

On Musical Onset Detection via the S-Transform

Nishal Silva

Dept. of Eng. and Mathematics
Sheffield Hallam University
Sheffield, UK
b3047941@my.shu.ac.uk

Chathuranga Weeraddana

Dept. of Electronic and Telecomm. Eng.
University of Moratuwa
Moratuwa, Sri Lanka
chathurangaw@uom.lk

Carlo Fiscione

Dept. of Networks and Systems Eng.
KTH Royal Institute of Technology
Stockholm, Sweden
carlofi@kth.se

Abstract—Musical onset detection is a key component in any beat tracking system. Existing onset detection methods are based on temporal/spectral analysis, or methods that integrate temporal and spectral information together with statistical estimation and machine learning models. In this paper, we propose a method to localize onset components in music by using the S-transform, and thus, the method is purely based on temporal/spectral data. Unlike the other methods based on temporal/spectral data, which usually rely short time Fourier transform (STFT), our method enables effective isolation of crucial frequency subbands due to the frequency dependent resolution of S-transform. Moreover, numerical results show, even with less computationally intensive steps, the proposed method can closely resemble the performance of more resource intensive statistical estimation based approaches.

Index Terms—Onset detection, beat tracking, music, S-transform, time frequency representation

I. INTRODUCTION

When a human hears music, an action which is almost subconscious is the rhythmic tapping of the foot. These taps are consistent with the *beat* of the music and is measured in beats-per-minute (bpm). The process of detecting beat locations in a music is called *beat tracking*. Beat tracking is a vital step in many studio and live music applications: for example, when a DJ should perform beat matching to play two songs successively. Beat matching is the adjustment of the tempo of one or multiple songs so that their beat locations overlap each other when played simultaneously. The same applies for an audio engineer whenever two instrument tracks are to be played in unison. In this case the audio engineer needs to know the beat locations in both tracks to create a smooth playthrough.

Beat of a music is maintained by a rhythm instrument. A beat usually corresponds to a rapid and unpredictable change in the underlying music signal. Therefore, a primary step in any beat tracking algorithm is to represent such changes, which is referred to as the *beat causing onsets* (BCO). However, isolating BCOs among others can be challenging.

Existing onset isolation (detection) algorithms, based on temporal and spectral analysis, do not usually yield good results when the beat of a music is not prominent. A primary cause of this is the masking off of important BCO components. Therefore, exploring generalized mechanisms for BCO detection in music, is important in theory, as well as in practice, and therefore deserve investigation. Blending the existing temporal

and spectral analysis methods with statistical estimation techniques yields more promising results, however, at the expense of significant computational complexity. Integrating temporal and spectral data of music with machine learning techniques (e.g., neural networks) is apparently the best among others. Such algorithms always rely on a substantial training phase in advance, in order to yield promising results.

In this paper, we propose a method which relies on the S-transform [1] for BCO detection. Unlike the existing methods based on statistical estimation techniques, our method does not rely on any *a priori* information of the underlying music. Moreover, unlike the state-of-the art machine learning algorithms, the proposed method does not require a training dataset. The proposed algorithm can be considered as a graceful trade-off between the performance and the computational complexity and resources required.

The choice of the S-transform, among other time-frequency representations (TFR), is motivated by the following:

- 1) The beat causing onsets are usually created by instruments with relatively lower frequencies [2],
- 2) S-transform provides a good concentration at lower frequencies [1].
- 3) S-Transform uses a frequency-dependent window dilation, which results a frequency-dependent resolution [1].

The first two points enable one to extract the power of rhythm instruments effectively. The last point plays a key role in the sense that, unlike the STFT, S-transform is not required to know the window size a priori. This facilitates, irrespective of the underlying frequencies of the rhythm instruments, a general implementation of proposed algorithms.

The rest of the paper is organized as follows. In Section II, we give a literature overview. Section III discusses our proposed algorithms for BCO detection. In Section VII, numerical results are presented. Section VIII concludes the paper.

II. LITERATURE

Several works have been investigated on BCO detection in music [3]–[23]. These can be split into methods based on temporal/spectral analysis, and more sophisticated methods which blend temporal/spectral data together with statistical estimation techniques and machine learning techniques. Temporal and spectral analysis methods generate a *time series*, usually called the *onset envelope function* (OEF), which contains information of the locations of BCOs. The OEF is then

used to compute the underlying bpm [5]. Methods based on statistical estimation and machine learning techniques rely on more resources, in addition to the pure temporal and spectral data, for locating BCOs, e.g., *a priori* information of the underlying music, training data sets [16, § 4].

Temporal analysis methods split the signal into frequency bands, for which amplitude envelopes are calculated and summed to obtain an OEF [3], [4]. Spectral analysis methods take into account, the change in spectral energy. These methods usually compute some form of a time-frequency representation (TFR), where the STFT is most common [3], [5], [10], [21], [22]. Different scaling methods such as *the Mel* [5], [6] and *the square root* [7] are used to avoid low amplitude components from being masked off. Either the summation [5], [7], the median [6], or the mean [8] is computed of the first order difference for each time bin to obtain an *OEF*.

A common limitation of the spectral analysis methods mentioned above is the poor detection of BCOs if the rhythm is less pronounced. This is due to masking off of BCO components of interest, or because spectral changes constituting to BCOs have not been identified accurately. This is the case in most classical, opera, soft pop and instrumental music [23]. In addition, the designs can be very sensitive to the algorithm parameters, e.g., window length [24].

The authors of [25] presents a comparison between several onset detection methods which were submitted to the ISMIR 2006 competition [26]. The methods discussed include the works presented by [4], [8], [20], [21] and several others. The authors show that the method proposed by [20], which maneuvers temporal/spectral data, together with statistical estimation techniques, outperforms other methods by a considerable margin.

Methods such as [16]–[19] uses a machine learning based approach where there is no computation of an OEF. Based on recent results of ISMIR [26], the research conducted in [17]–[19] appears to be the best among others. However, for machine learning algorithms, usually the existence of a reasonable training data set is necessary to achieve better accuracies.

III. PROPOSED METHOD, AN OVERVIEW

The proposed method is based on the discrete S-transform [1, § III]. Moreover, the overall method is divided into two sections;

- 1) Onset envelopes by band spitting.
- 2) Onset envelope isolation.

Recall that the existing methods rely on a single STFT-TFR followed by an associated onset envelope for beat detection. Intuitively, to get the benefits of frequency dependent resolution of S-transform, it is suggestive to split the TFR into several bands and to process different subbands separately [3], [4], [20]. Such a splitting and a processing can avoid or at least minimize the masking off and suppression of desired BCO information from undesired spectral information. Thus, we first consider a band splitting followed by an onset envelope computation for each band (Figure 1). Note that, of the

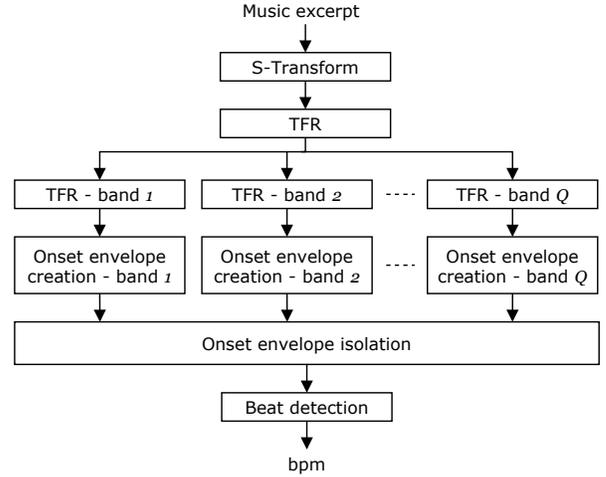


Fig. 1. Block diagrams of Proposed Method.

several onset envelopes present, the beat information may be encoded in some, depending on the rhythm instrument used. The challenge is then to pick the ‘best’ one that encodes the BCOs of the underlying music. This is the second stage of our proposed method, in particular the onset envelope isolation, see Figure 1.

In the sequel, we discuss in more detail, the computation of onset envelopes by band splitting [cf. § IV] and onset envelope isolation [cf. § V].

IV. ONSET ENVELOPES BY BAND SPITTING

Let us first outline the proposed algorithm for onset envelope computation. We assume that the musical excerpt is provided in mono format.

Algorithm 1

Input:

- Mono audio file, $\{x[n]\}_{n=0}^{N-1}$.
- Downsampling factor D , a positive even integer.
- Subband size, K such that $\lfloor (N-1)/D \rfloor = 2QK-1$ for some positive integer Q .

Steps:

- 1) Downsampling:

$$y[n] = x[nD], \quad n = 0, \dots, M-1,$$

where $M = 1 + \lfloor (N-1)/D \rfloor$.

- 2) Compute M -Discrete Fourier Transform $\{Y[k]\}_{k=0}^{M-1}$ of $\{y[n]\}_{n=0}^{M-1}$, where

$$Y[k] = \frac{1}{M} \sum_{n=0}^{M-1} y[n] \exp\left(-\frac{j2\pi nk}{M}\right).$$

- 3) Compute Discrete S-Transform matrix $F \in \mathbb{C}^{(M/2) \times M}$, whose (p, n) -th element is given by:

$$F[p, n] = \begin{cases} \sum_{m=0}^{M-1} Y[m+n] \exp\left(\frac{j2\pi mp}{N} - \frac{2\pi^2 m^2}{n^2}\right), & \text{if } n \neq 0 \\ \frac{1}{M} \sum_{m=0}^{M-1} y[m], & \text{otherwise,} \end{cases}$$

where $n = 0, \dots, M-1$ and $p = 0, \dots, M/2-1$. Define $S \in \mathbb{R}_+^{(M/2) \times M}$ as follows:

$$S(p, n) = |F(p, n)|, \quad \forall p, n.$$

4) Split S by rows,

$$S = [S_1^T \ S_2^T \ \dots \ S_Q^T]^T,$$

with $S_i \in \mathbb{R}^{K \times M}$ representing i -th block of S .

5) For each block S_i , compute the mean (over rows) $r_i \in \mathbb{R}^M$, i.e.,

$$r_i = K^{-1} S_i^T \mathbf{1},$$

where $\mathbf{1} \in \mathbb{R}^K$ is a K -vector with all ones.

Output:

- Onset envelopes: return $r_i \in \mathbb{R}^M$ associated with sub-band i , $i = 1, \dots, Q$.

Algorithm starts with a sampled musical excerpt denoted by the sequence $\{x[n]\}_{n=0}^{N-1}$. Note that, the smaller N or the duration T of the musical excerpt is, the lesser the computational burden of the algorithm. Therefore, the duration T can be chosen intelligently for efficient implementation of the algorithm. Note that, the tempo of a music can usually range from 60 bpm to 240 bpm [27]. Therefore, T can be on the order of few seconds to extract useful beat information. For example, even in the worst-case, i.e., when the musical excerpt is of 60 bpm, a $T = 4$ second musical excerpt can be used to capture 4 beats for further processing.

The downsampling factor D also plays a key role for efficient implementation of the algorithm [cf. step (1)]. In other words, the larger D is, the smaller M , and therefore, the lesser computational burden of the algorithm [cf. step (2), (3)]. A better choice for D can be argued by considering the frequencies of rhythm instruments. Note that the frequencies of rhythm instruments' typically range from 32Hz to 512Hz [28]. Thus, sampling is to be done at a rate no smaller than 1024Hz to avoid aliasing. Therefore, for a musical excerpt sampled at a rate $f_s = 44100$ Hz [28], $D = 40$ corresponds to a sampling frequency 1102.5Hz (≥ 1024 Hz) and $M = 4410$ samples in a $T = 4$ s period.

The idea of band splitting is essentially to extract the potentials of S-transform in a frequency dependent resolution. Thus, the choice of K is to be such that it is large enough to hold a sufficient spectral energy concentration to emphasize BCOs (if any). On the other hand, K should be small enough to minimize the masking off of important BCO information (if any) from spectral contents within the subband itself. Numerical experiences suggest that a K on the order of 200 for a

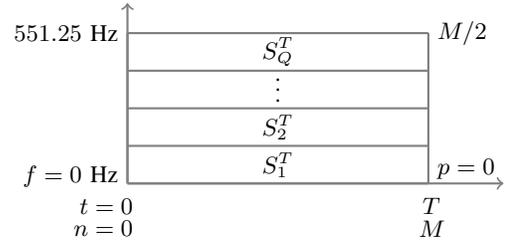


Fig. 2. Splitting of discrete S-transform matrix $S \in \mathbb{R}^{(M/2) \times M}$

$T = 4$ s period, or in other words, a subband width on the order of 50Hz is a good choice. Note that the subbands are indexed by $1, \dots, Q$ for simplicity.

A concise depiction of our considered TFR, in particular, the absolute discrete S-transform matrix $S \in \mathbb{R}^{(M/2) \times M}$ is shown in Figure 2, together with the considered splitting. Note that the TFR is plotted only for the range $0\text{Hz} \leq f \leq 551.25\text{Hz}$ and $0\text{ s} \leq t \leq T\text{ s}$, because the upper frequency band $551.25\text{Hz} < f \leq 1102.5\text{Hz}$ is just a repetition of S .

After having determined S and its splitting [cf. step (3), (4)], step (5) computes the onset envelopes of each subband. The output of the algorithm is the onset envelopes for each subband, which is used by the onset envelope isolation stage.

V. ONSET ENVELOPE ISOLATION

Given onset envelopes $r_i \in \mathbb{R}^M, i = 1, \dots, Q$, the task of the isolation stage is to choose *one* envelope that can potentially encode the BCO information. To this end, the key idea is to associate each r_i , with a real number b_i , so that, the bigger b_i is, the higher the likeliness of r_i carrying BCO information. Let us first outline the algorithm.

Algorithm 2

Input:

- Onset envelopes: $r_i \in \mathbb{R}^M, i = 1, \dots, Q$.
- Local maxima (peak) separation n_p .
- Threshold steps H .
- Isolation accuracy level $\epsilon > 0$.

Steps:

For each $i \in \{1, \dots, Q\}$,

- Normalization: compute \tilde{r}_i as $\tilde{r}_i = r_i / \|r_i\|_\infty$, where $\|\cdot\|_\infty$ is the ℓ_∞ norm.
- Upper envelope computation: Determine the upper envelope $u_i \in \mathbb{R}^M$ by using *cubic* spline interpolation over local maxima of \tilde{r}_i separated by at least n_p samples [29, § IV].
- Centering: Compute \hat{r}_i as

$$\hat{r}_i = [\tilde{r}_i - (1^T u_i / M) \mathbf{1}]_+,$$

where $\mathbf{1} \in \mathbb{R}^M$ is a M -vector with all ones and $[x]_+$ is the projection of x onto \mathbb{R}_+^M .

¹That is the vector obtained by taking the nonnegative part of each component of x and replacing each negative component with 0.

- 4) Thresholding and clustering: Divide equally, the range $\mathcal{H}_i = [0, \max(\hat{r}_i)]$ into H segments indexed by $\{1, \dots, H\}$.

For each segment $j \in \{1, \dots, H\}$

- Let threshold $h = l_j$, the lower level of segment j .
- Let $\mathcal{I} = \{k \mid (\hat{r}_i)_k \geq h\}$, the set of indexes whose associated components are larger than or equal to the threshold h .
- Determine the set partition $\{\mathcal{I}_m\}_{m=1}^{M_i}$ of \mathcal{I} such that, $\mathcal{I}_m \cap \mathcal{I}_{\bar{m}} = \emptyset \forall m, \bar{m}$ and the elements of any set are *consecutive*.
- Let $\{\bar{I}_m\}_{m=1}^{M_i}$ be the ordered sequence, where I_m is of mean of the elements of \mathcal{I}_m .
- Define $c_{ij} \in \mathbb{R}^{M_i-1}$ as follows:

$$c_{ij} = [I_{12}, I_{23}, \dots, I_{(M_i-2)(M_i-1)}, I_{(M_i-1)(M_i)}]^T,$$

where $I_{mn} = I_n - I_m$ and let $v_{ij} = (1^T c_{ij}) / \|c_{ij}\|_2$.

- 5) Define $b_i \in \mathbb{R}$ as follows:

$$b_i = \max_{j \in \{1, \dots, H\}} v_{ij}.$$

Output:

- Onset envelope isolation:

$$\mathcal{I}^* = \{i \mid |1 - b_i| \leq \epsilon, i \in \{1, \dots, Q\}\}.$$

- If $\mathcal{I}^* = \emptyset$, return an exception `Isolation Failure`,
Otherwise return r_{i^*} , where the partition index $i^* \in \mathcal{I}^*$.

The first step is a preconditioning step, where r_i is normalized to yield \tilde{r}_i . For an illustration, see Figure 3-(a). It is reasonable to assume that most of the relatively lower level amplitudes of \tilde{r}_i do not carry BCO information. Therefore, we consider only the amplitudes of \tilde{r}_i above some level. More specifically, the level is chosen to be the *mean* [Figure 3-(b), dotted curve] of the *upper envelope* u_i [Figure 3-(b), solid curve] determined at step (2). Step (3) removes the mean aforementioned from \tilde{r}_i to yield \hat{r}_i , cf. Figure 3-(c).

Note that the upper envelope u_i in step (2), computed by using cubic spline interpolation corresponds to some local maxima² of \tilde{r}_i whose separation is at least $n_p \in \mathbb{Z}$ samples. For example, Figure 3-(b) shows u_i of \tilde{r}_i in Figure 3-(a) for $n_p = 1$.

Step (1), (2), as well as (3) of the algorithm correspond to preconditioning of the input r_i . In contrast, step (4) is the key for envelope isolation, which capitalizes on a clustering of components of \hat{r}_i by using a thresholding mechanism. To see this, first suppose the range of frequencies of the underlying rhythm instrument overlaps with subband $i \in \{1, \dots, Q\}$. Thus, there is a high potential that \hat{r}_i contains nonzero components, which correspond to the BCOs. In addition, their neighboring components can also be nonzeros due to the *spectral leakage* caused by windowing. As a result, \hat{r}_i can

resemble a sequence as shown in Figure 3-(c), where there are clusters of nonzero components (*nonzero clusters*) that are separated by clusters of zero components (zero clusters). For example, \hat{r}_i in Figure 3-(c) has 3 nonzero clusters. Because of the periodicity of BCOs, the ‘distance’ between consecutive pairs of nonzero clusters should be the same. However, for any subband $\bar{i} \neq i$, the characteristics of the nonzero clusters mentioned above, do *not* apply. This is indeed the key to isolate subband i from others.

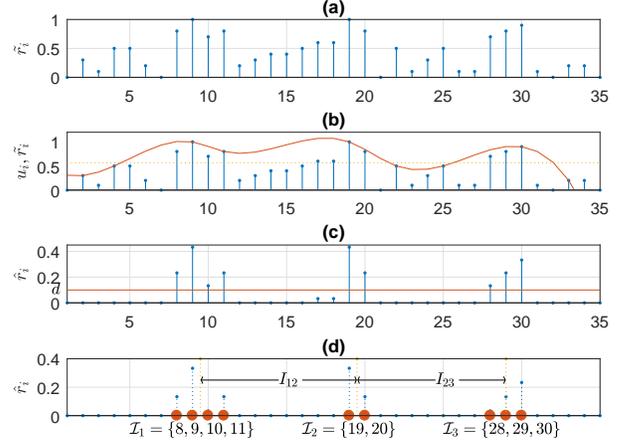


Fig. 3. Signatures of split frequency bands

Steps (4)-a to (4)-e, correspond to clustering and the distance computation between consecutive pairs of nonzero clusters of \hat{r}_i . First, a threshold h is given, cf. step (4)-a and Figure 3-(c). Then components of \hat{r}_i which are greater than or equal to h is isolated into \mathcal{I} , cf. step (4)-b. For example, Figure 3-(c) shows that $\mathcal{I} = \{8, 9, 10, 11, 19, 20, 28, 29, 30\}$. Step (4)-c partitions \mathcal{I} into subsets $\{\mathcal{I}_m\}_{m=1}^{M_i}$, where each subset corresponds to a nonzero cluster. For example, from Figure 3-(d), we have $M_i = 3$ subsets (one for each nonzero cluster), denoted $\mathcal{I}_1, \mathcal{I}_2$, and \mathcal{I}_3 , where $\mathcal{I}_1 = \{8, 9, 10, 11\}$, $\mathcal{I}_2 = \{19, 20\}$, and $\mathcal{I}_3 = \{28, 29, 30\}$. Step (4)-d computes the center of gravity of each subset, denoted $\{I_m\}_{m=1}^{M_i}$. Particularized to our example, we have $I_1 = 9.5$, $I_2 = 19.5$, and $I_3 = 29$, cf. Figure 3-(d). The distance between consecutive pairs of nonzero clusters are simply given by the $(M_i - 1)$ -vector $[I_{12}, I_{23}, \dots, I_{(M_i-2)(M_i-1)}, I_{(M_i-1)(M_i)}]^T$, cf. step (4)-e. This is illustrated in Figure 3-(d), where the distance between nonzero cluster 1 and 2 is I_{12} and that of nonzero cluster 2 and 3 is I_{23} . Finally, recall that the ‘distance’ between consecutive pairs of nonzero clusters should be the same if r_i contains BCOs. Mathematically, this corresponds to a larger *inner product* of vectors c_{ij} and $1 \in \mathbb{R}^{M_i-1}$. Therefore, step (4)-e computes such inner products denoted $\{v_{ij}\}_{j=1}^D$ and step (5) chooses the best.

At the end of step (5), associated with each subband, we have a real number b_i which characterizes the likeliness of r_i containing BCOs. Finally, for the specified isolation accuracy ϵ , isolated subband indexes are returned.

²We say $k \in \mathbb{Z}$ is a local maximum of $x \in \mathbb{R}^M$ whenever $(x)_{k-1} < (x)_k < (x)_{k+1}$, where $(x)_k$ represents the k -th component of x .

Finally, a potential BPM value is computed as

$$\text{BPM} = \lceil (1^T c_{ij}) \rceil / \text{length}(c_{ij}) \quad (1)$$

for some $i \in \mathcal{I}^*$, where $\lceil x \rceil$ represents the rounding of x to the nearest integer and $\text{length}(y)$ represents the length of vector y .

VI. COMPUTATIONAL COMPLEXITY

A vast majority of the existing methods use the STFT to obtain a TFR. The asymptotic complexity for the STFT is $\mathcal{O}(N \log N)$, where N is the samples used in the underlying FFT operations³ [30].

The discrete S-transform, on the other hand, has an asymptotic complexity of $\mathcal{O}(N^3)$ [31]. However, by exploiting structural properties, variants of discrete S-transforms, such as fast discrete orthonormal Stockwell transform can be computed, still in $\mathcal{O}(N \log N)$ [31, Theorem 6.1].

VII. RESULTS

This section compares the performance of the proposed method with the algorithms documented in [5] and [20], which we consider as benchmarks *A* and *B*, respectively. Algorithm in [5] can be considered to be superior among the methods based on pure temporal/spectral analysis methods [26]. On the other hand, the work by [20] is the best among methods that rely on temporal/spectral data, together with statistical estimation.

In our simulations, we consider two publicly available datasets - the Ballroom dataset, and the Songs dataset, which comprise of 698 and 465 song excerpts, respectively [25, § III-B]. The tempo, genre, and style distribution of the datasets are given by [25, § III].

Note that the sampling rate of each song excerpt is 44.1kHz. A downsampling factor of $D = 40$, a subband size $K = 1103$, and $Q = 10$ subbands are used as inputs to Algorithm 1. In the case of Algorithm 2, we use $n_p = 40$, $H = 100$, and $\epsilon = 10^{-3}$.

To exemplify the outputs of the proposed algorithms, we first consider an arbitrarily chosen classical music excerpt in the Songs dataset.

Figure 4 shows the output r_i , $i = 1, \dots, 10$ for the considered music excerpt. Results show that r_9 and r_{10} can apparently isolate the BCOs.

Figure 5 shows v_{ij} versus j for each subband i , $i = 1, \dots, Q$. Results indicate that v_{10j} yields values almost close to 1 for some thresholds l_j [cf. step (4)-a]. More specifically, Algorithm 2 returns $\mathcal{I}^* = \{10\}$, which corresponds to $b_{10} = 0.999895$ [cf. step (5)]. The resulting BPM is 88 [cf. (1)], which is identical to the ground-truth tempo.

To see the performance of the proposed algorithms on average, we ran the algorithms separately for each data set.

As discussed in [25], we considered the same two metrics to measure the accuracy of the system. In particular, we have

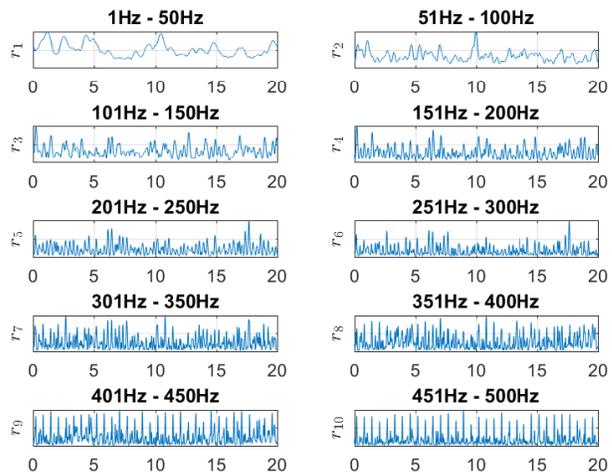


Fig. 4. Split frequency bands r_i for $i = 1, \dots, 10$

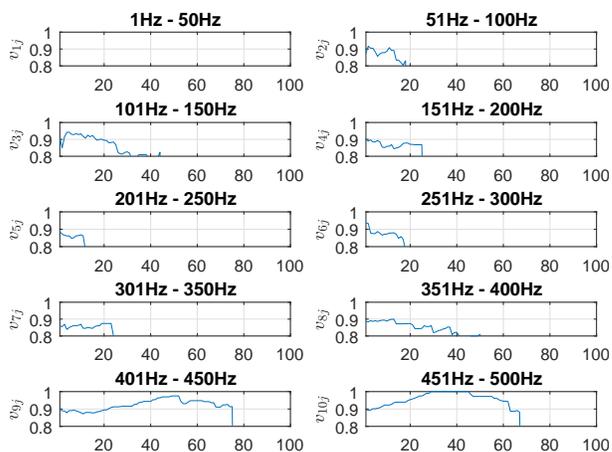


Fig. 5. v_{ij} versus j for each subband i , $i = 1, \dots, Q$

- **Accuracy 1:** The percentage of tempo estimates within 4% of the ground-truth tempo.
- **Accuracy 2:** The percentage of tempo estimates within 4% of either the ground-truth tempo, or half, double, three times, or one third of the ground-truth tempo.

Figure 6 depicts the results of percentage accuracies of the proposed algorithm compared with the two benchmarks. Results show that the proposed method has a better performance than [5]. Note that, this gain can be accomplished with the same computational complexity, cf. § VI. Results further show that the method proposed in [20] is superior to the proposed method. This is not surprising, because, unlike the proposed method, the algorithm in [20] relies on many computationally intensive operations, e.g., filtering, comb filter operations, discrete power spectral estimations, statistical estimation of period and phase of underlying time series within a hidden Markov model, among others. Therefore, results suggest that the proposed method holds an advantage in that it is less computationally intensive than the benchmark [20], yet with a comparable performance.

³e.g., the STFT window length.

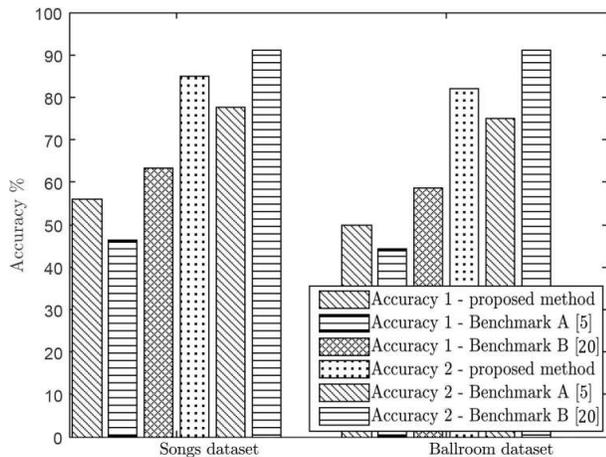


Fig. 6. Comparison of Accuracy 1 and Accuracy 2 values for both datasets

VIII. CONCLUSIONS

In this paper, a beat causing onset (BCO) detection method based on the S-transform has been proposed. The method provided an advantage over the approaches that are purely based on classic temporal/spectral analysis. The frequency dependent window dilation used in S-transform has been the key to yield such performances by exploiting better frequency resolution at lower frequencies, where BCOs generally occur. Compared to state-of-the-art algorithms, the proposed method is less resource intensive. For example, our method does not require any *a priori* information of the underlying music, unlike the statistical estimation based approaches. Moreover, the method does not require training datasets like in the methods based on state-of-the-art machine learning techniques. The result is a graceful trade-off between the performance and the required computational burden and the resources.

REFERENCES

- [1] R. Stockwell, L. Mansinha, and R. P. Lowe, "Localization of the complex spectrum: the S-transform," *IEEE Transactions on Signal Processing*, vol. 44, pp. 998–1001, Apr. 1996.
- [2] R. Marxer and J. Janer, "Low-latency bass separation using harmonic-percussion decomposition," in *International Conference on Digital Audio Effects Conference (DAFx-13)*, (Maynooth, Ireland), pp. 290–294, Sept. 2013.
- [3] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 6, (Phoenix, AZ, USA), pp. 3089–3092, Mar. 1999.
- [4] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, pp. 588–601, Apr. 1998.
- [5] D. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, 2007.
- [6] B. McFee and D. P. W. Ellis, "Better beat tracking through robust onset aggregation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (Florence, Italy), pp. 2154–2158, May 2014.
- [7] J. Laroche, "Efficient tempo and beat tracking in audio recordings," *Journal of the Audio Engineering Society (JAES)*, vol. 51, pp. 226–233, Apr. 2003.
- [8] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of musical signals," in *Proceedings of the 5th International Conference on Music Information Retrieval*, (Barcelona, Spain), pp. 158–164, Oct. 2004.

- [9] A. Stark, M. Davies, and M. Plumbley, "Real-time beat-synchronous analysis of musical audio," in *Proceedings of the 12th International Conference on Digital Audio Effects (DAFx-09)*, (Como, Italy), Sept. 2009.
- [10] F. Wu, T. Lee, J. Jang, K. Chang, C. Lu, and W. Wang, "A two-fold dynamic programming approach to beat tracking for audio music with time-varying tempo," in *Proceedings of the 12th International Society for Music Information Retrieval Conference, ISMIR 2011*, (Miami, FL, USA), pp. 191–196, Jan. 2011.
- [11] M. Goto and Y. Muraoka, "Beat tracking based on multiple-agent architecture a real-time beat tracking system for audio signals," in *Proceedings of the Second International Conference on Multiagent Systems*, (Kyoto, Japan), pp. 103–110, Dec. 1996.
- [12] Y. Shiu, P. C. Cho, and C. J. Kuo, "Robust online beat tracking with kalman filtering and probabilistic data association," *IEEE Transactions on Consumer Electronics*, vol. 54, pp. 1369 – 1377, Oct. 2008.
- [13] A. Cemgil, B. Kappen, P. Esain, and H. Honing, "On tempo tracking: Tempogram representation and kalman filtering," *Journal of New Music Research*, vol. 29, pp. 259–273, May 2000.
- [14] J. R. Zapata, E. P. Davies, and E. Gomez, "Multi-feature beat tracking," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, p. 816825, Apr. 2014.
- [15] N. Degara, E. A. Rua, A. Pena, S. Torres-Guijarro, M. Davies, and M. D. Plumbley, "Reliability-informed beat tracking of musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, p. 290301, Jan. 2013.
- [16] D. Fioocchi, "Beat tracking using recurrent neural network: a transfer learning approach," Master's thesis, Politecnico di Milano, Milan, Italy, 2017.
- [17] S. Bock and M. Schedl, "Enhanced beat tracking with context-aware neural networks," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, (Paris, France), p. 135139, Sept. 2011.
- [18] S. Bock, F. Krebs, and G. Widmer, "Accurate tempo estimation based on recurrent neural networks and resonating comb filters," in *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR 2015)*, (Malaga, Spain), p. 625631, Oct. 2015.
- [19] S. Bock, A. Arzt, F. Krebs, and M. Schedl, "Online real-time onset detection with recurrent neural networks," in *Proceedings of the 15th International on Digital Audio Effects (DAFx-12)*, (York, UK), p. 301304, Sept. 2012.
- [20] A. P. Klapuri, A. J. Eronen, and J. T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1832 – 1844, Sept. 2006.
- [21] S. Dixon, "Automatic extraction of tempo and beat from expressive performances," *Journal of New Music Research*, vol. 30, pp. 39–58, Aug. 2001.
- [22] M. Davies and M. Plumbley, "Context-dependent beat tracking of musical audio," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1009–1020, Mar. 2007.
- [23] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proceedings of the 5th Digital Audio Effects (DAFx-02) Conference*, (Hamburg, Germany), pp. 33–38, Nov. 2002.
- [24] J. Scargle, "Studies in astronomical time series analysis. ii - statistical aspects of spectral analysis of unevenly spaced data," *Astrophysical Journal*, vol. 263, pp. 835–853, Dec. 1982.
- [25] F. Gouyon, A. Klapuri, S. Dixon, M. Alonso, G. Tzanetakis, C. Uhle, and P. Cano, "An experimental comparison of audio tempo induction algorithms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 1832–1844, Sept. 2006.
- [26] "The international society of music information retrieval." [Online]. Available: <https://www.ismir.net/>.
- [27] W. Apel, *Harvard Dictionary of Music*. Cambridge, MA, USA: Harvard University Press, 1950.
- [28] J. Watkinson, *The Art of Digital Audio*. Oxford, UK: Focal Press, 2001.
- [29] C. de Boor, *A Practical Guide to Spline*, vol. 27, pp. 40–48. Boston, NY, USA: Springer, Jan. 1978.
- [30] J. Cooley and J. Tuckey, "An algorithm for the machine calculation of complex fourier series," *Mathematics of Computation*, vol. 19, pp. 297–301, Jan. 1965.
- [31] Y. Wang and J. Orchard, "Fast discrete orthonormal stockwell transform," *SIAM Journal on Scientific Computing*, vol. 31, pp. 4000–4012, Jan. 2009.