, and by extension in unreliable beat estimates. Alongside with the hard tracks, the authors provided 'easy' tracks, which result in good and reliable beat estimates. The entropy distribution for 'Hard' and 'Easy' categories are graphically set apart in Figure 3. In addition, we performed a two sample Welch's t-test that strongly rejected the null hypothesis of equal means of the two distributions at the 0.001 level. In other words, the beat tracking difficulty, and thereby the reliability of the beat estimates, that had been estimated using beat tracker disagreement, is also related on average to measurement of the rhythmogram entropy.

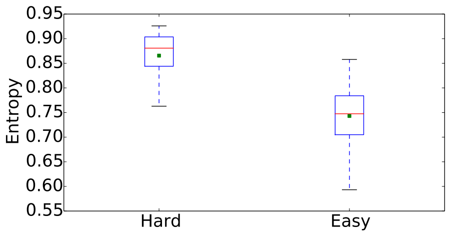

Figure 3. Entropy distribution for the dataset published in [5].

4. CONCLUSION

The results presented in this paper show that the entropy of a rhythmogram is statistically related to the reliability of the extraction of multiple high-level rhythm features, a higher entropy typically being related to lower feature extraction reliability. This descriptor therefore shows potential to provide a reliability or confidence value, therefore significantly increasing the features usability in complex systems and real-world applications.

5. REFERENCES

[1] Juan Pablo Bello and Jeremy Pickens. A Robust Mid-Level Representation for Harmonic Content in Music Signals. In *ISMIR*, volume 5, pages 304–311, 2005.

[2] Matthew EP Davies and Mark D. Plumbley. Context-dependent beat tracking of musical audio. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(3):1009–1020, 2007.

[3] Masataka Goto. An audio-based real-time beat tracking system for music with or without drum-sounds. *Journal of New Music Research*, 30(2):159–171, 2001.

[4] Peter Grosche, Meinard Mller, and Craig Stuart Sapp. What Makes Beat Tracking Difficult? A Case Study on Chopin Mazurkas. In *ISMIR*, pages 649–654, 2010.

[5] Andre Holzapfel, Matthew EP Davies, Jos R. Zapata, Joo Lobato Oliveira, and Fabien Gouyon. Selective sampling for beat tracking evaluation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(9):2539–2548, 2012.

[6] Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and Jose Fornari. Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimization. In *ISMIR*, pages 521–526. Citeseer, 2008.

[7] Elio Quinton, Christopher Harte, and Mark Sandler. Extraction of Metrical Structure from Music Recordings. In *Proc. of the 18th Int. Conference on Digital Audio Effects (DAFx). Trondheim, Norway, Nov 30 - Dec 3, 2015*, 2015.

[8] Balaji Thoshkahna, Meinard Muller, Venkatesh Kulkarni, and Nanzhu Jiang. Novel Audio Features for Capturing Tempo Salience in Music Recordings. In *Acoustics Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, Brisbane, Australia, 2015. IEEE.

[9] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Speech and Audio Processing, IEEE transactions on*, 10(5):293–302, 2002.

[10] George Tzanetakis, Georg Essl, and Perry Cook. Human perception and computer extraction of musical beat strength. In *Proc. DAFx*, volume 2, 2002.