Score-Aligned Polyphonic Microtiming Estimation

Xueyang Wang [1], Ryan Stables [1], Bochen Li [2] and Zhiyao Duan [2]

[1]DMTLab, Birmingham City University
[2]Audio Information Research Lab, University of Rochester

1: Abstract

Accurate estimation of note onset timing is important for music ensemble performance analysis and synthesis. In this study, we present a method for the detection of onsets from polyphonic mixtures, using score information. First, a MIDI score is aligned to the audio signal using dynamic time warping, and pitches of performed notes are refined using a multi-pitch estimation technique. Notes in a signal are then isolated using a spectral masking method, based on the average harmonic structure learned from each source. Onset timing is finally estimated by maximizing the time derivative of the energy curve of the note within an observation window. We show that this method significantly improves the onset timing estimation accuracy, measured by both the align rate and onset time deviation, and outperforms a state-of-art reference method.

2: Proposed Model

- **Synchronize the MIDI score**
  A DTW-based offline audio-score alignment algorithm
- **Refine the pitch**
  A score-informed multi-pitch estimation algorithm
- **Estimate microtimings**
  Analyze the polyphonic audio signal around the score-aligned positions

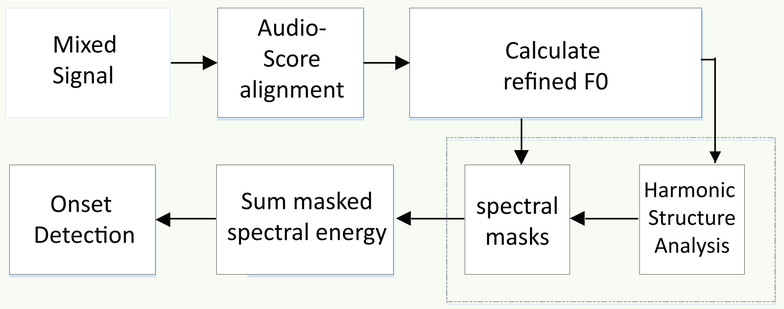

Figure 1: Illustration of the proposed model for microtiming estimation on polyphonic mixtures.

3: Harmonic Mask

Implement two types of mask:

- **H** - harmonic amplitudes: roll off at 12dB per octave (h^{th} harmonic $\rightarrow 1/h^2$).
- **AHS** - harmonic amplitudes: mean across all frames for each partial.

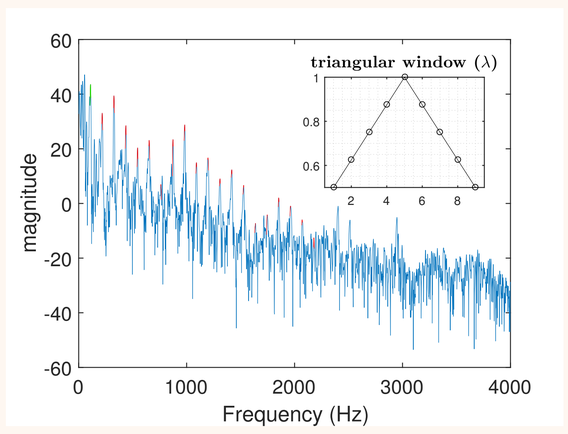

Figure 2: Calculation of a harmonic mask using triangular windows around harmonic peaks.

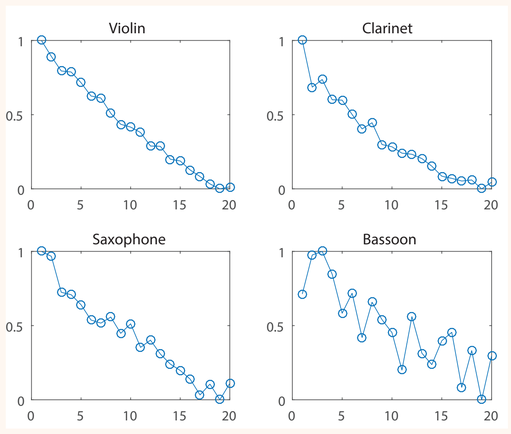

Figure 3: An example of AHS measurements from the Bach10 dataset.

4: Microtiming Approximation

- **Apply harmonic mask to the STFT** at an observation window $\mathbf{\Omega}_{k,n}$

  $k \rightarrow$ instrument index
  $n \rightarrow$ note onset position from score alignment result
  window centered at n, span to smallest inter-note interval M_k

- **Generate energy envelope**

$$E_{k,n} = H_{k,n}\mathbf{\Omega}_{k,n} \tag{1}$$

  $H_{k,n} \rightarrow$ harmonic mask (size: $1 \times N_{FFT}$)
  $\mathbf{\Omega}_{k,n} \rightarrow$ a block of STFT frames centered at n (size: $N_{FFT} \times M_k$)
  $E_{k,n} \rightarrow$ energy envelope for k-th instrument, n-th note (size: $1 \times M_k$)

- **Take the time derivative of the envelope**, and output the frame index with the highest value as estimated microtiming.

$$x_{k,n} = \max(E'_{k,n}) \tag{2}$$

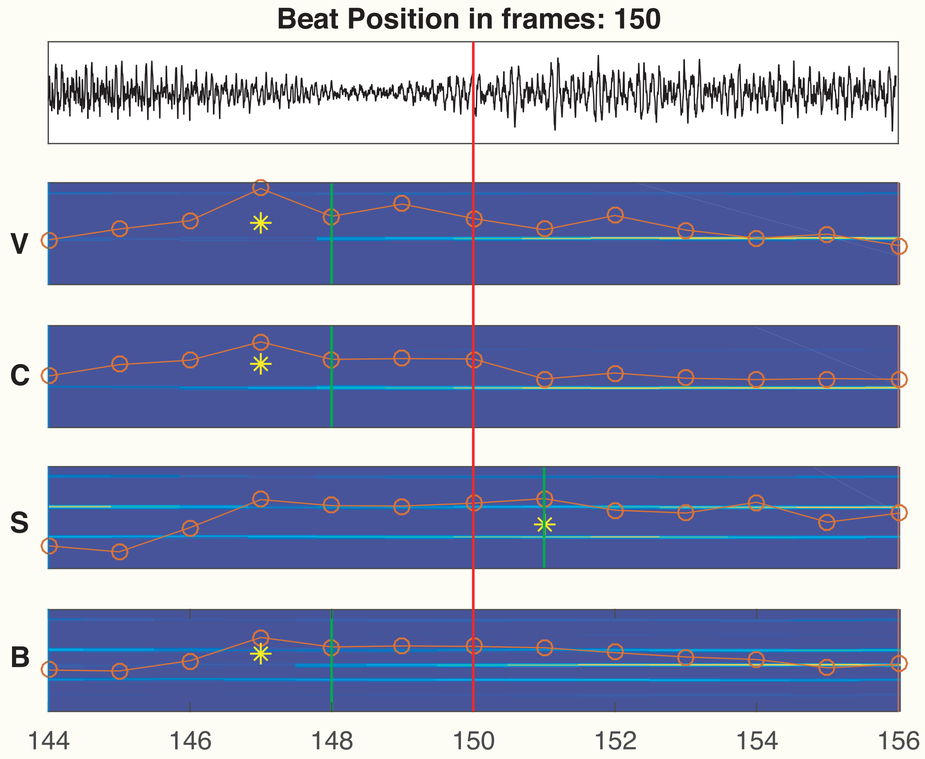

Figure 4: Onset refinement applied to concurrent events performed by four instruments in the Bach10 dataset [1]. Here, onsets are extracted from concurrent notes being performed by a violin (V), clarinet (C), saxophone (S) and bassoon (B). The background illustrates the masked STFT, the red vertical line represents the note-group location predicted by the score-alignment algorithm, the green vertical lines are ground truth onset annotations, and the yellow asterisks are predicted onset locations (note position - microtiming approximation).

5: Results

- Manually annotate onset timings of each track for evaluation
- Evaluation measures:

  - **Align Rate** $\rightarrow$ proportion of correctly aligned notes varying the tolerance thresholds
  - **Mean Timing Error** $\rightarrow$ absolute difference between aligned and ground-truth note positions

- Compare the proposed methods: **H** and **AHS**, with reference methods: **SA** (score alignment without microtiming analysis) and **Ref** (Miron et al. [2])

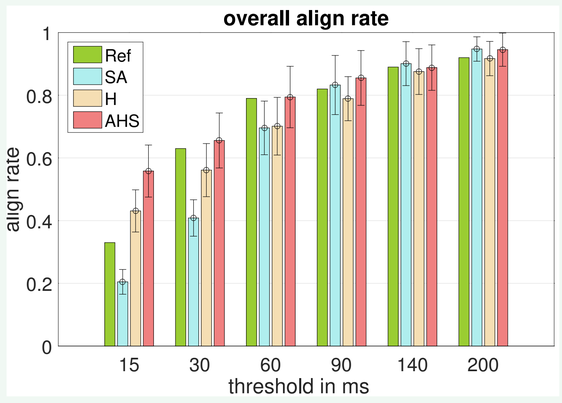

Figure 5: Align rate measured over a range of error thresholds for different methods.

- AHS exhibits best performance in most cases.
- AHS consistently outperforms other methods when the threshold is $\leq$ 90ms.
- AHS consistently has the lowest error.
- both H and AHS exhibit a significant improvement from the SA method ($p < 0.05$).

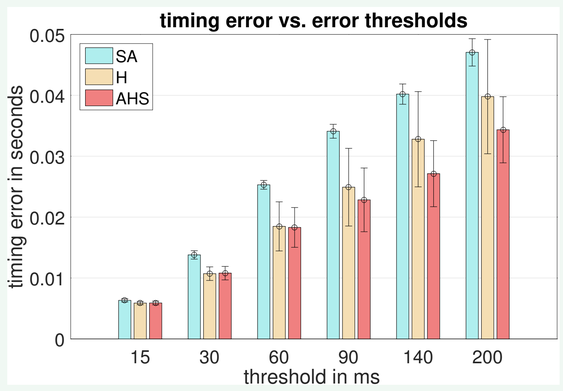

Figure 6: Mean timing error measured over a range of error thresholds for different methods.

6: Conclusion

In this study, we have presented a model for the estimation of accurate onset locations in polyphonic music mixtures using a DTW-based score-alignment method, with a harmonic spectral masking technique. We evaluated two methods for constructing the harmonic mask, one using an decaying harmonic series (H), and one based on the average harmonic structure of the instrument (AHS). When evaluated the models on a dataset of Bach chorales. Results showed that the AHS method for constructing a spectral mask improves both the alignment rate and the timing estimation of the score alignment algorithm significantly, and outperforms a state-of-art reference method.

References

[1] Z. Duan, B. Pardo, and C. Zhang, "Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2121–2133, 2010.

[2] M. Miron, J. J. Carabias-Orti, and J. Janer, "Audio-to-score alignment at the note level for orchestral recordings." in *ISMIR*, 2014, pp. 125–130.