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

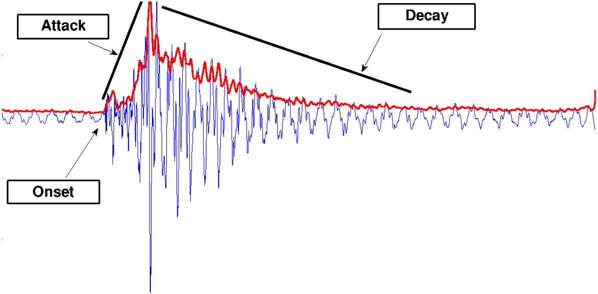
Krishna Subramani, Srivatsan Sridhar, Rohit M A, Preeti Rao

Electrical Engineering
Indian Institute of Technology Bombay, India

National Conference on Communications 2018

What is **Onset Detection**?

- Onset detection refers to the estimation of the timing of events in a music signal



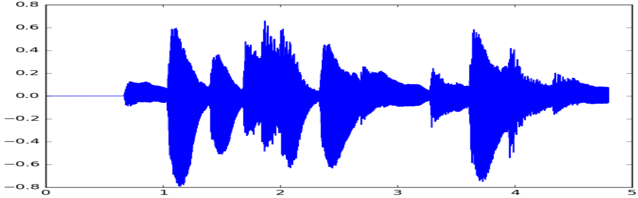
Envelope of a Musical Note[1]

- Depending on the musical instrument, onset detection poses distinct challenges

[1] http://lantana.tenet.res.in/music/stroke/on_de.png

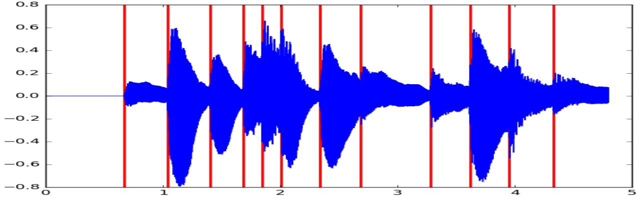
Examples

A simple example $\boxed{1}$



Examples

A simple example $\boxed{1}$

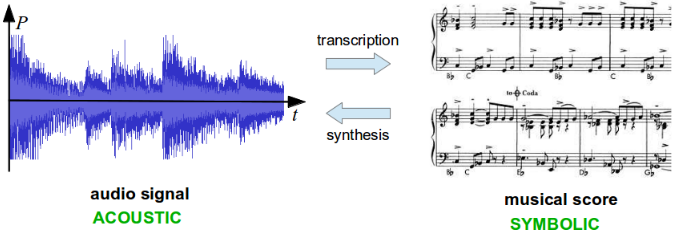


Challenges

1. Soft notes being shadowed by previous loud notes that have not decayed entirely $\boxed{1}$
2. Possible asynchrony between the individual notes played in a chord $\boxed{2}$
3. Notes occuring in rapid succession (fast tempo) $\boxed{3}$

Applications of Onset Detection

1. Automatic Music Transcription (AMT)



[2]

2. Music Pedagogy (Learning aids)

3. Music Recognition (Midomi, Shazam, etc.)

[2] http://www-etud.iro.umontreal.ca/ boulanni/amt.png

Literature Methods Review [1, 2, 3, 4, 5]

Energy (or Amplitude) Based

1. Analyze changes in signal's energy by calculating energy in windowed segments, and then computing energy difference, followed by peak picking

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

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

2. If successive onsets are weak in amplitude, this method will fail to detect them accurately because the energy increase is too little for such weak notes

Spectral Flux Based

1. Exploits changes in the signal's spectral distribution by calculating its Power Spectral Density (magnitude squared of its Short Time Fourier Transform)

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

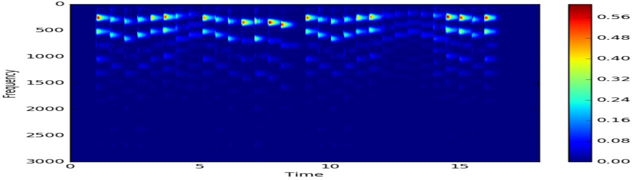
$$S_{xx}(n,k) = \left|X(n,k)\right|^2$$

2. Logarithmic Compression to emphasize high frequency transients

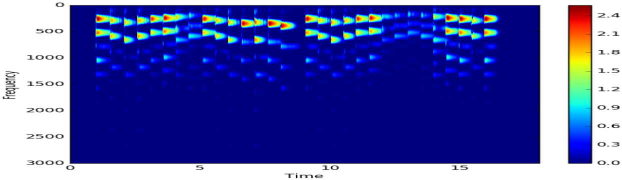
$$\gamma(S_{xx}(n,k)) := log(1 + c \cdot S_{xx}(n,k))$$

3. Spectral Flux, which is discrete derivative of the above

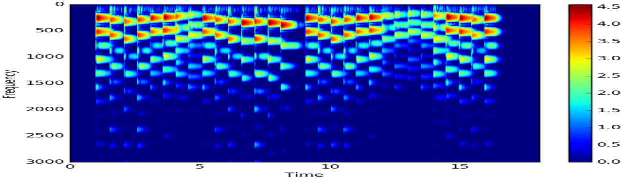
$$SF(n,k) := \left|\gamma(n+1,k) - \gamma(n,k)\right|_{\geq 0}$$



STFT



Logarithmic Compression (c=1000)



Logarithmic Compression (c=100000)

4. Finally, we add up all the frequency bins for a particular time instant, as this represents the total change in the power spectrum. The obtained array is our **novelty curve**

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

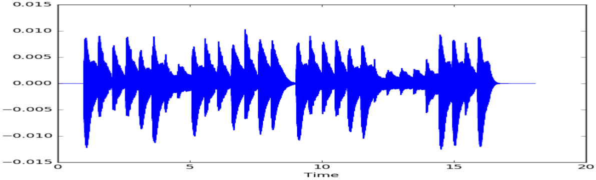
5. Spectral distribution can change considerably even for small energy changes,hence this method can pick up even relatively soft notes

4. Finally, we add up all the frequency bins for a particular time instant, as this represents the total change in the power spectrum. The obtained array is our **novelty curve**

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

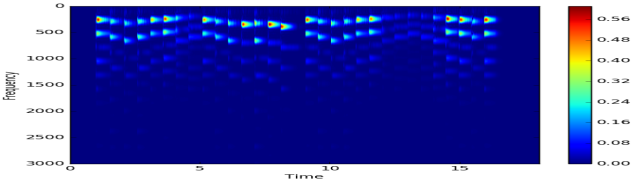
5. Spectral distribution can change considerably even for small energy changes,hence this method can pick up even relatively soft notes

- In our work, we present a modified version of the Spectral Flux based approach

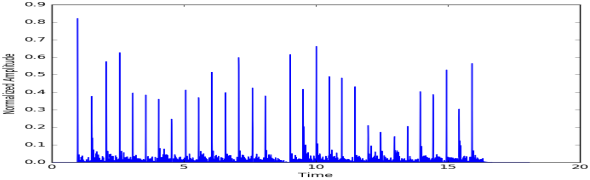


1

Audio Waveform



STFT

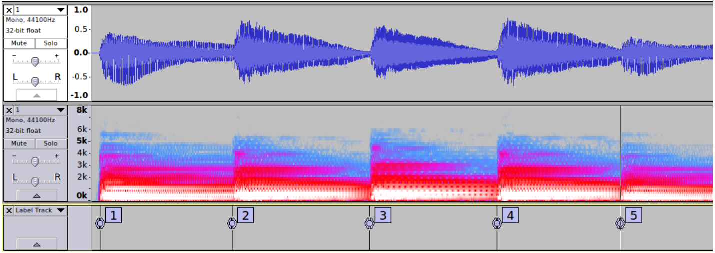


Novelty curve

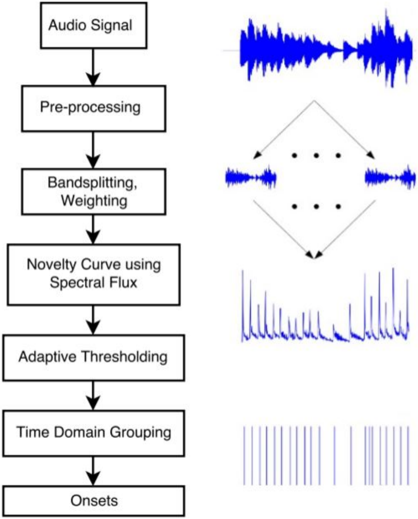
Dataset and Annotations

- 29 Piano pieces made available by West Valley College [6]
- The songs are between 20 and 60 seconds long, with the average duration being 34 seconds. The 29 pieces together contain 1934 note onsets
- Simple, medium-paced single-hand pieces to slightly expressive fast-paced pieces with dynamics and chords (sometimes with asynchrony)
- Onsets were manually marked on Audacity [7] by:
  1. Observing the spectrogram for changes
  2. Slowing down and listening to the audio

Annotation Process (using Audacity)



Proposed System



Flow of our Proposed System

Pre-Processing the Audio Signal

1. Low Pass Filtering (Cutoff = 6kHz), and re-sampling to 16kHz to remove high frequency noise and reduce computation time and memory
2. Normalization by one of the two following methods:
   2.1 Divide by the signal's maximum amplitude
   2.2 Find the window with the maximum average energy and divide throughout by this window's energy

Pre-Processing the Audio Signal

1. Low Pass Filtering (Cutoff = 6kHz), and re-sampling to 16kHz to remove high frequency noise and reduce computation time and memory

2. Normalization by one of the two following methods:
   2.1 Divide by the signal's maximum amplitude
   2.2 Find the window with the maximum average energy and divide throughout by this window's energy

- Both methods were tried, and method 2 detected more number of onsets

Band-Splitting and Weighting

- The filtered and normalized audio is split into 6 frequency bands which go from 0-6400Hz. This allows separate analysis for each frequency band
- The bands used are 0-200Hz, 200-400Hz ,400-800Hz, and so on. The octave separation between bands supports the logarithmic perception of frequencies
- 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)

$$NC(n) := \sum_{i=1}^{6} w_i \cdot NC_i(n) \qquad w_i := \frac{E_i}{\sum_{i=1}^{6} E_i}$$

Adaptive Thresholding

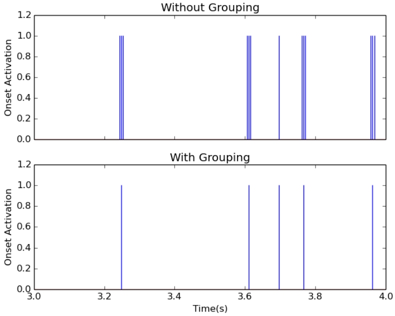
- Fixed threhsolding failed to detect soft onsets occurring immediately after a loud note
- This is because of the spectral change arising from the soft onset being over-shadowed by the strong and extended decay of the loud note strike
- This motivated us to relax the threshold for a few frames immediately after the frame containing a strong onset
- The variable threshold function, $t(n)$, a function of frame number n is defined as:

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

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

Time Domain Grouping

- Multiple onsets were detected at points where only one onset was expected
- We replaced multiple closely-spaced onsets caused due to one primary onset, with a single onset



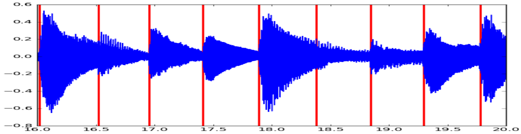
Results

- 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 and adaptive thresholding (constant threshold)

Algorithm	Precision	Recall	F-Measure
Benchmark SF	98.42	85.03	91.24
Constant Threshold	96.90	94.00	95.43
Adaptive Threshold	**97.52**	**96.62**	**97.07**

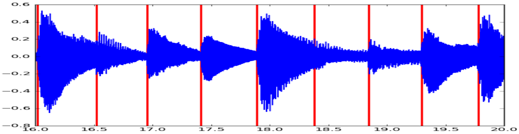
Improvement due to Bandsplitting

Ground truth annotation: $\boxed{1}$

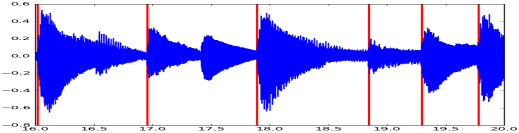


Improvement due to Bandsplitting

Ground truth annotation: $\boxed{1}$

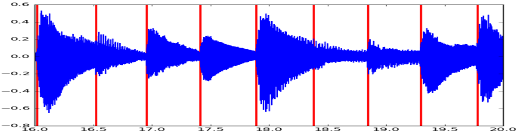


Benchmark spectral flux method: (Recall=78.18%)

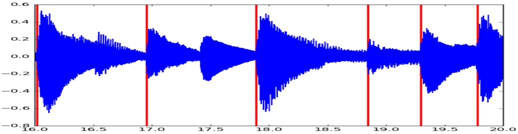


Improvement due to Bandsplitting

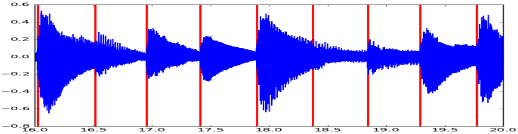
Ground truth annotation: $\boxed{1}$



Benchmark spectral flux method: (Recall=78.18%)

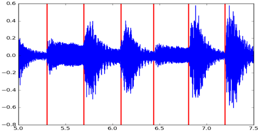


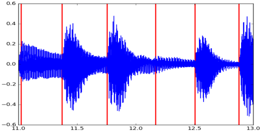
Bandsplitting, constant threshold: (Recall=96.36%)



Improvement due to Variable Thresholding

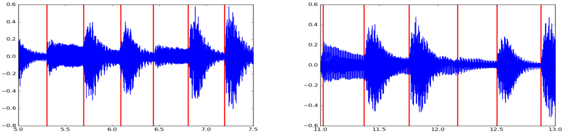
Ground truth annotation: 1 2



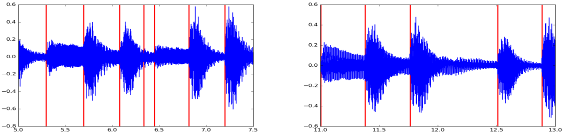


Improvement due to Variable Thresholding

Ground truth annotation: 1 2

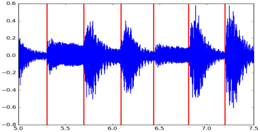


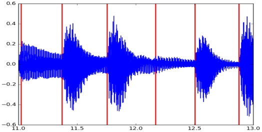
Constant Threshold (Precision=95.65%, Recall=95.65%):



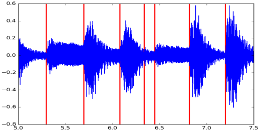
Improvement due to Variable Thresholding

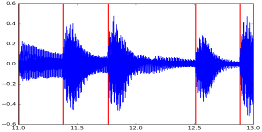
Ground truth annotation: 1 2



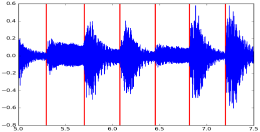


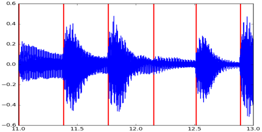
Constant Threshold (Precision=95.65%, Recall=95.65%):





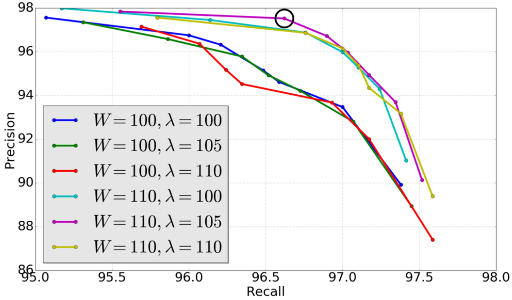
Variable Threshold: (Precision=100%, Recall=100%)





Choosing the Parameter Values for Adaptive Thresholding

- The parameters were chosen by plotting the precision and recall values for different parameter values

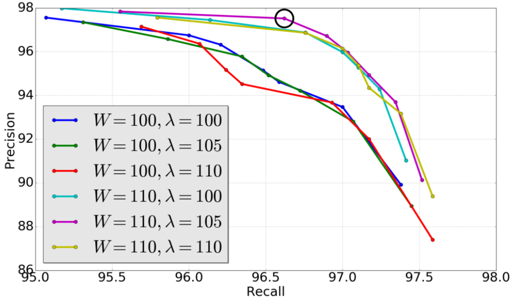


c, λ, W, h in the adaptive thresholding algorithm

$h = 1$ and $c = 0.08$ to 0.12 for each curve

Choosing the Parameter Values for Adaptive Thresholding

- The parameters were chosen by plotting the precision and recall values for different parameter values



c, λ, W, h in the adaptive thresholding algorithm

$h = 1$ and $c = 0.08$ to 0.12 for each curve

- The true capabilites of our method can be realized when the parameters for the model are learnt with an appropriate learning model

Conclusion and Future Work

The main distinctive features of our proposed system are:

1. Energy-weighted band splitting of the novelty curve
2. Adaptive thresholding
3. Grouping of multiple onsets

Conclusion and Future Work

The main distinctive features of our proposed system are:

1. Energy-weighted band splitting of the novelty curve
2. Adaptive thresholding
3. Grouping of multiple onsets

Further work to include:

1. Trying out the proposed method on more complex music from professional performances 1
2. Using Recurrent Neural Networks (Bidirectional LSTM's) or SVM based approaches to learn the parameters [8, 9, 10]
3. Extracting beat and tempo information from the music using the obtained onsets [11, 12]

References I

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

[2] 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.

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

[4] 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.

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

[6] "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.

References II

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

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.