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 design of systems for automatic audio feature extraction is a central aspect of the field of Music Information Retrieval. However, feature extraction systems often do not provide an indication of the reliability of the corresponding feature. Nevertheless, 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. In the present study we investigate the relationship between the entropy of a rhythmogram, which has been proposed as a descriptor of tempo salience in previous work, and the reliability of the extraction of multiple high level rhythm related features. The results show that this single descriptor is viable for simultaneously estimating the reliability of multiple rhythm features extraction. The results also provide quantitative insight that is consistent with qualitative observations extensively reported in the literature on a qualitative basis.

Index Terms— MIR, rhythm, meter, tempo, beat tracking

1. INTRODUCTION

The design of systems for automatic audio feature extraction is a central aspect of the field of Music Information Retrieval (MIR). 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, 2]. However, feature extraction systems often do not provide an indication of the reliability of the corresponding feature. Nevertheless, the provision of a reliability or confidence measure can be critical for usage of a given feature in complex systems and real-world applications [3]. In this paper we focus on the case of high-level rhythm features, namely tempo, beat positions and metrical structure.

It has been extensively reported in the 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 making use of expressive timing (e.g. rubato playing). There is relatively little effort in quantifying this, however. In an attempt to estimate related characteristics of the musical signal, the extraction of indicators such as 'beat strength' [4] and 'pluse clarity' [5] have been proposed. These studies provided a direct evaluation of estimation of 'beat strength' or 'pulse clarity' against human judgment, but 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 [6]. Goto used the difference of the power on the beat, and the power on other positions to assess the beat tracking difficulty of a song [7]. An alternative approach to beat tracking difficulty estimation is based on disagreement in a committee of beat trackers, so that a disagreement suggests a difficult case for beat tracking [8].

In recent work, Thoshkahna demonstrated that the entropy of a cyclic tempogram [9] can be used as an indicator of the tempo salience of a musical piece [10]. However, the tempogram feature captures multiple properties of the musical signal related to what have been reported as problematic for high level rhythm features extraction such as beat tracking or tempo estimation. Such properties could be expressive timing [8] (resulting in a widening of the horizontal lines in a rhythmogram) or strong syncopation (resulting in an overall blurring of the rhythmogram) [10]. In this paper we show that the entropy of a rhythmogram can be interpreted as a single estimate of the reliability of the automatic estimation of several high level rhythm features. In section 2 we describe the rhythm salience feature we used. The experiment and results are presented in section 3 and section 4 respectively, and the conclusions are drawn in section 5.

2. RHYTHM SALIENCE FEATURE

The rhythm salience feature used in this paper is derived from the processing proposed by Thoshkahna [10]. From the audio signal, an onset detection function is computed using the *Superflux* method, which improves on the spectral flux method [11] by incorporating robustness against vibrato using a maximum filter [12]. A rhythmogram $\mathcal{R}_F(t, f)$ is then generated as the Fourier transform based magnitude spectrogram of the onset detection function, using 12 seconds long Hanning windows and 0.2 seconds step. Although the terms 'tempogram' and 'rhythmogram' refer to the same processing, we will favour the term 'rhythmogram' in the remainder of this

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) and Omnifone

paper because this feature captures more information than just the tempo [13, 14]. The cyclic variation of the rhythmogram (which is not used in this study) consists in wrapping it over one octave so that pulses related by a power of two are identified, by analogy with the chromagram processing [9].

The columns of the rhytmogram are normalised with respect to the L^1 norm, and for a vector $x = (x_1, \cdots, x_m)$ in $\mathbb{R}^M$ representing the k^{th} rhythmogram frame, the entropy S_k is defined as:

$$S_k = \frac{\sum_{m=1}^{M} -x_m \log_2(x_m)}{\log_2(M)} \quad (1)$$

with M the number of frequency bins in the rhythmogram.

The entropy, commonly used as a measure of disorder, or uncertainty in a probabilistic framework, takes high values for uniform distributions of energy in vector x, and small values for highly organised, and therefore uneven, distributions. The former corresponds to musical signals having no consistent or clear pulse emerging in the analysis window and the latter corresponds to musical signals with very salient pulse [10]. Considering the length of the analysis windows, the consistency of the rhythmic properties of the music also impacts the results: large inconsistencies — that could correspond for example to heavy syncopation or rubato playing — would typically results in a blurring of the rhythmogram, and therefore in an increased entropy value. These inconsistencies are known for creating challenging conditions for high level rhythm features extraction, and therefore motivates our consideration of the entropy as a descriptor for feature extraction reliability in the present study.

3. EXPERIMENT

The experiment is structured as follows. First, several features are extracted from audio and the performance of the extraction is evaluated using standard metrics. Secondly, a measure of the rhythmic salience is computed for every track according to the method specified in section 2. We then investigate how it relates to feature extraction performance. It is important to note at this point that the aim of this paper is not to investigate nor improve feature extraction or evaluation methods — we refer to relevant literature for this purpose. Instead, we focus on the analysis of the relationship between rhythm salience and feature extraction performance.

We consider three feature extraction procedures, namely tempo and metrical structure estimation and beat tracking. Tempo estimation is performed using the Vamp plugin implementation[1] of a two-state context dependent algorithm [15]. The metrical structure is extracted based on prior work by Quinton et. al [13]. As per the beat tracking, the evaluation results are drawn from a prior study on beat tracking evaluation [8]. Two publicly available datasets are used to carry out the estimation of feature extraction reliability. The GTZAN dataset [16] is used in the case of tempo and metrical structure, alongside with the corresponding annotations for tempo[2] and metrical structure [13]. For these two features we use the track-level average values of rhythmic features. The tracks of the GTZAN dataset being 30 seconds long and of overall reasonable consistency (in other words they do not contain a lot of musical changes), the track-level average is a reasonable estimate of the track content.

For each track the estimated tempo is compared to the annotated tempo, and considered correct if they are equal within a tolerance window of 8% of the annotated value, consistently with the standard adopted in the MIREX audio tempo evaluation task[3]. We refer to the original publication for a detailed description of the metrical hierarchy feature extraction evaluation metrics [13]. For each track, the extracted feature consists of an estimate of the pulse rate of all the metrical levels present in the music. They are then compared with the corresponding annotations and the result is summarised by an F-measure. In both cases the hypothesis is that the feature extraction procedure is considered reliable if it consistently matches the human annotations. In the case of beat tracking, we rely on the difficulty assessment by disagreement in a committee of beat trackers [8]. The authors used this method to compose a dataset of 40 seconds long 'hard' and 'easy' musical excerpts. The hard tracks were chosen for their propensity to generate disagreement in the committee, that is to say unreliable beat estimates. Conversely, reliable beat estimates are consistently produced for 'easy' tracks. For each musical excerpt considered in this paper, we computed the rhythmogram entropy according to equation 1, and an average entropy value S is obtained by averaging the values for each frame S_k.

4. RESULTS

In this section we analyse the relationship between the rhythmogram entropy and the performance of rhythm features extraction algorithms, evaluated according to the methods described in section 3. The evaluation procedures being feature specific, the results are presented on a per-feature basis.

4.1. Metrical structure

We first investigate the existence of a linear correlation between the entropy and the performance F-measure for all the songs. The Pearson, Spearman and Kendall coefficients were computed and are presented in Table 2. Pearson coefficient is a measure of linear correlation and results in absolute values between 0 (no correlation) and 1 (maximum correlation). The very low value observed here does not reveal a significant linear correlation between entropy and the algorithm performance. Similarly, Spearman and Kendall coefficients produce absolute values between 0 and 1, measuring the monotonic

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

[2] http://www.marsyas.info/tempo/genres_tempos.mf

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

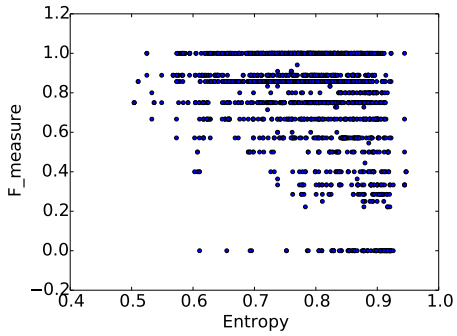

Fig. 1. Metrical structure feature extraction performance, given by the F-measure, against track mean entropy. Each dot on the graph represents the results of the evaluation for a track of the GTZAN dataset.

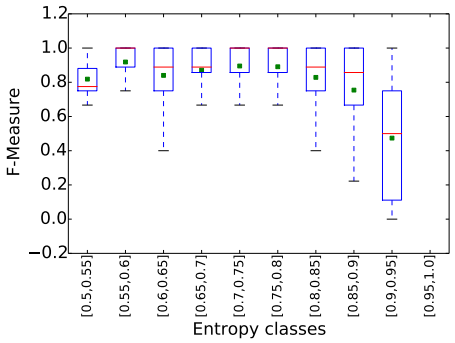

Fig. 2. F-measure distribution for each entropy class. Mean is represented by a green dot, and median by a red line.

relationship between entropy and algorithm performance. Again, low values suggest the absence of significant monotonic correlation. This evidence is graphically corroborated by the scatter plot of Figure 1. However, it is salient on this plot that the bottom left area contains little to no points, which seem to indicate a tendency in the distribution, despite the absence of linear correlation: tracks with low entropy tend to consistently lead to good performance while tracks with high entropy result in inconsistent performance.

In order to gain more statistical insight, data points are now grouped by entropy classes. Figure 2 shows the boxplot of the distribution of the feature extraction performance F-measure for each entropy class. Although the [0.6,0.65] class appears as a relative outlier, it suggests a tendency for the performance characterised by the F-measure to be relatively consistent up until the entropy reaches values around 0.8, and a clear decrease of both mean performance and performance consistency (characterised by the spread of the distribution) is observed. In order to assess the statistical significance of the drop in mean performance, we run a two sample Welch t test on F-measures distributions belonging to adjacent entropy classes. The results are shown in Table 1 and confirm that the decrease of mean performance observed for entropy values higher than 0.8 is statistically significant at the 0.001 level. The distributions for the two smaller entropy classes also exhibit apparently significant differences in their means ($p < 0.01$). The number of observations in these classe is very small (<10 in the lowest entropy class) and the overall mean F-measure remains very high as well as the spread of the distribution remains small. As a consequence, although the means of these two classes are different, the data still suggests both high performance and high performance consistency, with high mean and narrow distribution. The width of the distribution in the [0.6,0.65] entropy class is probably affected by a number of relatively mediocre performance outliers, as suggested by the scatter plot of Figure 1. Nevertheless, its mean appears not to be significantly different from the mean of the [0.65,0.7] class.

Table 1. Two sample Welsch's t-test p-values for the mean of F-measure of metrical hierarchy evaluation. Values rejecting the null hypothesis of equal means at the 0.001 level are in bold font.

Entropy classes	p-value
$[0.5, 0.55]$ and $[0.55, 0.6]$	0.004
$[0.55, 0.6]$ and $[0.6, 0.65]$	0.005
$[0.6, 0.65]$ and $[0.65, 0.7]$	0.117
$[0.65, 0.7]$ and $[0.7, 0.75]$	0.076
$[0.7, 0.75]$ and $[0.75, 0.8]$	0.644
$[0.75, 0.8]$ and $[0.8, 0.85]$	$\ll \mathbf{0.001}$
$[0.8, 0.85]$ and $[0.85, 0.9]$	$\ll \mathbf{0.001}$
$[0.85, 0.9]$ and $[0.9, 0.95]$	$\ll \mathbf{0.001}$

In conclusion, it appears that for entropy values higher than 0.8 (approximately), firstly the mean performance significantly decreases and secondly the consistency of performance also decreases, as suggested by the widening of the performance scores distribution. In other words, the reliability of the feature extraction significantly drops for high entropy values, while it remains relatively stable on the lower range.

4.2. Tempo

The evaluation of tempo extraction provides a dichotomy between correct and incorrect estimations. The resulting data is grouped in entropy classes so that some statistical information can be derived. The percentage of successful tempo estimation for each entropy class is given in Figure 3. The apparent trend in this data suggests that the tempo extraction accuracy decreases as the rhythmic entropy increases. The Pearson, Spearman and Kendall coefficients computed for the middle of the entropy class and the tempo accuracy for each class are given in Table 2. The Spearman and Kendall coefficient strongly reveal the monotonic relationship between entropy and tempo estimation accuracy. Moreover, the Pearson coef-

Table 2. Correlation coefficients between entropy and both mean tempo accuracy for an entropy class and metrical structure F-measure, alongside with the corresponding p-value.

	Tempo		Metrical Structure	
	Coefficient	p-value	Coefficient	p-value
Pearson	-0.947	0.0001	-0.282	<0.0001
Spearman	-1.0	0.0	-0.247	<0.0001
Kendall	-1.0	0.0002	-0.179	<0.0001

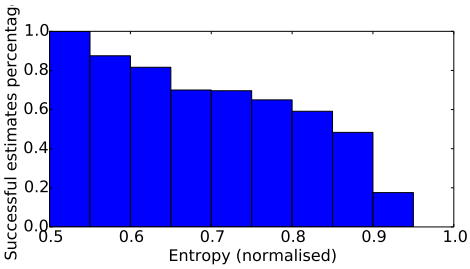

Fig. 3. Mean tempo extraction accuracy (proportion of correct estimations) for different entropy classes.

ficient suggests a good degree of negative linear correlation.

Tempo usually represents the rate of a metrical level, and an 'octave error' occurs when the algorithm produces a tempo estimate that is typically half or twice of the annotated tempo in the case of duple meter (a third or three times in the case of triple meter). As such, the 'octave error' estimate effectively corresponds to a different metrical level than the one which the annotated tempo is associated to. Therefore, incorporating tolerance to octave error in the evaluation procedure implicitly relates to the estimation of a part of the metrical structure. Interestingly, if the evaluation metric used is changed from the strict condition of equality between the estimated and annotated tempo, as used above, to a metric that also counts an 'octave error' (by ratios of either 1/3, 1/2, 2 or 3) as a correct tempo estimate, the distribution of percentage of 'correct' estimates exhibits a shape very similar to the distribution of average F-measure in the case of metrical structure extraction, as shown by the comparison of Figure 4 and Figure 2. Here again, the performance appears to be relatively stable from lowest entropy class up to a critical value (around 0.8) from which the performance drops.

4.3. Beat tracking

As a product of the assessment of beat tracking difficulty by beat trackers disagreement, Holzapfel *et al.* composed a dataset of 'hard' tracks for beat tracking [8]. 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 5. In addition, we performed

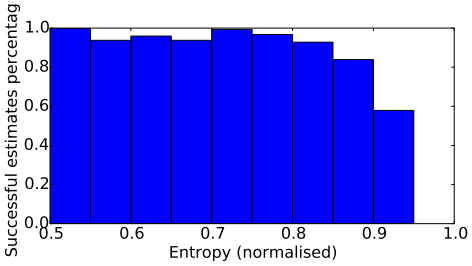

Fig. 4. Mean tempo extraction accuracy (proportion of correct estimations) for different entropy classes, also considering an octave error by a factor 1/3, 1/2, 2 or 3 as correct tempo estimate.

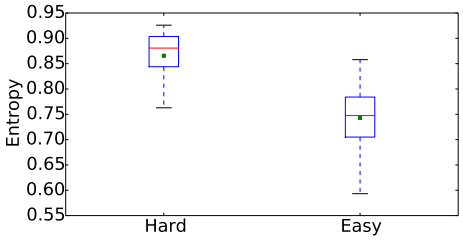

Fig. 5. Entropy distribution for the dataset published by Holzapfel *et al.* [8].

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, which means that 'easy' tracks tend to have a significantly smaller entropy than 'hard' tracks. This suggests that the entropy measurement is correlated with the results obtained by Holzapfel *et al.* [8]. 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.

5. CONCLUSIONS

In this paper we have investigated the relationship between the entropy of a rhythmogram derived from the audio and the reliability of the extraction of several high level rhythm features. We considered as reliable a feature extraction that performs consistently well. Providing a reliability or confidence value alongside an extracted feature significantly increases its usability in complex systems and real-world applications. The results show that the entropy 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. Given that the rhythmogram entropy is computed directly from the audio and does not depend on the feature extraction method, it is a valuable asset for the production of a reliability value, even for features for which a confidence value was not initially provided.

6. REFERENCES

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

[2] Daniel PW Ellis and Graham E. Poliner, "Identifying-cover songs' with chroma features and dynamic programming beat tracking," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*. 2007, vol. 4, pp. IV–1429, IEEE.

[3] Garth Griffin, Youngmoo E. Kim, and Douglas Turnbull, "Beat-sync-mash-coder: A web application for real-time creation of beat-synchronous music mashups," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. 2010, pp. 437–440, IEEE.

[4] George Tzanetakis, Georg Essl, and Perry Cook, "Human perception and computer extraction of musical beat strength," in *Proc. DAFx*, 2002, vol. 2.

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

[6] Peter Grosche, Meinard Müller, and Craig Stuart Sapp, "What Makes Beat Tracking Difficult? A Case Study on Chopin Mazurkas.," in *ISMIR*, 2010, pp. 649–654.

[7] Masataka Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[8] Andre Holzapfel, Matthew EP Davies, José R. Zapata, João Lobato Oliveira, and Fabien Gouyon, "Selective sampling for beat tracking evaluation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 9, pp. 2539–2548, 2012.

[9] Peter Grosche, M. Müller, and Frank Kurth, "Cyclic tempogram, a mid-level tempo representation for music signals," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. 2010, pp. 5522–5525, IEEE.

[10] Balaji Thoshkahna, Meinard Müller, 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.

[11] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B. Sandler, "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 1035–1047, 2005.

[12] Sebastian Böck and Gerhard Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. of the 16th Int. Conf. on Digital Audio Effects (DAFx). Maynooth, Ireland (Sept 2013)*, 2013.

[13] 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.

[14] Mi Tian, Gyorgy Fazekas, Dawn AA Black, and Mark Sandler, "On the use of the tempogram to describe audio content and its application to Music structural segmentation," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. 2015, pp. 419–423, IEEE.

[15] Matthew EP Davies and Mark D. Plumbley, "Context-dependent beat tracking of musical audio," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1009–1020, 2007.

[16] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *Speech and Audio Processing, IEEE transactions on*, vol. 10, no. 5, pp. 293–302, 2002.