MULTI-FEATURE MODELING OF PULSE CLARITY: DESIGN, VALIDATION AND OPTIMISATION

Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, Jose Fornari[*]

Finnish Centre of Excellence in Interdisciplinary Music Research,
University of Jyväskylä
Finland
<first.last>@campus.jyu.fi

ABSTRACT

Pulse clarity is considered as a high-level musical dimension that conveys how easily in a given musical piece, or a particular moment during that piece, listeners can perceive the underlying rhythmic or metrical pulsation. The objective of this study is to establish a composite model explaining pulse clarity judgments from the analysis of audio recordings, decomposed into a set of independent factors related to various musical dimensions. To evaluate the pulse clarity model, 25 participants have rated the pulse clarity of one hundred excerpts from movie soundtracks. The mapping between the model predictions and the ratings was carried out via regressions. More than three fourth of listeners' rating variance can be explained with a combination of periodicity-based and non-periodicity-based factors.

1. INTRODUCTION

This study is focused on one particular high-level dimension that may contribute to the subjective appreciation of music: namely *pulse clarity*, which conveys how easily listeners can perceive the underlying pulsation in music. The notion of pulse clarity is considered in this study as a subjective measure that participants were asked to rate whilst listening to a given set of musical excerpts. The aim is to model these behavioural responses using signal processing and statistical methods. Specific descriptors have been designed, that indicate diverse characteristics related to the amplitude envelope and its periodicities. The estimation of these primary representations, on the other hand, is based on a compilation of state-of-the-art research in this area. The objective of the experiment is to select the best combinations of primary representations and secondary descriptors correlating with listeners' judgements. This paper presents a subset of the model developed in [1].

2. COMPUTING THE ONSET DETECTION FUNCTION

In the analysis presented in this paper, several models for onset or beat detection and/or tempo estimation have been partially integrated into one single framework. Beats are considered as prominent energy-based onset locations, but more subtle onset positions (such as harmonic changes) might contribute to the global rhythmic organisation as well.

2.1. Amplitude envelope

When the onset detection curve is computed by way of envelope extraction, the audio signal is usually decomposed first into bands.

- This decomposition can be performed using a bank of filters ("filterbank" in figure 1), featuring between six [2], and more than twenty bands [3]. Filterbanks used in the models are Gammatone ("Gamm." in table 1) and two sets of non-overlapping filters : "Scheirer" [2] and "Klapuri" [3]. The envelope is extracted from each band through signal rectification and low-pass filtering. The low-pass filtering is implemented using either a simple auto-regressive filter ("IIR") or a convolution with a half-Hanning window ("half-Hanning") [2, 3].
- Another method consists in computing a spectrogram ("spectrum") and reassigning the frequency ranges into a limited number of critical bands [4]. The frame-by-frame succession of energy along each separate band, usually resampled to a higher rate, yields envelopes.

Important note onsets and rhythmical beats are characterised by significant rises of amplitude in the envelope. In order to emphasize those changes, the envelope is differentiated ("diff"). Differentiation of the logarithm ("log") of the envelope has also been advocated [3, 4]. The differentiated envelope can be subsequently half-wave rectified ("HWR") in order to focus only on the increase of energy. The half-wave rectified differentiated envelope can be summed to the non-differentiated envelope, with a specific λ weight fixed here to the value .8 proposed in [4] ("HWR=.8" in table 1).

2.2. Frequency-based strategy

Instead of focusing on the temporal evolution of the global energy, frequency-based onset detection curve describe temporal changes in the spectral distribution of the signal.

- One method consists in computing the spectral flux ("flux"), i.e., the distance between spectra computed on successive frames.
- Another method consists in computing distances not only between strictly successive frames, but also between all frames in a temporal neighbourhood of pre-specified width [5]. Inter-frame distances[1] are stored into a similarity matrix, and a "novelty" curve is computed by means of a convolution

[*] This work has been supported by the European Commission (Brain-Tuning FP6-2004-NEST-PATH-028570), the Academy of Finland (project 119959) and the Center for Advanced Study in the Behavioral Sciences, Stanford University. We are grateful to Tuukka Tervo for running the listening experiment.

[1] In our model, this novelty-based method is applied to frame-decomposed autocorrelation ("autocor").

along the main diagonal of the similarity matrix with a Gaussian checkerboard kernel [6]. Intuitively, the novelty curve indicates the positions of transitions throughout the temporal evolution of the spectral distribution.

2.3. Energy along frames

A simpler evaluation of the temporal evolution of energy consists in computing the root-mean-square energy ("rms") of each successive frame of the signal.

3. NON-PERIODIC CHARACTERISATIONS OF PULSE CLARITY

Some characterisations of the pulse clarity might be estimated from general characteristics of the onset detection curve that do not relate to periodicity.

3.1. Articulation

Articulation, describing musical performances in terms of *staccato* or *legato*, may have an influence in the appreciation of pulse clarity. One candidate description of articulation is based on Average Silence Ratio (ASR), indicating the percentage of frames that have an RMS energy significantly lower than the mean RMS energy of all frames [7]. The ASR is similar to the low-energy rate [8], except the use of a different energy threshold: the ASR is meant to characterize significantly silent frames. This articulation variable has been integrated in our model, corresponding to predictor "*ART*" in Figure 1.

3.2. Attack characterization

Characteristics related to the attack phase of the notes can be obtained from the amplitude envelope of the signal.

- Local maxima of the amplitude envelope can be considered as ending positions of the related attack phases. A complete determination of the attack requires therefore an estimation of its starting position, through an extraction of the preceding local minima using an appropriate smoothed version of the energy curve. The main slope of the attack phases [9] gives one candidate ("*ATT1*") for the prediction of pulse clarity.
- Alternatively, attack sharpness can be directly collected from the local maxima of the temporal derivative of the amplitude envelope ("*ATT2*") [4].

4. PERIODIC CHARACTERISATION OF PULSE CLARITY

Besides low-level characterization of dynamics developed in envelopes, pulse clarity seems to related more specifically to the degree of periodicity exhibited in these envelopes.

4.1. Pulsation estimation

The periodicity of the onset curve can be assessed via autocorrelation ("autocor") [10]. If the onset curve is decomposed into several channels, as is generally the case for amplitude envelopes, the autocorrelation can be computed either in each channel separately, and summed afterwards ("sum after"), or it can be computed from the summation of the onset curves ("sum bef."). A more refined method consists in summing adjacent channels into a lower number of wider band ("sum adj."), on each of which is computed the autocorrelation, further summed afterwards ("sum aft.") [4].

Peaks indicate the most probable periodicities. In order to model the perception of musical pulses, most perceptually salient periodicities are emphasized by multiplying the autocorrelation function with a resonance function ("reson."). Two resonance curve have been considered, one presented in [11] ("res1" in table 1), and a new curve developed for this study ("reson2"). In order to improve the results, redundant harmonic in the autocorrelation curve can be reduced by using an enhancement method ("enhan.") [12].

4.2. Previous work: Beat strength

One previous study on the dimension of pulse clarity [13] – where it is termed *beat strength* – is based on the computation of the autocorrelation function of the onset detection curve decomposed into frames. The three best periodicities are extracted. These periodicities – or more precisely, their related autocorrelation coefficients – are collected into a histogram. From the histogram, two estimation of beat strength are proposed: the `SUM` measure sums all the bins of the histogram, whereas the `PEAK` measure divides the maximum value to the main amplitude.

This approach is therefore aimed at understanding the global metrical aspect of an extensive musical piece. Our study, on the contrary, is focused here on an understanding of the short-term characteristics of rhythmical pulse. Indeed, even musical excerpt less than a few seconds long can easily convey to the listeners a strong sense of rhythmicity.

4.3. Statistical description of the autocorrelation curve

For that purpose, the analysis is focused on the analysis of the autocorrelation function itself, and tries to extract from it any information related to the dominance of the pulsation.

- The most evident descriptor is the amplitude of the main peak ("*MAX*"), i.e., the global maximum of the curve. The maximum at the origin of the autocorrelation curve is used as a reference in order to normalize the autocorrelation function. In this way, the actual values shown in the autocorrelation function correspond uniquely to periodic repetitions, and are not influenced by the global intensity of the total signal. The global maximum is extracted within a frequency range corresponding to perceptible rhythmic periodicities, i.e. for the range of tempi between 40 and 200 BPM.
- The global minimum ("*MIN*") gives another aspect of the importance of the main pulsation. The motivation for including this measure lies in the fact that for periodic stimuli with a mean of zero the autocorrelation function shows minima with negative values, whereas for non-periodic stimuli this does not hold true.
- Another way of describing the clarity of a rhythmic pulsation consists in assessing whether the main pulsation is related to a very precise and stable periodicity, or if on the contrary the pulsation slightly oscillates around a range of possible periodicities. We propose to evaluate this characteristic through a direct observation of the autocorrelation function. In the first case, if the periodicity remains clear and stable, the autocorrelation function should display

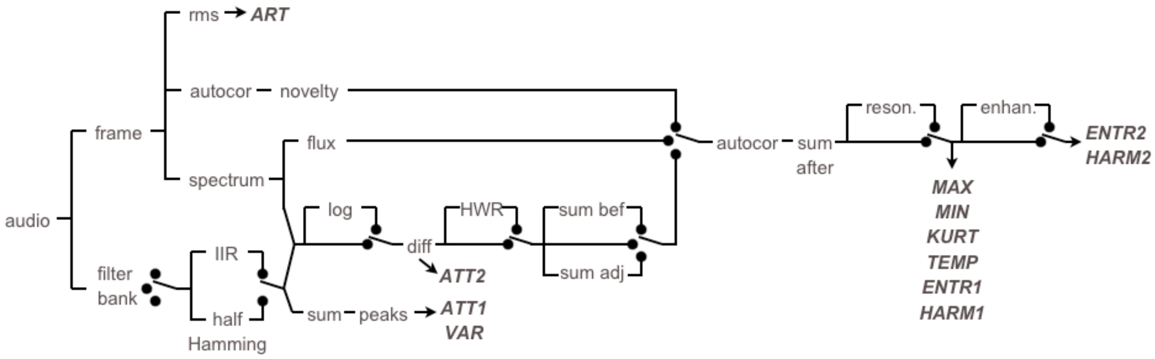

Figure 1: *Flowchart of operators of the compound pulse clarity model, where options are indicated by switches.*

a clear peak at the corresponding periodicity, with significantly sharp slopes. In the second and opposite case, if the periodicity fluctuates, the peak should present far less sharpness and the slopes should be more gradual. This characteristic can be estimated by computing the kurtosis of the lobe of the autocorrelation function containing the major peak. The kurtosis, or more precisely the excess kurtosis of the main peak ("*KURT*"), returns a value close to zero if the peak resembles a Gaussian. Higher values of excess kurtosis correspond to higher sharpness of the peak.

- The entropy of the autocorrelation function ("*ENTR1*" and "*ENTR2*") characterizes the simplicity of the function and provides in particular a measure of the peakiness of the function. This measure can be used to discriminate periodic and non-periodic signals. In particular, signals exhibiting periodic behaviour tend to have autocorrelation functions with clearer peaks and thus lower entropy than non-periodic ones.
- Another hypothesis is that the faster a tempo ("*TEMP*") is, the more clearly it is perceived by the listeners. This conjecture is based on the fact that fast tempi imply a higher density of beats, guiding the rhythmic understanding of the listeners more tightly.

4.4. Harmonic relations between pulsations

The clarity of a pulse seems to decrease if pulsations with no harmonic relations coexist. We propose to formalize this idea as follows. First a certain number of peaks are selected from the autocorrelation curve. Let the list of peak lags be $P = \{l_i\}_{i\in[0,N]}$, and let the first peak l_0 be the one considered as the main pulsation. The list of peak amplitudes is $\{p(l_i)\}_{i\in[0,N]}$.

A peak will be inharmonic if the remainder of the euclidian division of its lag with the lag of the main peak (and the inverted division as well) is significantly high. This defines the set of inharmonic peaks $\overline{H}$:

$$\overline{H} = \left\{ i \in [0,N] \;\middle|\; \begin{array}{ll} l_i \in [\alpha l_0, (1-\alpha)l_0] & \pmod{l_0} \\ l_0 \in [\alpha l_i, (1-\alpha)l_i] & \pmod{l_i} \end{array} \right\} \qquad (1)$$

where α is a constant tuned to .15 in our implementation.

The degree of harmonicity is hence decreased by the cumulation of the autocorrelation coefficients of the non-harmonic peaks:

$$\texttt{HARM} = \exp\left(-\frac{1}{r}\frac{\sum_{i\in\overline{H}} p(l_i)}{p(l_0)}\right) \qquad (2)$$

where r is another constant set to 4.

5. MAPPING MODEL PREDICTIONS TO LISTENERS' RATINGS

In order to assess the validity of the models predicting pulse clarity judgments presented in the previous section, an experimental protocol has been designed. 25 musically trained participants rated the clarity of the pulse on one hundred 5-second excerpts using a computer interface that randomized the excerpt orders individually [14]. These ratings were considerably homogenous (Cronbach alpha of 0.971) and therefore the mean ratings will be utilized in the following analyses.

The major factors correlating with the ratings are indicated in table 1. The best predictor is the global minimum of the autocorrelation function, with a correlation r of .59 with the ratings. For the following variables, κ indicates the highest cross-correlation with any factor of better r value. A low κ value indicates a good independence of the related factor, with respect to the other factors considered as better predictors.

Table 1: *Majors factors correlating with pulse clarity ratings, in decreasing order of correlation r with the ratings. Factor with cross-correlation κ exceeding .6 have been removed.*

var	r	κ	parameters
MIN	.59		Klapuri, half Hamming, log, HWR, sum bef., reson1
KURT	.42	.55	Scheirer, IIR, sum aft.
HARM1	.40	.53	Scheirer, IIR, log, HWR, sum aft.
ENTR2	-.4	.54	Klapuri, IIR, log, HWR=.8, sum bef., reson2
MIN	.40	.58	Flux, reson1

Table 2 shows the result of the stepwise regression between the ratings and all computed variables. A compound model of pulse clarity can be obtained through a linear combination of six of the best factors, explaining up to 76 % of the variability of listeners' ratings.

Table 2: *Result of stepwise regression between pulse clarity ratings and all computed variables, with accumulated adjusted variance r^2 and standardized β coefficients.*

step	var	r^2	β	parameters
1	*MIN*	.48	1.57	Scheirer, half Hamming, sum bef.
2	*HARM2*	.68	-.81	Spectrum, log, sum adj.
3	*TEMP*	.76	.64	Gamm., half Hamming, HWR=.8, sum aft., res2

6. *MIRTOOLBOX* 1.2

The whole set of algorithms used in this experiment has been implemented using *MIRtoolbox*[2][15]: the set of operators available in the version 1.1 of the toolbox have been improved in order to incorporate a part of the onset extraction and tempo estimation approaches presented in this paper. The different paths indicated in the flowchart in figure 1 can be implemented in *MIRtoolbox* in alternative ways:

- The successive operations forming a given process can be called one after the other, and options related to each operator can be specified as arguments. For example,

```
    a = miraudio('myfile.wav')
f = mirfilterbank(a,'Scheirer')
  e = mirenvelope(f,'HalfHann')
                            etc.
```

- The whole process can be executed in one single command. For example, the estimation of pulse clarity based on the MIN heuristics computed using the implementation in [3] can be called this way:

```
mirpulseclarity('myfile.wav',
           'Min','Klapuri99')
```

- A linear combination of best predictors, based on the stepwise regression[3] can be used as well. The number of factors to integrate in the model can be specified.
- Multiple paths of the pulse clarity general flowchart can be traversed simultaneously. At the extreme, the complete flowchart, with all the possible alternative switches, can be computed as well. Due to the complexity of such computation[4], optimization mechanisms limit redundant computations.

The routine performing the statistical mapping – between the listeners' ratings and the set of variables computed for the same set of audio recordings – is also integrated in the new version of *MIRtoolbox*. This routines includes an optimization algorithm that automatically finds optimal Box-Cox transformations [16] of the data ensuring that their distributions becomes sufficiently gaussian, which is a prerequisite for correlation estimation.

[2] Available at http://www.jyu.fi/music/coe/materials/mirtoolbox

[3] The final multi-feature model available in the latest version 1.2 of *MIRtoolbox* actually results from more advanced results of this study [1].

[4] In the complete flowchart, the number of individual variables exceeds 6000.

7. REFERENCES

[1] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "(paper in preparation)," in *Proc. Intl. Conf. on Music Information Retrieval*, Philadelphia, PA, Sep. 14-18 2008.

[2] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 588–601, 1998.

[3] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proc. Intl. Conf. on Acoust. Speech Sig. Proc.*, Phoenix, Arizona, Mar. 15-19 1999, pp. 3089–3092.

[4] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio Speech Langage Proc.*, vol. 14, no. 1, pp. 342–355, 2006.

[5] J. P. Bello, S. Abdallah L. Daudet, C. Duxbury, M. Davies, and M. Sandler, "A tutorial on onset detection in music signals," *Tr. Speech Audio Proc.*, vol. 13, no. 5, pp. 1035–1047, 2005.

[6] J. Foote and M. Cooper, "Media segmentation using self-similarity decomposition," in *Proceedings of SPIE Storage and Retrieval for Multimedia Databases*, 2003, number 5021, pp. 167–175.

[7] Y. Feng, Y. Zhuang, and Y. Pan, "Popular music retrieval by detecting mood," in *Proc. Intl. ACM SIGIR Conf. on Res. Dev. Information Retrieval*, Toronto, Canada, Jul. 28-Aug. 1 2003, pp. 375–376.

[8] J. J. Burred and A. Lerch, "A hierarchical approach to automatic musical genre classification," in *Proc. Digital Audio Effects (DAFx-03)*, London, UK, Sep. 8-11 2003, pp. 344–349.

[9] G. Peeters, "A large set of audio features for sound description (similarity and classification) in the cuidado project (version 1.0)," Tech. Rep., Ircam, 2004.

[10] J. C. Brown, "Determination of the meter of musical scores by autocorrelation," *J. Acoust. Soc. Am.*, vol. 94, no. 4, pp. 1953–1957, 1993.

[11] P. Toiviainen and J. S. Snyder, "Tapping to bach: Resonance-based modeling of pulse," *Music Perception*, vol. 21, no. 1, pp. 43–80, 2003.

[12] T. Tolonen and M. Karjalainen, "A computationally efficient multipitch analysis model," *IEEE Trans. Speech Audio Proc.*, vol. 8, no. 6, pp. 708–716, 2000.

[13] G. Tzanetakis, G. Essl, and P. Cook, "Human perception and computer extraction of musical beat strength," in *Proc. Digital Audio Effects (DAFx-02)*, Hamburg, Germany, Sep. 26-28, 2002, pp. 257–61.

[14] O. Lartillot, T. Eerola, P. Toiviainen, and J. Fornari, "Multi-feature modeling of pulse clarity from audio," in *Proc. Intl. Conf. on Music Perception and Cognition*, Sapporo, Japan, Aug. 25-29 2008.

[15] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proc. Digital Audio Effects (DAFx-07)*, Bordeaux, France, Sep. 10-15 2007, pp. 237–244.

[16] G. E. P. Box and D. R. Cox, "An analysis of transformations," *J. Roy. Stat. Soc.*, , no. 26, pp. 211–246, 1964.