Energy-Weighted Multi-Band Novelty Functions for Onset Detection in Piano Music

Krishna Subramani, Srivatsan Sridhar, Rohit M A, Preeti Rao
Department of Electrical Engineering
Indian Institute of Technology Bombay, India
Email: {krishna.subramani,srivatsan}@iitb.ac.in

***Abstract*—Onset detection refers to the estimation of the timing of events in a music signal. It is an important sub-task in music information retrieval and forms the basis of high-level tasks such as beat tracking and tempo estimation. Typically, the onsets of new events in the audio such as melodic notes and percussive strikes are marked by short-time energy rises and changes in spectral distribution. However, each musical instrument is characterized by its own peculiarities and challenges. In this work, we consider the accurate detection of onsets in piano music. An annotated dataset is presented. The operations in a typical onset detection system are considered and modified based on specific observations on the piano music data. In particular, the use of energy-based weighting of multi-band onset detection functions and the use of a new criterion for adapting the final peak-picking threshold are shown to improve the detection of soft onsets in the vicinity of loud notes. We further present a grouping algorithm which reduces spurious onset detections.**

I. Introduction

Music information retrieval is an active field of research where computational methods are applied to extract musically relevant attributes from either symbolic scores of the music or, more commonly, directly from the music audio signal. The applications are far ranging, from music recommendation systems and musical instrument identification to pedagogy and musicology research. Signal processing and machine learning techniques are applied to obtain descriptors of high level information related to melody, harmony, rhythm and timbre{[1], [2], [3]}

The rhythmic aspect of music lies in the notions of tempo and meter and, in turn, on the perceived beat. The tracking of the beat of the music comes relatively easily to human listeners but requires sophisticated computation for automatic extraction. The regularity of the low-level musical events, such as note onsets, in time gives rise to the perception of beats. The accurate detection of note onsets is also important in automatic music transcription. Depending on the musical instrument of interest, note onset detection poses distinct challenges. For example, the singing voice can be among the more challenging due to the variety of note onset types arising from the use of lyrics and dynamics.

In general, note onsets are easier to detect in percussive music due to the sharp transients and bursts of energy caused by the striking or plucking gestures in their playing. Although the piano is regarded as a pitched percussive instrument characterized by the presence of sharp onsets, there are some serious challenges to be addressed due to the dynamics and ornamentation that is characteristic of expressive piano playing:

1) The presence of soft notes which may not be marked by large enough energy rises, moreover being shadowed by previous loud notes that have not decayed entirely (soft notes frequently occur in the accompaniment part played by the left hand)
2) Notes can occur in very rapid succession in portions
3) Possible asynchrony between the individual notes played in a chord, leading to dispersed energy in the chord onset

The problem of onset detection is quite old, with research dating back nearly two decades{[4],[5]}. Commonly used energy, spectral magnitude and phase based techniques have been reviewed thoroughly in {[6],[7],[8]}. Most onset detection methods are essentially about detecting either energy or spectral changes between successive short-time windows of the signal. Further, taking the specific acoustic characteristics of the instrument and playing style into account is expected to lead to superior performance in onset detection.

In this work, we consider a widely applied method for note onset detection based on spectral magnitude changes, 1.e. spectral flux [7]. The introduction of multi-band processing of spectral flux is investigated for the case of piano onsets on a data set of hand-labeled piano excerpts representing the expected typical variety of onsets. A weighting function to combine the outcomes in the multiple bands is presented and the resulting novelty function is considered for adaptive thresholding for onset time detection. We further present a grouping algorithm to reduce spurious onsets. We begin with an investigation of the simpler energy based method, in order to appreciate the motivation for the spectral flux method more clearly.

II. Extracting the Novelty Curve

A. Energy (Amplitude) Based Detection

This technique is an implementation of the ideas discussed in {[1],[6],[7],[8],[9]}. It involves analyzing the signal for sudden changes in energy. As signal energy is proportional to amplitude squared, this method basically involves analyzing the derivative of the squared amplitude of the signal. Two methods were tested here:

1) Square the signal, take its discrete derivative and rectify, to only consider energy increases as potential

978-1-5386-1224-8/18/$31.00 ©2018 IEEE

onset candidates. This method is not very useful, and is vulnerable to music with high frequency content, as the novelty curve obtained will fluctuate a lot.

$$\Delta_{Energy}(n) := |E(n+1) - E(n)|_{\geq 0} \quad (1)$$

Here, Δ_{Energy} is the energy difference, and $E(n)$ is the energy (square of the amplitude).

2) Use windowing to obtain the energy of the signal in successive small windows, and then compute the changes in energy in these windows. This method works better than the above because, rather than directly computing the envelope, we are averaging the energy in the window, and then computing the discrete derivative. This can eliminate the rapid fluctuations.

$$E_w(n) := \sum_{m=-M}^{m=M} |x(n+m)W(m)|^2 \quad (2)$$

Here, $E_w(n)$ is the frame-wise energy (also called short time energy), $x(n)$ is the audio signal and $W(n)$ is an appropriate windowing function.
The energy difference is then computed using (1) .

One thing that should be noted for the energy-based method is that it works well only in music mainly composed of strong onsets (preferably by percussive or other energetic instruments). In case successive onsets are weak in amplitude, this method will fail to detect them accurately because the energy increase is too less for such weak notes. The main limitation of the energy-based detection is that it does not incorporate the changes in the spectral content of the signal, but rather only uses gross energy changes

B. Spectral Flux Based Novelty Curve

This technique is also based on the methods reviewed in {[1],[6],[7],[8]}. First, we find the Short Time Fourier Transform (STFT) of the audio signal, and obtain the squared magnitude of the STFT, which is basically the power spectrum of the signal.

$$X(n,k) := \sum_{m=0}^{N-1} w(m)x(m + n \cdot H)e^{-j2\pi km/N} \quad (3)$$

$$S_{xx}(n,k) = |X(n,k)|^2 \quad (4)$$

Here, $X(n,k)$ is the STFT of the audio signal for a frame number n and frequency bin k, and $S_{xx}(n,k)$ is the signal's short time power spectral density. $w(m)$ is a window of the frame size N samples, and H is the hop size between two frames. Because of the way in which Discrete Fourier transforms are computed, the number of frequency bins of importance is $K = N/2$.
We may also perform logarithmic compression [1], which can help us use the high frequency transients that occur at a note onset, by emphasizing them. We should be careful, because this method can also introduce spurious peaks by emphasizing noise as well.

$$\gamma(S_{xx}(n,k)) := log(1 + c \cdot S_{xx}(n,k)) \quad (5)$$

Here, in $\gamma(X(n,k))$, each element in $X(n,k)$ is replaced by $log(1 + c \cdot X(n,k))$ where c is the compression factor.
We then take the discrete derivative of the above signal, and rectify it (considering only intensity increases).

$$SF(n,k) := |\gamma(n+1,k) - \gamma(n,k)|_{\geq 0} \quad (6)$$

$SF(n,k)$ represents the spectral flux of the signal. It essentially characterizes the spectral changes in the signal.
Finally, we add up all the rows for a particular time instant, as this represents the total change in the power spectrum. The obtained array is our desired novelty curve.

$$NC(n) := \sum_{k=0}^{N/2-1} SF(n,k) \quad (7)$$

This method is expected to work well for soft onsets as well, because even if the energy associated with a change is small, its spectral distribution can change considerably. Hence, this method can pick up even relatively soft notes.

III. Dataset and Annotation

To test our algorithm, we have used a set of 29 music files made available by West Valley College in their Audio Exercises Course[10]. The songs are between 20 and 60 seconds long (with an exception of one 105 second piece), with the average duration being 34 seconds. The 29 pieces together contain 1934 note onsets. Although the set is from an introductory course directed towards beginners, it contains a fairly diverse set of songs, ranging from simple, medium-paced single-hand pieces to slightly expressive fast-paced pieces with dynamics and chords (sometimes with asynchrony). And while this set excludes complex solo piano pieces like sonatas, it provides a good starting point to evaluate existing methods and identify the precise nature of issues encountered, if any, even with these simple pieces, thus motivating the way forward.

The files in the dataset were mp3 files, with no annotated onsets. Hence the onsets were manually marked by the authors on Audacity®(v 2.1.3)[11] as outlined below:

1) Spectrograms of the piano music files were observed in Audacity, and distinct changes in notes were marked over the audio. (At a note onset, a discontinuity in the spectrogram is visually observable).
2) The music file was slowed down and played repeatedly to get a good estimate of when any notes were being played in the specific interval.
3) Estimation by listening to narrow segments was adopted when the spectrogram discontinuity was not localised enough.

IV. Proposed System

We implemented the two mentioned methods (energy and spectral flux) on Python using Essentia [12], an open source library for audio signal analysis. On applying both the above methods on our dataset, we observed the following:

1) The spectral flux method gives more prominent peaks in the novelty curve and detects a significantly larger number of onsets than the energy envelope method.

2) Because of the usage of a fixed threshold on our novelty curves for peak picking, a significant portion of the onsets fail to get detected.
3) Multiple onsets are observed around the time instants where a single onset was expected.

Based on the first observation, we chose to use the spectral flux method to propose modifications to, and try and improve it.

On investigating the techniques mentioned in [13], [14], [15], [16], [17], we were inspired to adopt a multi-band approach to onset detection. Appropriate splitting of the frequency content into bands has been realized through the use of auditory filters in [14], [15], [16], a conjugate quadrature filter bank in [13], and a set of 4 contiguous bands from 0-10 kHz in [17]. Further, the novelty curves obtained individually in each of the bands are then combined by a weighted sum. [13] uses a set of weights which assign greater precedence to higher frequency sub-bands than the lower ones, and [14] weights the onset candidates in a given segment by the maximum value of the smoothed log-amplitude envelope of that segment. In [15] and [17] however, the band-wise novelty curves are input to a neural network and a probabilistic onset evaluator network, respectively, with an intention to avoid the use of weighting and thresholding methods.

The shortcomings mentioned earlier and past work in literature thus motivated the following improvements: A set of sub-bands based on piano octaves was used, along with weights proportional to the band's total energy in the whole song. Further, adaptive thresholding was implemented so that soft onsets could be detected more reliably. Finally, a grouping algorithm was used to merge multiple closely-spaced onsets that appear instead of a single onset.

The following section explains the four main stages in the proposed method in detail. Fig. 1 depicts these stages schematically.

A. Pre-Processing the audio signal

As a first step, the audio files are passed through a low pass filter with a cutoff frequency of 6000 Hz and re-sampled to 16 kHz to avoid the effect of higher frequency noise in the obtained novelty curve, and to reduce computation time and memory. Also, different audio files can have different signal parameters depending on how they have been recorded or synthesized. They should hence be normalized before any further analysis, so that a universal algorithm can be used across different audio files. We tried two different normalization methods, as described below.

1) Divide by the signal's maximum amplitude to normalize the amplitude.
2) Find the window with the maximum energy and divide throughout by this window's energy.

The former works when we want to adjust for the amplitude level of the song as a whole. The latter works better even if there is an increase of amplitude in one particular window (as in a window with a prominent onset). In this case, the former method would have made the weaker onsets elsewhere even weaker. This latter method was experimentally observed to work better by detecting a larger number of onsets.

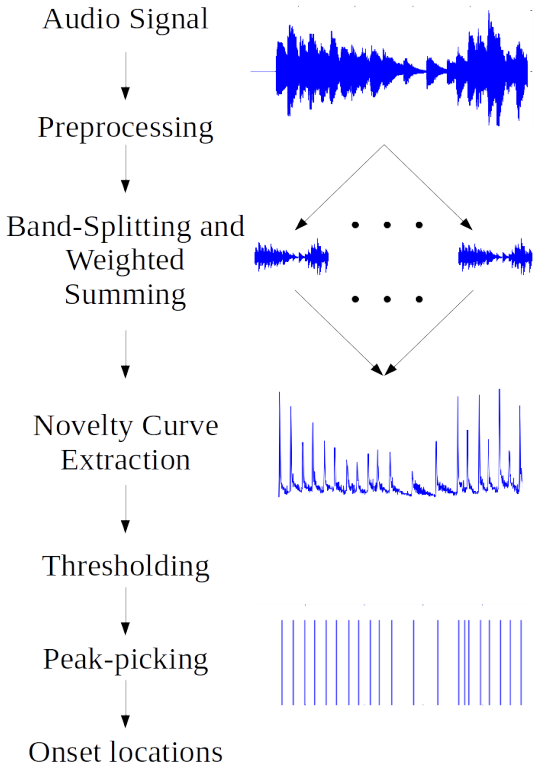

Fig. 1. Flow of analysis for onset detection

B. Band-Splitting and Weighting

The filtered and normalized audio is split into 6 frequency bands which go from 0Hz to 6400Hz. This splitting is as per the 8 standard piano octave bands. The first band from 0-200Hz contains the first 3 octaves, and the bands 200-400Hz, 400-800Hz, 800-1600Hz, 1600-3200Hz, and 3200-6400Hz, contain the remaining 5 octaves. Each of these 5 bands approximately contains the fundamental frequencies of the notes going from the A of one octave to the G of the next octave. The fundamental frequencies of the standard piano notes occur between 27.5 Hz and and 4186 Hz. A splitting based on musical octaves allows to adjust the method in a musically intuitive manner, by analysing which octaves are played louder or softer, for instance. A novelty curve for each of these sub-bands is then computed based on the spectral flux method discussed above (Eqs. 3-7). The novelty curve of each sub-band is weighted by the energy in that sub-band (in the whole song) as a fraction of the net energy in all the sub-bands(in the whole song). Such a weighting scheme helps detect softly played notes with energy content in specific frequency bands. For instance, in pieces containing a mix of low and high octave notes, band-wise energy weighting improves the detection of low frequency note onsets which

are often played very softly and are hence hard to detect.

$$NC(n) := \sum_{i=1}^{6} w_i \cdot NC_i(n) \tag{8}$$

$$w_i := \frac{E_i}{\sum_{i=1}^{6} E_i} \tag{9}$$

Here, $NC(i)$ is the novelty curve computed for the i^{th} frequency band, and $w(i)$ is the weighting coefficient as defined above. E_i is the energy content of the whole song in the i^{th} band. It was seen that the weights are larger and vary across songs for the first 3 frequency bands, but are smaller and fairly constant for the higher 3 bands. This is because, the songs in the dataset most often contain notes in the lower 3 bands, with the exact content in these bands varying depending on the dynamics of the song.

The weighting scheme described above returns a global weight per band for the entire duration of the song. A more adaptive approach with weights computed using the derivative of short-time energy instead of the entire signal's energy was also experimented with. While this method proved effective in the detection of extremely soft onsets, it did not offer an improvement in performance over the entire dataset. This was because of the considerable amount of parameter tuning required in the post processing of the novelty curve of every audio signal, to make the energy-derivative curve possess sharp enough peaks to serve as an appropriate weighting function.

C. Thresholding

The novelty curve obtained after adding all frequency bands, is first normalized by dividing by its maximum value. Those time instants are chosen as note onsets where the local peak in the novelty curve crosses a given threshold. A drawback of using a fixed threshold was the missed detection of soft onsets occurring immediately following a loud note. This is explained by the spectral change arising from the soft onset being over-shadowed by the strong and extended decay of the loud note strike. This motivated us to relax the threshold for a few frames immediately after the frame containing a strong onset. This thresholding method is different from the adaptive thresholding used in [6] and [7] which modify the threshold based on a moving average of the novelty curve. The method in this work uses the difference in the moving average, to focus on the soft onsets. The variable threshold function, $t(n)$, a function of frame number n is defined as:

$$t(n) := c + \lambda \cdot \{g(n) - g(n-h)\} \tag{10}$$

$$g(n) := \sum_{i=n}^{i=n+W} NC(i) \tag{11}$$

Here, c is a fixed threshold value, λ is a scaling factor, and $g(n)$ is a sum in a window of length W frames after the frame n. The time duration between consecutive frames depends on the hop size used in the spectral flux method (Eq.(3)) (which is 5 ms in this work). A frame is chosen as an onset-frame if the value of the novelty curve in that frame is above the corresponding value of $t(n)$.

The difference $g(n) - g(n-h)$ is negative for h frames after a strong onset, which reduces the threshold, resulting in better detection of soft onsets in that frame. This thresholding method not only increases the correctly detected onsets but also decreases the false positives, as the threshold becomes higher after a period without onsets (difference is positive or almost zero), thus rejecting small non-onset peaks in that period. Actual onsets after a period of silence show very high peaks in the novelty curve, so the higher threshold still captures them. This can be inferred from the results shown later, in Table 1.

The final values of the parameters were chosen after observing the precision and recall values for different values of the parameters, as described in Section V ahead.

D. Grouping

One of the problems observed was that multiple onsets were detected at points where only one onset was expected. This was happening because of some rapid fluctuations in the novelty curve at the onset points instead of just a single peak. To address this problem, we designed a time domain grouping algorithm to replace multiple closely-spaced onsets caused due to one primary onset, with a single onset. This is similar to the 'temporal integration' step in [14]. It works by forming clusters of onsets, and adding an onset to the cluster if it lies within a window of 30 ms from the previous onset that was added to the cluster. Thus, it essentially clusters onsets which lie too close to each other and are not likely to represent distinct onsets, and outputs the average time instant of the onsets in the cluster (to account for the estimation error). The allowed time gap was chosen by observing that there were onsets as close as 30 ms in the dataset used (in case of asynchronous onsets of the notes of a chord). There is thus one caveat - if two successive onsets actually occur too close in time, they may be wrongly grouped into a single onset. However this is extremely rare for a time gap of less than 30 ms.

Figure 2 shows the effect of the grouping algorithm. The multiple closely spaced lines in the upper graph, which represent multiple-onset detection, have been grouped to yield a single onset instant, as shown in the lower graph.

V. Testing and Results

1) The methods mentioned above were tested on the dataset, and the corresponding novelty curves, and time instants of onsets were obtained. As a preliminary evaluation measure, the onset locations were tested by listening to the audio files superimposed with beeps at the onset locations.
2) To perform a more thorough evaluation, an onset evaluation algorithm was created (inspired from the algorithm used by the IEEE Signal Processing Cup)[18] to compare the detected onsets and annotations. The percentage of undetected onsets, false positives and false negatives were determined.
3) We compared the performance of our proposed algorithm against a benchmark SF (spectral flux) algorithm, based on the spectral flux method itself, but without the band-splitting, adaptive thresholding and grouping.

On comparing our results with the benchmark, we obtained a significant improvement in the number of onsets that were detected. One such case is highlighted in figure 3, which shows a 5 second clip of one of the more complex pieces of the dataset, with red dotted lines indicating the ground truth onset locations and the blue solid lines indicating the onsets determined by the algorithm.

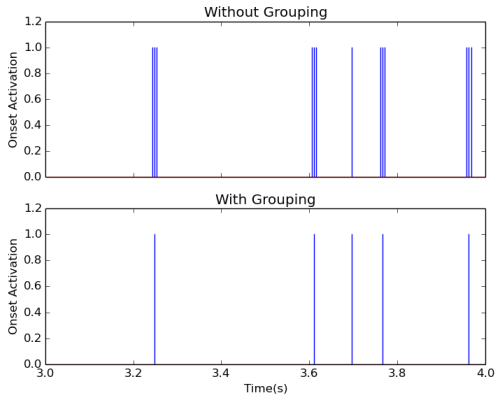

Fig. 2. Time Domain Grouping

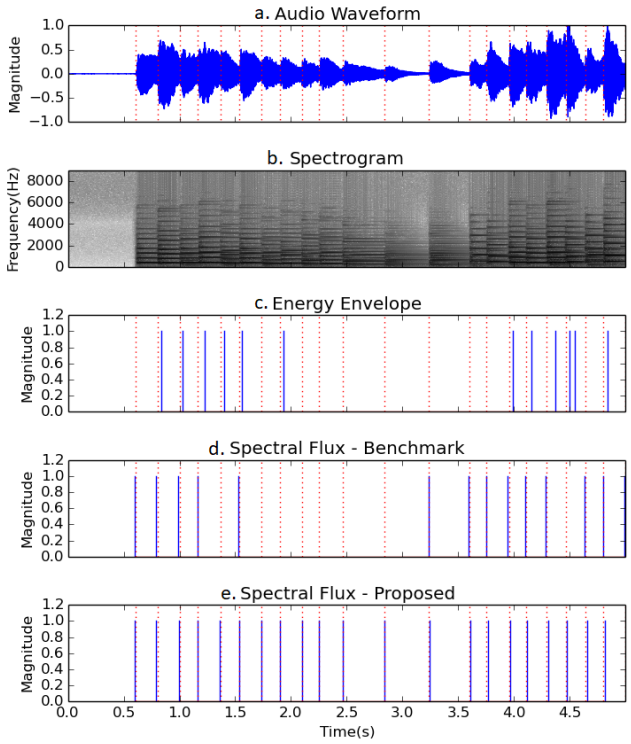

Fig. 3. Comparison of different detection functions for a 5s section

We use the Precision, Recall and F-Measure to compare the average performance of both the algorithms.

$$Precision = \frac{tp}{tp + fp} \tag{12}$$

$$Recall = \frac{tp}{tp + fn} \tag{13}$$

$$F - Measure = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{14}$$

with tp being the number of true positives, fp being the number of false positives and fn the number of false negatives. Table 1 shows the average values of precision and recall computed over the entire dataset.

The proposed algorithm was run for different values of the parameters used in the variable thresholding method, and the precision and recall values were obtained for each of them. A section of the precision vs recall plot obtained is shown in Fig. 4, comparing a set of 6 curves obtained from Eq.(11) at two values of W - the lower 3 curves at W = 100 and the upper 3 at W = 110. On experimenting with several values of W, the performance was observed to deteriorate for values both greater and lesser than W = 110(The reason for this sharp dependence in W has to be investigated further). Hence the other parameters were chosen at this value of W = 110 such that they maximized the recall without significantly reducing the precision from its maximum value for this algorithm (it can be seen from the plot that precision saturates at close to 98%). Increasing the recall beyond this by a fraction of a percent decreases precision by 2-3%, thus discouraging us from choosing those points. The results corresponding to the optimum set of parameter values are shown in the Table 1 below.

The proposed method with a constant threshold gives a 9% increase in the recall value, with a small increase in the number of false positives, which is indicated by the 1.5% drop in precision. This demonstrates the sensitivity of the proposed method to soft onsets, as hypothesised. Additionally, using an adaptive threshold further increases both precision and recall values, thereby reducing the number of both false positives and negatives, as mentioned in Section IV-C.

Algorithm	Precision	Recall	F-Measure
Benchmark SF	98.42	85.03	91.24
Constant Threshold	96.9	94.0	95.43
Adaptive Threshold	97.52	96.62	97.07

Table1: Results comparing the regular SF method(Benchmark SF) with the proposed SF method using both constant and adaptive thresholding (for the optimum set of parameters)

VI. Challenges Faced

Although our proposed algorithm does perform better than the benchmark for detecting relatively softer onsets, there are still cases where our algorithm fails to detect onsets. One particular case of interest is Song no. 25 in the dataset [10], which contains a repeating series of extremely soft onsets in the lower octave played after a strongly played note in the higher octave. These notes are in fact barely audible to the ear, and can only be perceived by the listener based on their recurring pattern. The other limitation that remains is the false positives ratio (2.48% of the detected onsets). However, it is observed that most of these occur only as groups of multiple onsets around a single onset.

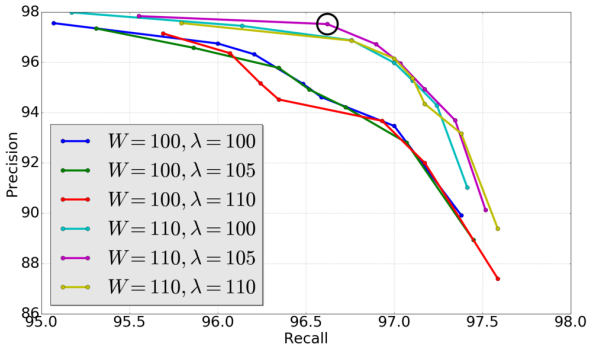

Fig. 4. Precision vs Recall (%) plot for various values of parameters c, λ, W, h in the adaptive thresholding algorithm with h set to 1 and c varying between 0.08 and 0.12 for each curve. The encircled point shows the performance obtained for the set of chosen parameters

VII. Conclusion and Future Plans

The main distinctive features of this proposed system were the energy-weighted band splitting of the novelty curve, adaptive thresholding, and the grouping of spurious onsets. An implementation of the energy-weighted band splitting alone has increased recall from 85% to 94%, thus demonstrating its success in detecting most of the soft onsets which were earlier undetected. However there was about 1.5% decrease in the precision as well. On adding the adaptive thresholding and grouping methods, the recall increases further to 96.6% and the precision also increases slightly to 97.5%. Thus, the methods mentioned in this work have helped detect a much greater number of onsets correctly with a small increase in false detections.

Further work along the same lines to include -

1) Testing the proposed methods on more complex music from professional performances
2) Using a combination of magnitude and phase information [7] (complex domain based onset detection)
3) Using Recurrent Neural Networks i.e. Bidirectional Long Short Term Memories [19], [20] or Support Vector Machine based approaches [21] to obtain higher efficiency with onset detection
4) Extracting beat and tempo information from the music using the obtained onsets[22], [23]

References

[1] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[2] J. S. Downie, "Music information retrieval," *Annual review of information science and technology*, vol. 37, no. 1, pp. 295–340, 2003.

[3] P. Herrera-Boyer, G. Peeters, and S. Dubnov, "Automatic classification of musical instrument sounds," *Journal of New Music Research*, vol. 32, no. 1, pp. 3–21, 2003.

[4] C. Tait and W. Findlay, "Wavelet analysis for onset detection," in *Proceedings of the International Computer Music Conference*, pp. 500–503, International Computer Music Association, 1996.

[5] P. Masri, *Computer modelling of sound for transformation and synthesis of musical signals.* PhD thesis, University of Bristol, 1996.

[6] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on speech and audio processing*, vol. 13, no. 5, pp. 1035–1047, 2005.

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

[8] M. Muller, D. P. Ellis, A. Klapuri, and G. Richard, "Signal processing for music analysis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1088–1110, 2011.

[9] P. Grosche, *Signal processing methods for beat tracking, music segmentation, and audio retrieval*. PhD thesis, Grosche, Peter, 2012.

[10] "MUSIC 30A/B: Beginning Piano - Eckstein Audio Exercises by West Valley College on Apple Podcasts." https://itunes.apple.com/us/podcast/music-30a-b-beginning-piano-eckstein-audio-exercises/id380860116?mt=2, accessed 21-01-2018.

[11] "Audacity® software is copyright © 1999-2017 Audacity Team. The name Audacity® is a registered trademark of Dominic Mazzoni."

[12] D. Bogdanov, N. Wack, E. Gómez, S. Gulati, P. Herrera, O. Mayor, G. Roma, J. Salamon, J. R. Zapata, X. Serra, *et al.*, "Essentia: An audio analysis library for music information retrieval.," in *ISMIR*, pp. 493–498, 2013.

[13] C. Duxbury, M. Sandler, and M. Davies, "A hybrid approach to musical note onset detection," in *Proc. Digital Audio Effects Conf.(DAFX,02)*, pp. 33–38, 2002.

[14] J. Ricard, "An implementation of multi-band onset detection," *integration*, vol. 1, no. 2, p. 10, 2005.

[15] M. Marolt, A. Kavcic, and M. Privosnik, "Neural networks for note onset detection in piano music," in *Proceedings of the 2002 International Computer Music Conference*, 2002.

[16] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 6, pp. 3089–3092, IEEE, 1999.

[17] G. P. Nava, H. Tanaka, and I. Ide, "A convolutional-kernel based approach for note onset detection in piano-solo audio signals," in *Int. Symp. Musical Acoust. ISMA*, pp. 289–292, 2004.

[18] I. T. Matthew Davies, "IEEE Signal Processing Cup Beat Evaluator," 2016. https://piazza.com/ieee_sps/other/sp1701/resources, accessed 21-01-2018.

[19] F. Eyben, S. Böck, B. Schuller, and A. Graves, "Universal onset detection with bidirectional long-short term memory neural networks," in *Proc. 11th Intern. Soc. for Music Information Retrieval Conference, ISMIR, Utrecht, The Netherlands*, pp. 589–594, 2010.

[20] H. Wen, "Onset detection for piano music transcription based on neural networks,"

[21] G. E. Poliner and D. P. Ellis, "A discriminative model for polyphonic piano transcription," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. 1, pp. 154–154, 2007.

[22] P. Grosche and M. Müller, "A mid-level representation for capturing dominant tempo and pulse information in music recordings.," in *ISMIR*, pp. 189–194, 2009.

[23] 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.