SCORE-ALIGNED POLYPHONIC MICROTIMING ESTIMATION

Xueyang Wang, Ryan Stables

Digital Media Technology Lab
Birmingham City University
xueyang.wang@bcu.ac.uk
ryan.stables@bcu.ac.uk

*Bochen Li, Zhiyao Duan**

Dept. of Electrical and Computer Engineering
University of Rochester, NY 14627, USA
bli23@ur.rochester.edu
zhiyao.daun@rochester.edu

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.

Index Terms— Microtiming, onset detection, score alignment, multi-pitch estimation

1. INTRODUCTION

In musical timing research, a number of tasks ranging from ensemble performance analysis [1] to the synthesis and modelling of musical groups [2] rely on high-resolution onset annotations, captured from multiple musicians performing simultaneously. In most cases, these annotations are captured empirically from isolated instrument recordings, as to reduce bleed from other sound sources. This means that datasets are limited and capturing this information is often time-consuming. Accurate timing information from large corpuses of musical performance data (e.g. music archives and digital libraries) would therefore be beneficial as it would reduce the requirements on data collection and annotation. Extracting this information is currently unrealistic as group performances are generally mixed-down into a single channel recording and no annotations are provided.

Identifying timing parameters from polyphonic recordings is particularly challenging due to interference from other sources in the mix. Onset detection functions [3] often perform well in isolation, but typically rely on clear changes in spectral or temporal energy [4, 5]. For polyphonic recordings, beat tracking [6] systems are able to accurately locate beat-markers at groups of concurrent temporal events, and if musical notation is available, score-alignment systems [7] are able to map between symbolic and signal-level representations of musical signals.

Audio-score alignment has been an active research topic for decades, in which early approaches were based on offline frameworks such as Dynamic Time Warping (DTW) [7, 8]. In this case, the algorithm requires access to the entire audio stream before the process starts, which limits the system's potential applications to pre-recorded music. To enable real-time applications such as live music performance following, methods based on online DTWs [9], and Hidden Markov Models (HMM) [10] were proposed. These algorithms however, are more challenging to design and typically achieve lower alignment accuracies.

Microtiming in the context of score-alignment is considered to be the asynchrony of an event with respect to an alignment point. These asynchronies have been investigated using DTW-based offline techniques [11, 12, 13, **?**] and are considered to represent the expressive characteristics of a performer. In this study, we focus on extracting accurate onset times from polyphonic recordings using score alignment to first locate note-positions, then a refinement process is applied to each event to identify the asynchrony of each instrument from the note position. Experiments on polyphonic pieces show that the proposed method significantly outperforms a state-of-the-art method. The organization of this paper is as follows: Section 2 introduces the proposed model, including the score alignment process, methods for isolating sources in a polyphonic mixture, and microtiming estimation. Section 3 presents the experimental procedure, and Sections 4 and 5 present the results and discussion respectively.

2. PROPOSED MODEL

To estimate microtiming in polyphonic mixtures, we propose a model (shown in Fig. 1), which first uses a DTW-based

*BL and ZD were partially funded by the National Science Foundation Grant No. 1741472.

offline audio-score alignment algorithm to map the note positions of a MIDI score to frames in the STFT of an audio signal. We then estimate the refined fundamental frequency f_0 of each note around the score-notated pitch using a multi-pitch estimation algorithm. Notes with the same score-notated onset are mapped to the same audio position, and we analyze the polyphonic audio signal around this position to estimate the microtimings of these notes.

To approximate the microtiming of each note n, an observation window ω_n comprising M_k frames is centered around the mapped note onset position, and a harmonic spectral mask (H_n) is constructed for each source. We calculate the energy envelope from the extracted signal across the window by summing the product of the mask and the corresponding magnitude spectrum at each frame in an STFT. The refined microtiming position is then represented by a peak in the half-wave rectified first-order derivative of the envelope, which corresponds to a significant increase in harmonic energy.

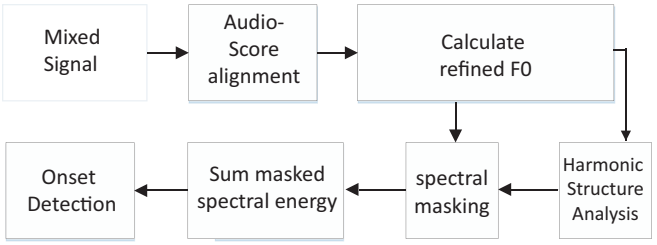

Fig. 1. Illustration of the proposed model for microtiming estimation on polyphonic mixtures.

2.1. Score Alignment Process

We adopt a commonly used offline Dynamic Time Warping (DTW) approach [7] to find the optimal alignment between audio and score, in which we use chroma features to represent the harmonic content of music. To calculate the chroma features for each audio frame, the magnitude spectrum is projected onto 12 dimensions representing the semitones in an octave. From this, a discrete audio chromagram $\mathbf{Ch}_a \in \mathbb{R}_+^{C \times L_{audio}}$ is derived, where $C = 12$ is the number of pitch classes and L_{audio} is the number of audio frames. To calculate chroma features of the symbolic score, we segment it into short time frames, and for each frame a binary 12-d vector is calculated, which indicates the presence (1) or absence (0) of a pitch. The mean length of the score frames is 0.025 beats, which is a similar scale to the audio frame hop size (10 ms), given a default tempo at 150 BPM. Similarly, a discrete score chromagram can be presented as $\mathbf{Ch}_s \in \mathbb{R}_+^{C \times L_{score}}$, where L_{score} is the number of frames in the score. We use the Euclidean distance between each pair of audio and score chroma vectors to derive a local distance matrix, then obtain a global distance matrix D. D is initialized to have $D(0,0) = 0$, $D(a,0) = D(0,b) = \inf$, for all $a = 1, \cdots, L_{score}$ and $b = 1, \cdots, L_{audio}$. It can then be iteratively calculated by:

$$D(a,b) = d(a,b) + \min \begin{cases} D(a-1,b) \\ D(a,b-1) \\ D(a-1,b-1) \end{cases} . \quad (1)$$

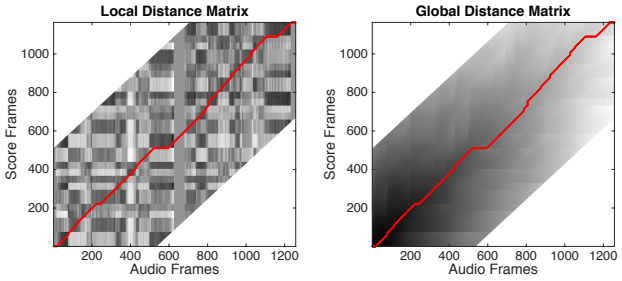

Fig. 2. The local distance matrix and the global distance matrix from the Dynamic Time Warping algorithm. The optimal alignment path is marked as the red line.

Fig. 2 illustrates the local and global distances for an excerpt from our dataset. The optimal alignment path is traced back from (L_{score}, L_{audio}) to $(1,1)$ through the global distance calculation process. To accelerate the DTW computations, we only search alignment paths within a Sakoe-Chiba band of the distance matrix, which runs along the main diagonal with a fixed width of 5 seconds.

2.2. Multi-pitch Estimation

Multi-pitch estimation is a challenging task, which can be alleviated by utilizing score information. This is known as score-informed pitch estimation or pitch refinement. A well-aligned score can provide integer MIDI pitch numbers to serve as reference pitches for each audio frame, where the main task for pitch refinement is then to estimate the pitch deviation caused by the variation in tuning or intonation (i.e., vibrato, portamento) in real performances.

We apply the multi-pitch estimation algorithm from [14]. A maximum likelihood approach based on spectral modeling of the peak and non-peak regions is used, where the power spectrum is the observation and the pitches are the parameters to be estimated. Instead of implementing the algorithm on the whole spectrum, we restrict the search space to a radius of half a semitone around the reference pitch for each note, and pitches are estimated using a greedy search strategy [10].

2.3. Harmonic Mask

For this study, we consider microtiming to be the asynchrony between the n^{th} score-aligned note group position, and the actual onset location for each source at position n. To identify this, harmonic masks comprising triangular windows around integer multiples of the refined f_0 at each of the score-aligned

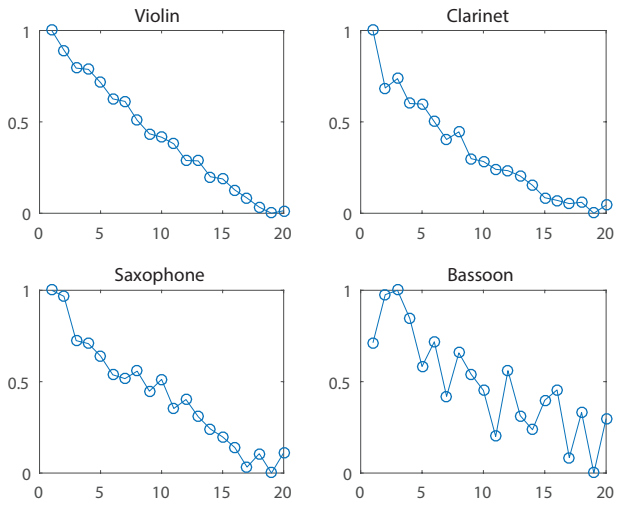

Fig. 3. An example of AHS measurements from our dataset, for four instruments in the Bach10 dataset.

note positions are constructed for each active instrument. In this study, we implement two types of mask and then evaluate the efficacy of each with regard to the timing accuracy of the model. The magnitudes of harmonics in the first mask decay such that the h^{th} harmonic has a magnitude of $1/h^2$. This is equivalent to a 12 dB/octave decay rate of harmonic energy. Each mask (H) is therefore a function of f_0, at note position n, as defined by the score alignment technique.

Due to the musical relationship of notes in the same key, pitch trajectories between instruments performing together are often correlated. This is caused by overlapping spectral components due to multiple instruments performing harmonically related pitches at note position n. With an instrument-invariant harmonic mask, the spectral energy for each frame also has the potential to be highly correlated. To mitigate the error caused by correlated masks, we construct a second mask by extracting the Average Harmonic Structure (AHS) of each instrument [15]. This is done by finding the mean of each partial in the harmonic series for a corresponding instrument, across all of the active STFT frames in a performance. We do this by weighting the spectrum with triangular windows centered around integer multiples of the refined f_0, each with a bandwidth of 40Hz. An example of AHS is shown in Fig. 3, in which the derived AHS is captured from four instruments in a string quartet.

2.4. Microtiming Approximation

To apply the harmonic mask to the STFT, an observation window Ω_n is centered around each note position, identified by the score alignment process in Section 2.1, i.e., the time span of Ω_n is $[n_t - \frac{M_k}{2}, n_t + \frac{M_k}{2}]$. The window length M_k is determined by the smallest inter-note interval of instrument k. We

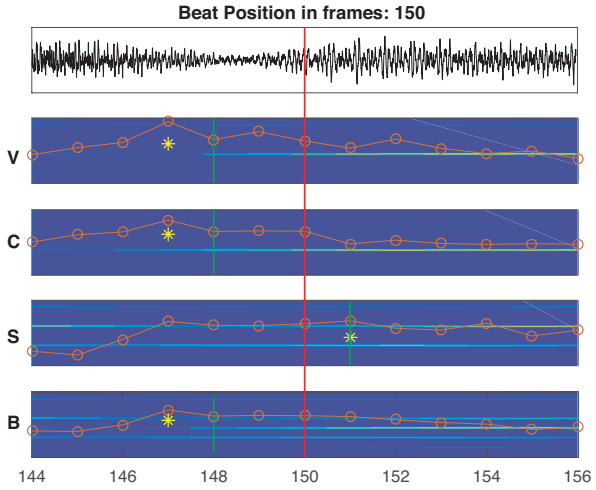

Fig. 4. Onset refinement applied to concurrent events performed by four instruments in the Bach10 dataset. Here, the background illustrates the masked STFT, the red vertical line represents the note-group location predicted by the score-alignment algorithm, the green vertical lines are ground truth onset annotations, and the yellow asterisks are predicted onset locations.

then use this to generate an energy envelope for each of the k sources in the mix, as shown in Eq. (2):

$$\mathbf{E_n} = \mathbf{H_n \Omega_n}, \tag{2}$$

where $\mathbf{H_n}$ is a ($k \times N_{FFT}$) matrix representing the harmonic mask, and $\mathbf{\Omega_n}$ is a ($N_{FFT} \times M_k$) matrix representing a block of STFT frames, centered around note n. The output $\mathbf{E_n}$ is then an M_k-length envelope for each of the k-instruments active at note position n. The microtiming of instrument k at note n is then estimated by maximizing the first derivative with regards to time of the k^{th} row of $\mathbf{E_n}$, defined as $x_{n,k} = \max(\mathbf{E'_{n,k}})$.

Fig. 4 shows this onset refinement process using four instruments in a Bach10 chorale dataset [14]. Here onsets are extracted from concurrent notes being performed by a violin (V), clarinet (C), saxophone (S) and bassoon (B). The figure shows the score-aligned note-group location, the ground truth annotations, and the refined onset locations. For each subplot, the energy envelopes $\mathbf{E'_{n,k}}$ are overlaid onto the corresponding harmonically-masked STFT block.

3. EXPERIMENT

To gauge the performance of the model, we use the Bach10 dataset, which has 10 J.S. Bach four-part chorales, all performed by quartets [14]. Each of the recordings comprise violin, clarinet, tenor saxophone and bassoon parts. The multitrack recordings were labeled with onset timing metadata, using a temporal energy-based onset detector, and corrected by

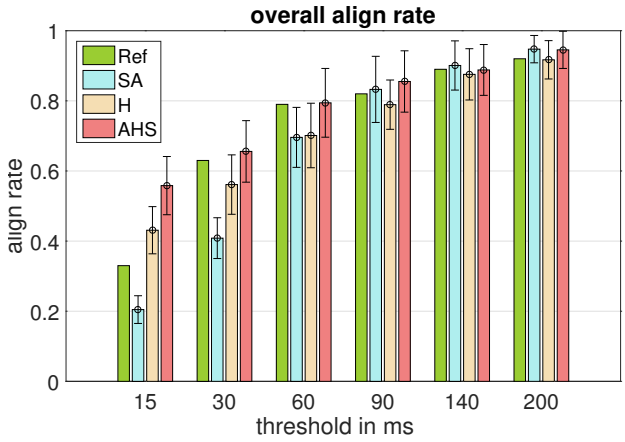

Fig. 5. Align rate measured over a range of error thresholds for two masking methods (H, AHS), the output of the score alignment process (SA), and a reference method (Ref [5]).

hand. Whilst the dataset was annotated for a previous study, we applied additional correction to the ground truth annotations due to some inaccuracies.

To evaluate the performance of our model, we use two performance measures. The *align rate* metric (as defined in [16]) is a method for quantifying the number of correctly aligned note positions to the ground-truth annotations. This is implemented by measuring the proximity of a predicted event to a ground-truth event in time. The align rate is then the portion of all events with an absolute error $|e|$ which is less than a pre-defined threshold. To test the reliability of our method, we compare it to a reference system developed by Miron et al [5], across a range of error thresholds between 15 and 200ms.

To evaluate the variability of the system, we calculate the *mean timing error*. To do this, the absolute difference between each of the refined onsets predicted by the algorithm and the corresponding ground truth is computed. The mean score is then calculated using Eq. 3.

$$e\bar{r}r = \frac{1}{J}\sum_{j=0}^{J-1}|x_j - \hat{x}_j|, \tag{3}$$

Here x_n is the ground-truth onset location at note n, and $\hat{x}_n$ is the predicted onset location.

4. RESULTS & DISCUSSION

In Fig. 5, we present the results of the align rate measurements, for thresholds of 15, 30, 60, 90, 140, and 200ms. Here, the results are averaged across all instruments and recordings in the dataset, and compared against the raw output of the score alignment process (SA), and a state-of-art reference method [5] (Ref). The results show that the AHS method performs the highest in the majority of cases, with over 56% of onsets correctly aligned at an error threshold of 15ms, over 65% aligned at 30ms, and 95% aligned at 200ms. Results also suggests that when compared to the output of the score alignment algorithm and the $1/h^2$ harmonic mask, the AHS method works particularly well when $|e| \leq 90$ms.

The timing error (TE) of each method, illustrated in Fig. 6, is measured using the same error thresholds and averaged over all instruments and recordings. Here, all timing errors which are lower than a given threshold are presented. A reference method for timing error is not included as comparable results were not available. The results show that the AHS method consistently has the lowest error, when compared against the score alignment output and the $1/h^2$ harmonic mask. When measured at 90ms (86% correctly aligned events), the mean timing error is less than 25ms, and at 200ms (95% correctly aligned events), the mean timing error is less than 35ms. In all cases above a 15ms error threshold, both of the masking methods show significant improvement ($p < .05$) over the output of the score alignment algorithm, suggesting they successfully improve the overall accuracy of the score alignment process.

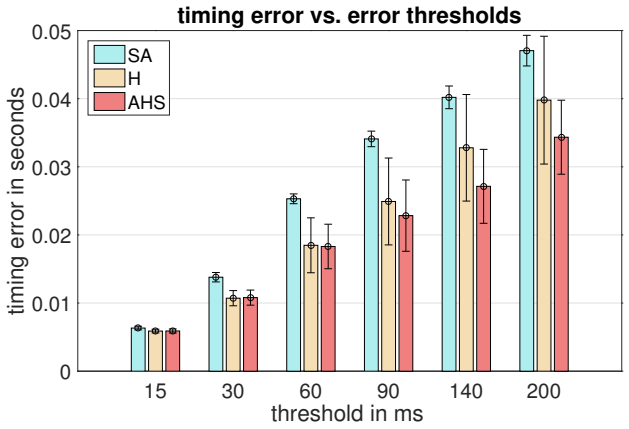

Fig. 6. Mean timing error measured over a range of error thresholds for two masking methods (H, AHS) and the output of the score alignment algorithm (SA).

5. CONCLUSION

In this study, we present a model for the estimation of accurate onset locations in polyphonic music mixtures using a DTW-based score-alignment method, with harmonic spectral masking technique. We evaluate two methods for constructing the harmonic mask, one using an decaying harmonic series, and one based on the average harmonic structure of the instrument. When evaluated on a dataset of Bach chorales, results show 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.

6. REFERENCES

[1] Alan M Wing, Satoshi Endo, Adrian Bradbury, and Dirk Vorberg, "Optimal feedback correction in string quartet synchronization," *Journal of The Royal Society Interface*, vol. 11, no. 93, pp. 20131125, 2014.

[2] Ryan Stables, Satoshi Endo, and Alan Wing, "Multiplayer microtiming humanisation using a multivariate markov model.," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2014, pp. 109–114.

[3] Juan Pablo Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark 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.

[4] Simon Dixon, "Onset detection revisited," in *Proceedings of the 9th International Conference on Digital Audio Effects*. Citeseer, 2006, vol. 120, pp. 133–137.

[5] Marius Miron, Julio José Carabias-Orti, and Jordi Janer, "Audio-to-score alignment at the note level for orchestral recordings.," in *ISMIR*, 2014, pp. 125–130.

[6] Masataka Goto, "An audio-based real-time beat tracking system for music with or without drum-sounds," *Journal of New Music Research*, vol. 30, no. 2, pp. 159–171, 2001.

[7] Nicola Orio and Diemo Schwarz, "Alignment of monophonic and polyphonic music to a score," in *Proceedings of the International Computer Music Conference (ICMC)*, 2001.

[8] Roger B Dannenberg and Christopher Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM*, vol. 49, no. 8, pp. 38–43, 2006.

[9] Simon Dixon, "Live tracking of musical performances using on-line time warping," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2005.

[10] Zhiyao Duan and Bryan Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[11] Johanna Devaney and Daniel PW Ellis, "Handling asynchrony in audio-score alignment," in *Proceedings of the International Computer Music Conference (ICMC)*, 2009.

[12] Johanna Devaney, "Estimating onset and offset asynchronies in polyphonic score-audio alignment," *Journal of New Music Research*, vol. 43, no. 3, pp. 266–275, 2014.

[13] Bernhard Niedermayer and Gerhard Widmer, "A multipass algorithm for accurate audio-to-score alignment.," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2010, pp. 417–422.

[14] Zhiyao Duan, Bryan Pardo, and Changshui 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.

[15] Zhiyao Duan, Yungang Zhang, Changshui Zhang, and Zhenwei Shi, "Unsupervised single-channel music source separation by average harmonic structure modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 766–778, 2008.

[16] Arshia Cont, Diemo Schwarz, Norbert Schnell, and Christopher Raphael, "Evaluation of real-time audio-to-score alignment," in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2007.