INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Music Tempo Estimation Using Sub-band Synchrony

Shreyan Chowdhury, Tanaya Guha, and Rajesh M. Hegde

Indian Institute of Technology, Kanpur

shreyan0311@gmail.com, tanaya@iitk.ac.in, rhegde@iitk.ac.in

Abstract

Tempo estimation aims at estimating the pace of a musical piece measured in beats per minute. This paper presents a new tempo estimation method that utilizes coherent energy changes across multiple frequency sub-bands to identify the onsets. A new measure, called the sub-band synchrony, is proposed to detect and quantify the coherent amplitude changes across multiple sub-bands. Given a musical piece, our method first detects the onsets using the sub-band synchrony measure. The periodicity of the resulting onset curve, measured using the autocorrelation function, is used to estimate the tempo value. The performance of the sub-band synchrony based tempo estimation method is evaluated on two music databases. Experimental results indicate a reasonable improvement in performance when compared to conventional methods of tempo estimation.

Index Terms: Music Tempo Estimation, Sub-band Synchrony

1. Introduction

The experience of listening to music is an engagement of our mind with the elements of music that broadly comprise melody, harmony, rhythm and timbre [1]. Rhythm, in its most generic sense, refers to all temporal aspects of a musical work, including periodicity, pace, syncopation, and the perception of repetition of a musical piece with time. A perceptually fundamental aspect of rhythm is the pace or *tempo*, the most salient component of which is the beat or *tactus* [2, 3]. The tapping behavior of humans (where one taps at the tactus rate) is considered a reliable indicator of tempo perception [4, 5, 6]. The tempo of a musical piece is defined as the rate of the tactus pulse, typically expressed in ‘beats per minute’ (BPM) [7]. This is considered as a quantified measure of musical speed [8, 9]. Tempo estimation has applications in music production and mixing, music classification [10, 11], and audiovisual synchronization [12].

One popular approach to music tempo estimation involves the detection of discrete *onsets*, and using the inter-onset interval to measure the tempo [13, 14]. Musical onsets refer to the discrete events that indicate the beginning of the notes or percussive events [15]. Onsets in music often act as the *phenomenal accents*, which are events that emphasize a moment in music, and are important in the role of meter perception [16]. An onset detection method based on acoustic energy flux was proposed by Laroche [17]. Fitzgerald used median filtering to separate the percussive events from the non-percussive ones to detect the percussive onsets [18]. In a comprehensive study, Bello et al. [15] provided a bottom-up description of the basic approach to onset detection using amplitude envelope. In another study, Dixon [19] discussed the pragmatic methods of onset detection including phase deviation and complex domain based methods, followed by an exhaustive evaluation. Dan Ellis used the Mel spectrogram to detect the onsets and to estimate tempo [20], while Dixon [14] proposed an energy based onset detector. In another work, Dixon also implemented periodicity computations in band-limited signals derived from the main audio signal to compute onsets and metrical structure [21]. Multiple frequency bands and comb resonators were used for tempo estimation by Klapuri [16]. Filterbank approaches have also been investigated in the past [22, 23]. Recent methods have focused on using neural networks to determine the beat onset curve and comb filtering to capture the periodicity [24]. Periodicity capture has also been done using autocorrelation methods [25].

This paper proposes a tempo estimation method that relies on the coherent changes across multiple frequency sub-bands to identify the onsets. We detect and quantify the coherent amplitude changes across multiple sub-bands, and derive a measure called the *sub-band synchrony*. The idea of multiband processing for tempo estimation is not new. It has been previously used for detection of periodicity [23] and metrical structure [21]. Our method differs from these works in the way we decompose the music signal (in a larger auditory band), and in proposing a new technique for identifying the points of coherent changes across the spectrum. This allows us to track harmonic as well as the percussive changes, leading to a more accurate onset curve, which is then used for tempo estimation. We evaluate our proposed tempo estimation method on two databases (one publicly available database [26] and the other created by the authors) comprising of music from different genres. Our method demonstrates competitive results as compared to the existing methods, in addition to being simplified.

2. Sub-band Synchrony based Tempo Estimation

In this section, we describe the proposed *sub-band synchrony* tempo estimation method. Our method looks for coherent changes across multiple sub-bands of a music signal to identify onsets. The periodicity of the resulting onset curve is measured using the generalized autocorrelation function, and the tempo is inferred from the periodicity thereafter. The steps of our proposed method are described below in detail.

2.1. Sub-band Decomposition

Let us consider a given digital music signal $s[n], n = 1, 2, ..., N$, where N is the total number of samples in the signal, and $s[n]$ is the amplitude of the signal at sample n. A gammatone filter bank, denoted as $g_k[n]$ [27], is used to decompose $s[n]$ into K sub-bands. Gammatone filters are chosen as they are widely used as approximations of auditory filters in the human auditory system [28].

$$s_k[n] = g_k[n] * s[n], \qquad k = 1, 2, ..., K \tag{1}$$

where $*$ indicates convolution, and $s_k[n]$ is the output of the filter bank. The impulse response of a gammatone filter is taken to be (from [27])

$$g_k[n] = an^{p-1}e^{-2\pi bn}\cos(2\pi f_k n + \phi) \tag{2}$$

http://dx.doi.org/10.21437/Interspeech.2017-1000

where p is the filter order, b is a bandwidth parameter, f_k is the k-th filter center frequency and ϕ is the phase. In our experiments, splitting into around $K = 40$ sub-bands is observed to yield most accurate results, though the exact number is not critical.

Next, we compute the envelopes for sub-band signals. Each signal is squared so as to demodulate the input signal by using the input itself as its carrier wave. This means that half the energy of the signal is pushed up to higher frequencies and half is shifted down toward DC.

$$\tilde{s}_k[n] = 2 \times [s_k[n]]^2 \tag{3}$$

To maintain the correct amplitude scaling, the signal is amplified by a factor of two, and its square root is taken to reverse the scaling distortion that resulted from squaring the signal. The signal is then downsampled by a factor of 4 to reduce the sampling frequency. To prevent aliasing, an finite impulse response (FIR) decimation is used, which applies a low pass filter before downsampling the signal.

$$\tilde{s}_{k,\text{dec}}[n] = \text{dec}(\tilde{s}_k[n], 4) \tag{4}$$

where $\text{dec}(x, n)$ denotes the decimation function decimating x by a factor of n.

The next step is to use an envelope detection function to each of the sub-bands. We compute the envelope by passing $s_k[n]$ through a low pass filter as follows.

$$E_k[n] = \sqrt{h[n] * \tilde{s}_{k,\text{dec}}[n]} \tag{5}$$

where $E_k[n]$ denotes the final sub-band envelope for band k, and $h[n]$ is the impulse response of a two-degree low pass filter used as a smoothing function. The resulting envelopes are used to detect the onsets. Fig. 1(a) shows a sample music signal and Fig. 1(b) shows its 40 corresponding envelopes.

2.2. Onset Detection using Sub-band Synchrony

In order to detect the changes in a spectrum that correspond to onsets, it is necessary to compute a measure of temporal changes of the envelopes. To achieve this, we take the derivative of all the envelopes.

$$D_k[n] = \frac{\mathrm{d}}{\mathrm{d}n} E_k[n] \tag{6}$$

where $D_k[n]$ is the derivative of the envelope $E_k[n]$.

At every onset event it is expected that the sub-bands will reflect a coherent disturbance resulting in a coherent change in its local energy. This coherent change in the sub-band energy is called sub-band synchrony. This phenomenon is also evident from Fig. 1(c), where the derivatives of the sub-band envelopes are close to zero across all sub-bands at the non-onset positions, and those at the onset positions exhibit greater magnitude and variability. To quantify sub-band synchrony, we look at statistical properties across the frequency bands over time. Two statistical estimates are computed from $D_k[n]$ for onset detection.

Sub-band synchrony mean (SBS$_\mu$): From Fig. 1, we observe that the points of onset give rise to higher derivative values across sub-band envelopes. Thus, the mean (Eq. 7) of the derivative values across sub-band envelopes is expected to be higher at the onset points.

$$\mu[n] = \frac{1}{M} \sum_{k=1}^{M} D_k[n] \tag{7}$$

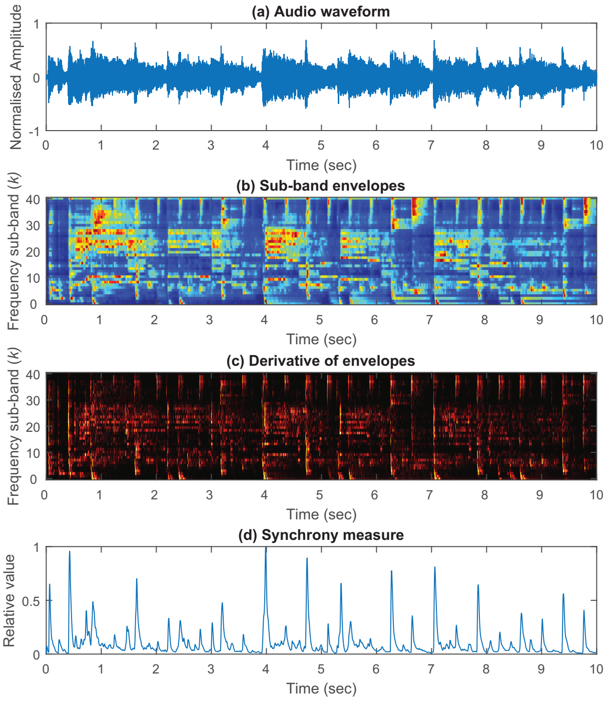

Figure 1: *The proposed onset detection process for a music sample: (a) the music signal, (b) 40 sub-band envelopes, (c) derivative of each subband, (d) onset envelope obtained after taking mean of the derivative envelopes at each time frame*

where M is the number of sub-bands. The mean computed at each time frame yields the onset curve (see Fig. 1(d)).

Sub-band synchrony variance (SBS$_\sigma$): Statistical variances of $D_k[n]$ across frequency bands are computed at each time frame, as given in Eq. 8. We expect variances at onset event instances to be higher than non-onset event instances because the magnitude of $D_k[n]$ for k's in which there are prominent changes are much more than for those in which there are relatively less prominent changes. For example, at the onset of a played middle-C piano note, many mid-range bands will show a coherent change, but the magnitude in bands near the extreme high and low frequencies will remain close to zero. At non-onset time frames, *all* band magnitudes will remain close to

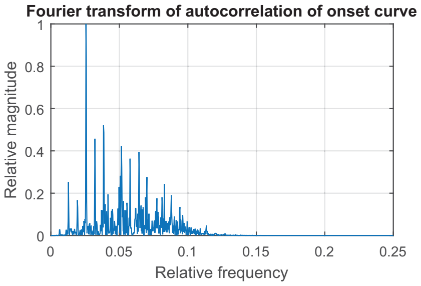

Figure 2: *Fourier transform of autocorrelation of onset curve*

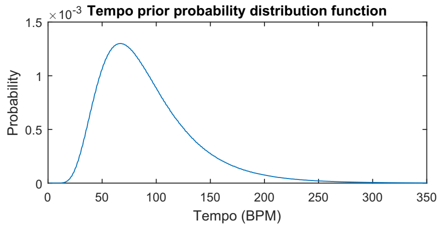

Figure 3: *Prior probability distribution function for tempo values*

zero. The variances is computed as follows:

$$\sigma[n] = \frac{1}{M}\sum_{k=1}^{M}(D_k[n] - \mu[n])^2 \tag{8}$$

where M is the number of sub-bands, and $\mu[n]$ is the mean across bands (refer to equation 7). Finally the onset curve $\dot{S}[n]$ is obtained as follows.

$$\dot{S}[n] = \begin{cases} \mu[n], & \text{for SBS}_\mu \\ \sigma[n], & \text{for SBS}_\sigma \end{cases} \qquad n = 1, 2, ..., N \tag{9}$$

2.3. Tempo Estimation

The tempo of the given music signal is computed from the periodicity of the onset curve $\dot{S}[n]$. The autocorrelation approach is utilized to calculate the periodicity. Autocorrelation is defined as the cross-correlation of a signal with itself at a certain lag. The periodicity curve is obtained by computing the autocorrelation for all lags from $-N$ to N. We thus have the periodicity curve $R[n]$ calculated as:

$$R[l] = \sum_{n=1}^{N}\dot{S}[n]\dot{S}[n-l] \qquad l = -N, -(N-1), ..., N \tag{10}$$

where l is the lag. Next, the fast Fourier transform (FFT) of $R[l]$ is computed, as shown in Fig. 2, to find the frequency distribution of its oscillations. The peaks denote the dominant frequencies in the Fourier spectrum of the periodicity function $R[l]$. The Fourier spectrum so obtained is then scaled by a prior tempo probability distribution function to eliminate unrealistic tempo estimates. This prior tempo probability distribution function is derived from a model of tactus periods of actual songs measured by many authors [2]. The model for tactus periods is defined as

$$P(\tau) = \frac{1}{K}\exp\left\{-\frac{1}{2\sigma^2}\left[\log_{10}\left(\frac{\tau}{\mu}\right)\right]^2\right\} \tag{11}$$

where $P(\tau)$ is the probability of the tactus period of an actual song being τ; μ denotes *moderate pulse period* and is typically around 600 ms; σ is the standard deviation of the logarithm of the pulse period and typically has a value of about 0.2. $\frac{1}{K}$ is the normalization constant. From this, for each tactus period τ (in seconds), we can read off the probability and assign that to the tempo corresponding to tactus period τ. The probability distribution function of tempos can thus be derived, and is shown in Fig. 3.

After scaling, the frequency at which the peak is obtained is finally converted to the tempo value (in BPM) as follows.

$$\text{tempo}_{\text{SBS}} = f_p \times \frac{f_{max}}{N_{\text{FFT}}} \times 60 \tag{12}$$

Table 1: *Summary of the two databases used in our work*

	ISMIR2004 [26]	IITK-MT
Number of songs	465	230
Duration per song	20 sec	10 or 30 sec
Max tempo (in BPM)	242	334
Min tempo (in BPM)	24	54

Table 2: *Results on the ISMIR2004 database*

Metric	MFCC [20]	FB [29]	SBS$_\mu$	SBS$_\sigma$
ϵ (in %)	**43.26**	29.68	23.66	21.08
ϵ_{scaled} (in %)	75.70	60.65	**77.42**	71.18
RMSE	10.88	14.54	**8.76**	9.30

where f_p is the fast Fourier transform (FFT) point at which the peak occurs, f_{max} is the Nyquist frequency of the audio (i.e. for an audio with sampling rate 44.1 kHz, $f_{max} = 22.05$ kHz), and N_{FFT} is the number of points in the FFT.

3. Performance Evaluation

The performance of the proposed tempo estimation method is evaluated on two datasets namely *ISMIR2004 database* [26] and the IITK-MT database. Table 1 contains a summary of the datasets.

ISMIR2004 database [26]: This database consists of 465 songs (duration 20 secs each) with approximately constant tempos.

IITK-MT database: This database is created by the authors using 230 song excerpts (duration 10 or 30 seconds each) from four different genres (Western classical, Indian classical, popular music, and rock music). This also includes an annotated dataset by Dan Ellis [20].

3.1. Evaluation Metrics

Three metrics are used in performance evaluation. The first is denoted by ϵ, which computes the percentage of estimated tempos falling in a 4% band around the ground-truth tempos [26]. The second metric denoted as ϵ_{scaled}, takes into account the octave tempo deviations, and computes the percentage of estimated tempos falling within 4% of 1, 2, 1/2, 3, 1/3 times the ground truth tempo. The third metric is the root mean squared error (RMSE) of the estimated tempo. Taking into account octave deviations this error is computed between the estimated tempo and the nearest of 1, 2, 1/2, 3, 1/3 of the ground truth values.

Using the above metrics, the sub-band synchrony tempo estimation method is compared with two other methods namely Mel Frequency Cepstral Coefficients (MFCC)-based [20] and filter bank (FB)-based [29] tempo estimation. Results are summarized in Tables 2 and 3. The (relative) error distribution for the proposed and the compared methods are presented in Fig. 4.

Relative errors are calculated as $((\text{tempo}_{\text{estimated}} - \text{tempo}_{\text{actual}})/\text{tempo}_{\text{actual}})$. The results in Tables 2 and 3 show

Table 3: *Results on the IITK-MT database*

Metric	MFCC [20]	FB [29]	SBS$_\mu$	SBS$_\sigma$
ϵ (in %)	58.26	51.74	**65.22**	58.70
ϵ_{scaled} (in %)	84.35	63.04	**85.65**	79.13
RMSE	15.13	22.74	**11.78**	14.44

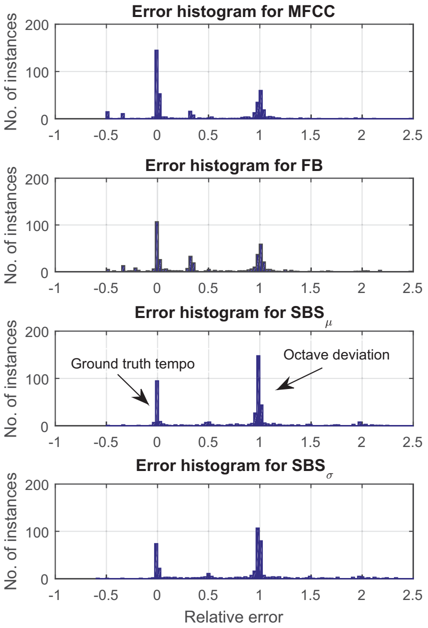

Figure 4: *Error histograms for the three evaluated methods. Errors greater than 2.5 have been truncated in these charts to save space.*

that the proposed sub-band synchrony-based tempo estimation method outperforms the other two methods when octave (double) tempo deviations are taken into account. The metric ϵ_{scaled} takes into account octave deviations since the metrical level is subjective and not always clearly agreed upon by human listeners [26].

4. Conclusion

A sub-band synchrony based tempo estimation method is presented in this work. A measure of sub-band synchrony which detects onsets by locating coherent changes across different frequency sub-bands is developed. This method is able to track harmonic as well as percussive changes leading to more accurate onset detection, and subsequently better tempo estimation. Experimental results indicate that this method performs reasonably better than two existing methods. Future work will investigate real-time tempo estimation using higher order statistical measures obtained from sub-band synchrony.

5. References

[1] L. B. Meyer, *Explaining music: Essays and explorations*. Univ of California Press, 1973.

[2] R. Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music perception*, pp. 409–464, 1994.

[3] M. R. Jones and M. Boltz, "Dynamic attending and responses to time." *Psychological review*, vol. 96, no. 3, p. 459, 1989.

[4] J. Snyder and C. L. Krumhansl, "Tapping to ragtime: Cues to pulse finding," *Music Perception*, vol. 18, no. 4, pp. 455–489, 2001.

[5] D. Epstein, *Shaping time: Music, the brain, and performance*. Wadsworth Publishing Company, 1995.

[6] C. Drake, L. Gros, and A. Penel, "How fast is that music? the relation between physical and perceived tempo," in *Int. Conf. Music Percept. Cognit*, 1999.

[7] A. J. Eronen and A. P. Klapuri, "Music tempo estimation with k-nn regression," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 18, no. 1, pp. 50–57, Jan 2010.

[8] S. Quinn and R. Watt, "The perception of tempo in music," *Perception*, vol. 35, no. 2, pp. 267–280, 2006.

[9] D. Moelants, "Perception and performance of aksak metres," *Musicae Scientiae*, vol. 10, no. 2, pp. 147–172, 2006.

[10] C. Xu, N. C. Maddage, and X. Shao, "Automatic music classification and summarization," *IEEE Trans. Speech and Audio Proc.*, vol. 13, no. 3, pp. 441–450, 2005.

[11] Z. Fu, G. Lu, K. M. Ting, and D. Zhang, "A survey of audio-based music classification and annotation," *IEEE Trans. Multimedia*, vol. 13, no. 2, pp. 303–319, 2011.

[12] M. Goto and Y. Muraoka, "An audio-based real-time beat tracking system and its applications," in *Int. Computer Music Conf.*, 1998, pp. 17–20.

[13] M. E. Davies, P. M. Brossier, and M. D. Plumbley, "Beat tracking towards automatic musical accompaniment," in *Audio Engineering Society Convention 118*, 2005.

[14] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *J. of New Music Research*, vol. 30, no. 1, pp. 39–58, 2001.

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

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

[17] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *J. of the Audio Engineering Society*, vol. 51, no. 4, pp. 226–233, 2003.

[18] D. Fitzgerald, "Harmonic/percussive separation using median filtering," 2010.

[19] S. Dixon, "Onset detection revisited," in *Int. Conf. on Digital Audio Effects (DAFx-06)*, 2006, pp. 133–137.

[20] D. P. Ellis, "Beat tracking by dynamic programming," *J. of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.

[21] S. Dixon, E. Pampalk, and G. Widmer, "Classification of dance music by periodicity patterns." in *ISMIR*, 2003, pp. 159–165.

[22] O. Lartillot, "Mirtoolbox 1.3. 4 users manual," *Finnish Centre of Excellence in Interdisciplinary Music Research, University of Jyväskylä, Finland*, 2011.

[23] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *J. of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.

[24] S. Böck, F. Krebs, and G. Widmer, "Accurate tempo estimation based on recurrent neural networks and resonating comb filters." in *ISMIR*, 2015, pp. 625–631.

[25] G. Percival and G. Tzanetakis, "Streamlined tempo estimation based on autocorrelation and cross-correlation with pulses," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1765–1776, 2014.

[26] J. Zapata and E. Gómez, "Comparative evaluation and combination of audio tempo estimation approaches," in *Audio Engineering Society Conf. Semantic Audio*. Audio Engineering Society, 2011.

[27] R. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Meeting of the IOC Speech Group on Auditory Modeling at RSRE*, vol. 2, no. 7, 1987.

[28] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," *Auditory physiology and perception*, vol. 83, pp. 429–446, 1992.

[29] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Int. Conf. Digital Audio Effects*, 2007, pp. 237–244.