HANDLING ASYNCHRONY IN AUDIO-SCORE ALIGNMENT

Johanna Devaney
DDMAL, Schulich School of Music
Center for Interdisciplinary Research in Music
Media and Technology (CIRMMT)
McGill University
Montreal, QC, Canada
devaney@music.mcgill.ca

Daniel P.W. Ellis
LABROSA
Dept. of Electrical Engineering
Columbia University
New York, NY, USA
dpwe@ee.columbia.edu

ABSTRACT

Aligning a canonical score to an audio recording of a musical performance can provide very good information about the timing of individual notes. However, a score representation frequently treats multiple note events as simultaneous, whereas in reality different performers will start notes at slightly differing times, and these timing details may be significant in the analysis of performance and expression. Using an example of a four-part *a cappella* vocal piece where each voice was recorded separately, we compare note onset and offset times obtained by manual annotation to three difference types of alignment: forced alignment of each part individually to its corresponding track, simultaneous alignment of the polyphonic score to the full audio, and independent alignment of single parts to the polyphonic audio. In each case, we examine the kinds of errors that occur. We discuss how standard dynamic time warping may be extended so that it retains the advantages of polyphonic alignment while allowing ostensibly simultaneous notes to have different onset and offset times.

1. INTRODUCTION

Music alignment techniques have been a topic of interest for the music information retrieval community over the past decade. Alignment allows MIDI data, which contains the pitches and ordering events in audio data but that are fixed to a pre-specified meter and tempo, to be adjusted to match that in a recorded or live performance. Research in this area is divisible into two distinct groups: causal and non-causal. Causal systems are generally used in realtime applications, such as score following, while non-causal systems are more typically used for research applications that do not require online processing. Non-causal, offline alignment systems typically achieve greater accuracy in estimating note onsets and offsets, as the entire signal is available before the alignment is calculated.

Our interest in music alignment is to serve as a proxy for polyphonic transcription. We are interested in determining the exact timings and frequencies of the performed pitches in *a cappella* polyphonic vocal music. This task is more challenging than standard transcription, where the goal is to obtain an estimate of the pitch, but it is also more tractable in that we have the musical scores of the performed pieces.

Studies of music performance have demonstrated that there are typically asynchronies among performers for events that are notated as simultaneities in the score [1]. In order to achieve alignment accuracy for note onsets and offsets of all of the notes in the score, the asynchronies must be accounted for. This paper demonstrates the limits of using existing approaches to MIDI-Audio alignment for this application and discusses how to address the issue of these asynchronies.

2. EXISTING APPROACHES

A range of techniques have been used to address the issue of alignment, including HMMs and more generalized graphical models [2,3,4,5], sparse coding [6], support vector machines [7], and, most frequently, dynamic time warping (DTW) [8,9,10,11]. A general overview of alignment systems is available in [11].

Graphical model-based approaches have proved highly successful for realtime applications. In the Raphael's work [4], a note-level pitch-based probabilistic dynamical model represents tempo variation and note-by-note deviations. Peeling et al. [5] attempt to improve Raphael's results by training for the specific pitches and timbres present in the audio and using a 'score pointer' that is flexible enough to account for unexpected events. Preliminary evaluations undertaken for the current paper have demonstrated that existing online graphical model techniques did not perform as well as DTW-based offline techniques for *a cappella* polyphonic vocal music.

This study focuses on the utility of the DTW approach for this task, and proposes ways in which to improve it. DTW allows for the alignment of similar linear patterns, or sequences, moving at different rates, and thus is an obvious solution for the problem of temporally aligning MIDI representations to audio recordings of actual performances. Through DTW, the two sequences are warped to match each other using a cost function that accounts for the number of insertions and deletions necessary to align the sequences. The time warp can be represented visually as a path through a similarity matrix, as demonstrated in Figure 1.

In standard DTW, both the MIDI and audio files are reduced to a set of features, which are then used for alignment. As these features are generally spectral, the MIDI file must first be converted to audio, or some type of spectral-like representation, for feature extraction. The question of which features are the most appropriate for this task has been the topic of some debate in the literature. For this project we used a combination of peak structural distance, following from Orio and Schwartz [9], and cosine distance, from Turetsky and Ellis [10], as these features provided the best results for recordings of *a cappella* polyphonic vocal ensembles.

In order to account for differences between the offset of one note and the onset of the next, as may occur at phase endings and with more detached articulations, we inserted an optional silence between each of the notes.

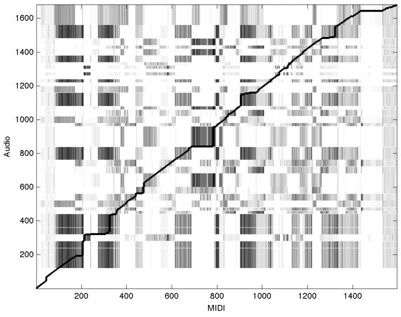

Figure 1. A similarity matrix with the DTW path indicated in black.

3. EVALUATION OF THE DTW APPROACH

3.1. Test Data

The evaluation of the standard DTW approach was performed on a hand-annotated forty-second excerpt of multi-tracked recordings of the Kyrie from Machaut's *Notre Dame Mass*. The benefit of using these multi-tracked recordings is that tests can be performed with individual as well as composite tracks. Also, the note onsets and offsets in the test set could be manually annotated to be used as ground truth relatively easily and with a high degree of accuracy. The score of this excerpt is shown in Figure 2.

3.2. Evaluation Method

Three different tests were performed on the test data: in the first we aligned each line to the monophonic recording of each part, in the second we aligned the four parts simultaneously to the polyphonic composite of the individual parts, and in the third we aligned the individual lines to the polyphonic composite. The first test allowed the DTW alignment algorithm to perform under the simplest circumstance, where all of the harmonic information in the signal was related to each note in the MIDI file. The second test presented the algorithm with more material to align, where simultaneous score events were treated as single events with a single time in the alignment. The third test evaluated whether aligning each vocal line individually allows for more accurate timing estimates for each line within a polyphonic recording.

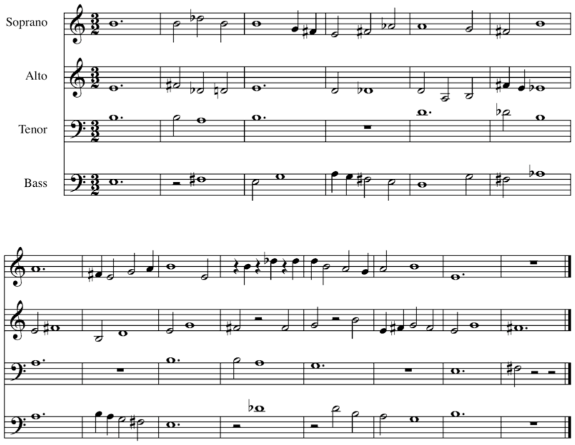

Figure 2. Musical excerpt used for testing.

Our evaluation metric looks at the note onset and offset alignment estimates against manually annotated ground truth. In order to assess the accuracy of the alignment, we considered two measures. The first tallies the number of alignments that are within 100ms of the ground truth's onsets and offsets (Table 1) and details the average amount that the alignments in each component of each test were off from the ground truth and their standard deviation (Table 2).

3.3. Results

These results demonstrate that the simultaneous alignment (Test 2) performs comparably to the individual alignment (Test 1). At times, the simultaneous alignment outperforms the individual alignment, this is due to the fact that the need to match multiple notes constrains the DTW algorithm and reduces the likelihood of it getting temporarily lost. Figures 3 and 4 show that in both tests the alignment algorithm is able to consistently find the relevant notes in the audio signal, but that the determination of the exact location of onsets and offsets is not always accurate. Also, as noted above, the simultaneous alignment cannot accurately account for the asynchronies between simultaneously performed notes because only a single time warp is created. All notes that occur simultaneously in the score are assigned the same onset and offset time. Figure 5 provides a visual example of the problem with this approach. Around 13.3 sec there is notated simultaneity between the soprano and the bass; the alignment is locked to the onset of the soprano's note, which, in performance, is about 30–40 ms behind the onset of the bass' note. Also, the offset of the tenor note occurs approximately 100ms before the other voices' offsets.

Vocal Part (# of notes)		**Test 1 Individual**	**Test 2 Composite Simultaneous**	**Test 3 Composite Individual**
Soprano (31)	***On***	7 (22%)	8 (26%)	8 (26%)
	Off	22 (71%)	21 (26%)	18 (58%)
Alto (30)	***On***	6 (20%)	10 (33%)	7 (23%)
	Off	20 (67%)	14 (70%)	17 (57%)
Tenor (14)	***On***	4 (29%)	6 (42%)	3 (21%)
	Off	7 (50%)	9 (64%)	2 (14%)
Bass (24)	***On***	5 (21%)	16 (67%)	8 (33%)
	Off	14 (58%)	15 (62%)	9 (38%)
Totals (99)	***On***	31 (31%)	40 (40%)	26 (26%)
	Off	63 (64%)	59 (60%)	46 (46%)

Table 1. The number of onsets and offsets predicted by the alignment within 100ms of the ground truth.

Vocal Part (# of notes)		**Test 1 Individual**		**Test 2 Composite Simultaneous**		**Test 3 Composite Individual**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Sop (31)	***On***	0.163	0.144	0.146	0.096	0.237	0.254
	Off	0.092	0.063	0.086	0.056	0.185	0.267
Alto (30)	***On***	0.194	0.146	0.182	0.153	0.229	0.195
	Off	0.154	0.224	0.179	0.174	0.165	0.216
Ten. (14)	***On***	0.206	0.232	0.124	0.082	1.419	1.598
	Off	0.327	0.082	0.074	0.059	1.815	1.579
Bass (24)	***On***	0.132	0.065	0.098	0.093	0.228	0.342
	Off	0.108	0.102	0.110	0.119	0.298	0.668
All	***On***	0.171	0.146	0.142	0.117	0.612	0.836
	Off	0.147	0.331	0.118	0.124	0.693	0.975

Table 2. Mean and standard deviation in seconds between the onset and offset set alignments and the ground truth.

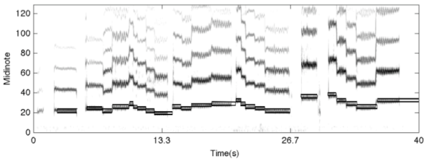

Figure 3. Test 1: Overlay of alignment of a single line aligned to a single voice.

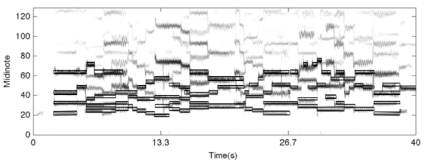

Figure 4. Test 2: Overlay of alignment for all four lines aligned simultaneously to a composite signal.

As noted above, one way of addressing the asynchrony is to align the lines one at a time against the composite signal (Test 3). Figure 6 shows the main drawback of this approach, which is that the alignment algorithm can easily become lost when aligning a single line in the presence of multiple voices. We therefore must consider extending the DTW approach for simultaneous alignment (Test 2) in order to address the issue of onset asynchrony in notated simultaneities.

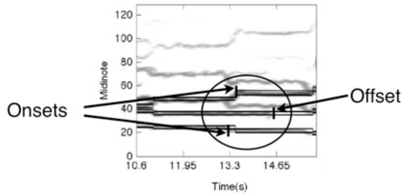

Figure 5. Test 2: An example of a performance asynchrony for a notated simultaneity.

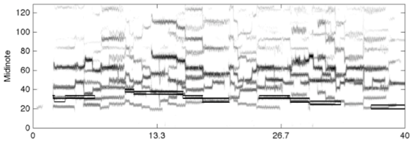

Figure 6. Test 3: Overlay of alignment for a single line aligned to a composite signal of all the voices.

4. EXTENDING THE STANDARD DTW APPROACH TO ADDRESS ASYNCHRONY

Given only polyphonic audio, we are currently faced with a compromise between on the one hand aligning the full polyphonic score, which is most likely to succeed since all the lines reinforce the overall alignment, but which is unable to identify small asynchronies that are not present in the score, and on the other hand aligning individual lines to the full audio, which allows the timing of each line to vary independently, but is highly prone to gross alignment errors due to the dense polyphonic "interference". A better approach would be to have an alignment procedure that required all lines to stay roughly aligned, but allowed (and measured) some asynchrony in events that are simultaneous in the score.

We are currently developing such a system that operates as follows: First, DTW is applied to the full polyphonic score to get a rough alignment in which notated simultaneities are forced to be simultaneous. We then refine each transition in turn by realigning just the portion of the audio in-between the centers of each note concurrence (i.e., spanning at most one transition per voice) with a system being able to accommodate asynchronous transitions. Instead of DTW, which represents only a single possible sequence, we use a more general graph of states (i.e., a hidden Markov model

(HMM)) that can include alternative paths to accommodate the alternative possible sequencing of transitions. If we have N voices, and require each voice to go through three states in sequence (initial note - transition state (e.g., silence) - final note), there are 3^N possible combinations of notes, each of which needs a separate state in our HMM graph. For N=4, this gives 81 states which, while large, is easily computed. The HMM structure allows for weighting transitions between all pairs of states; because the initial and final states are given (all voices in initial note, and all voices in final note, respectively), and because the sequence of events in each voice is constrained, the graph is actually far more sparse than the (3^N)^2 possible transitions in an unconstrained graph, and the best path can quickly be found by the analogous dynamic programming algorithm for an HMM model, the Viterbi algorithm [13]. Figure 7 shows an example of the graph of transitions for two notes (N=2, giving 9 states), and includes transitions that are added if we allow the possibility that each voice will skip the transition/silence state.

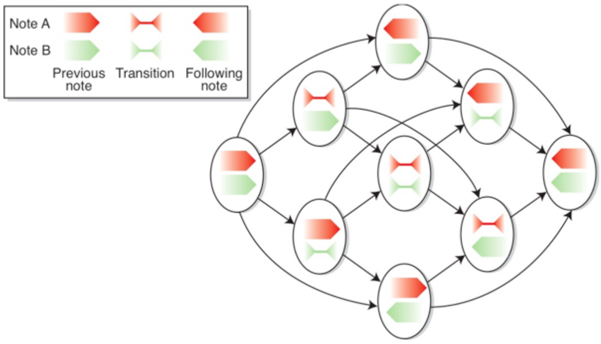

Figure 7. A graph of all of the possible transitions between two-note simultaneities.

5. CONCLUSIONS

This paper considered the problem of MIDI-Audio alignment for the particularly challenging idiom of polyphonic *a cappella* vocal music. While DTW-based approaches are the most robust in this application, they are not a complete solution. When aligning all of the voices simultaneously, they are unable to account for asynchronies in notated simultaneities, and aligning one line at a time against a polyphonic signal with this technique is not a viable option since the alignment is easily thrown off. We discussed how standard DTW-based approaches need to be extended in order to account for these asynchronies, and proposed an extension to these approaches that would to address these problems.

6. ACKNOWLEDGEMENTS

We would like to thank Chris Raphael and Paul Peeling for providing code for evaluation purposes, and Ichiro Fujinaga for his feedback at various stages of this project. This work was supported by the Center for Research in Music Media and Technology (CIRMMT) and the Social Sciences and Humanities Research Council of Canada (SSHRC).

7. REFERENCES

[1] Palmer, C. "Music Performance". *Annual Review of Psychology*. 1997, 48, pp. 115–38.

[2] Cano, P., A. Loscos, & J. Bonada. "Score-performance Matching Using HMMs", in *Proceedings of the International Computer Music Conference*, LOCATION 1999, pp. 441–4.

[3] Orio, N., & F. Déchelle. "Score Following Using Spectral Analysis and Hidden Markov Models", in *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001, pp. 151–4.

[4] Raphael, C. "A Hybrid Graphical Model for Aligning Polyphonic Audio with Musical Scores", in *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 387–94.

[5] Peeling, P., T. Cemgil, & S. Godsill. "A Probabilistic Framework for Matching Music" Representations", in *Proceedings of the International Conference on Music Information Retrieval*, Vienna, Austria, 2007, pp. 267–72.

[6] Cont, A. "Realtime Audio to Score Alignment for Polyphonic Music Instruments using Sparse Non-Negative Constraints and Hierarchical HMMs", in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, 2006, pp. 185–8

[7] Shalev-Shwartz, S., J. Keshet, & Y. Singer. "Learning to Align Polyphonic Music", in *Proceedings of the International Conference on Music Information Retrieval*, Barcelona, Spain, 2004, pp. 381–94.

[8] Hu, N., R., Dannenberg, & G. Tzanetakis. "Polyphonic Audio Matching and Alignment for Music Retrieval", in *Proceedings of the IEEE Workshop on Audio and Signal Processing to Audio and Acoustics*, New Paltz, USA, 2003, pp. 185–8.

[9] Muller, M., F. Kurth, & M. Clausen. "Audio Matching via Chroma-Based Statistical Features", in *Proceedings of the International Conference on Music Information Retrieval*, London, UK, 2005, pp. 288–95.

[10] Orio, N., & D. Schwarz "Alignment of Monophonic and Polyphonic Music to a Score", in *Proceedings of the International Computer Music Conference*, Havana, Cuba, 2001, pp.129–32.

[11] Turetsky, R., & D.P.W. Ellis. "Ground-Truth Transcriptions of Real Music from Force-Aligned MIDI Syntheses", in *Proceedings of the International Conference on Music Information Retrieval.* Baltimore, USA, 2003, pp. 135–41.

[12] Dannenberg, R., & C. Raphael. "Music Score Alignment and Computer Accompaniment". *Communications of the ACM.* 2006, 49(8), pp. 39–43.

[13] Rabiner, L.R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*. 1989, 77(2), pp. 257–89.