Handling Asynchrony in Audio-Score Alignment

Johanna Devaney
McGill University
devaney@music.mcgill.ca

Daniel P.W. Ellis
Columbia University
dpwe@ee.columbia.edu

DDMAL Distributed Digital Music Archives & Libraries Lab

Lab ROSA Laboratory for the Recognition and Organization of Speech and Audio

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology

McGill Schulich School of Music École de musique Schulich

Columbia University in the City of New York

Fonds de recherche sur la société et la culture Québec

Social Sciences and Humanities Research Council of Canada

Conseil de recherches en sciences humaines du Canada

Canada

Introduction

Description of MIDI/Audio alignment

Evaluation of Dynamic Time Warping

Future Work

Conclusions

INTRODUCTION

- MIDI-Audio alignment can be considered a solved problem for many applications
- There are typically asynchronies in musical performance for events that are notated as simultaneous in the score (Palmer 1996)
- Current methods are unable to account for these asynchronies
- How can existing approaches be extended in order to account for this?

Introduction

- Data in studies of musical performance is typically obtained through:
  - manual annotation of audio recordings
  - performances on specialized equipment

- This work is motivated by our interest in studying intonation in vocal ensembles
  - we plan to use alignment as a proxy for polyphonic transcription

MIDI/Audio Alignment

- MIDI data is adjusted to match the temporal characteristics of the audio
- Alignment can be done in real-time or offline
  - Real-time applications include score following
  - Offline applications include digital libraries and database searches
- Offline systems have the advantage of the entire signal being available before the alignment is calculated

MIDI/Audio Alignment

- A brief history...
  - ICMC - Dannenberg (1984) and Vercoe (1984)
    - Dannenberg made use of dynamic programming
  - Puckette (1995) - singing voice
  - Grubb and Dannenberg (1997) - singing voice/stochastic
  - Raphael (1999) - hidden Markov model

MIDI/Audio Alignment

- Dynamic Time Warping (DTW) and hidden Markov models (HMMs) approaches perform comparably
- We chose to use DTW as the basis of this research project
- Typical implementations of both methods produce a single time warp
  - A single time warp is problematic because it treats notated simultaneities as single events

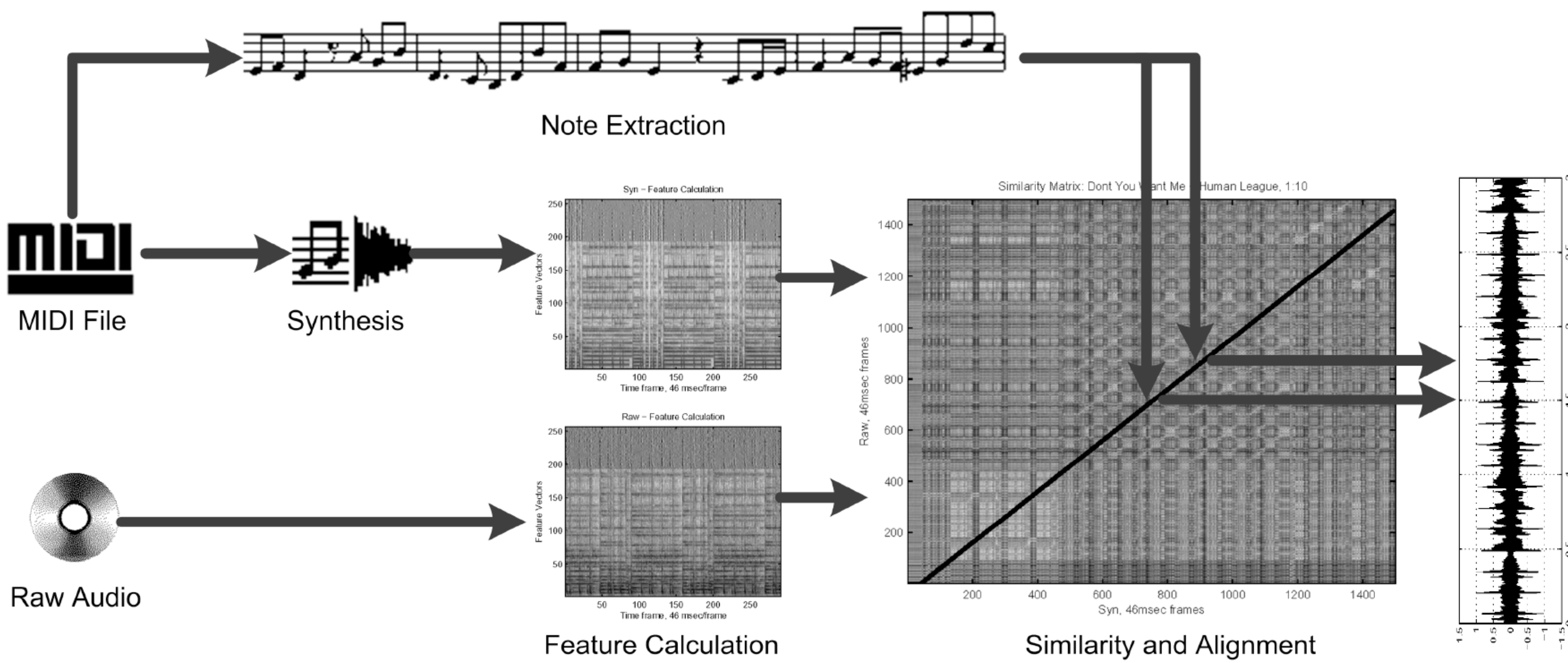
Dynamic Time Warping

- Dynamic Time Warping (DTW) is a constrained method that allows for the alignment of similar sequences moving at different rates

- First the audio and the MIDI are converted to sets of features
  - peak structure distance (Orio and Schwartz 2001)
  - chromagrams (Hu, Dannenberg, and Tzanetakis 2003)
  - cosine distance (Turetsky and Ellis 2003)

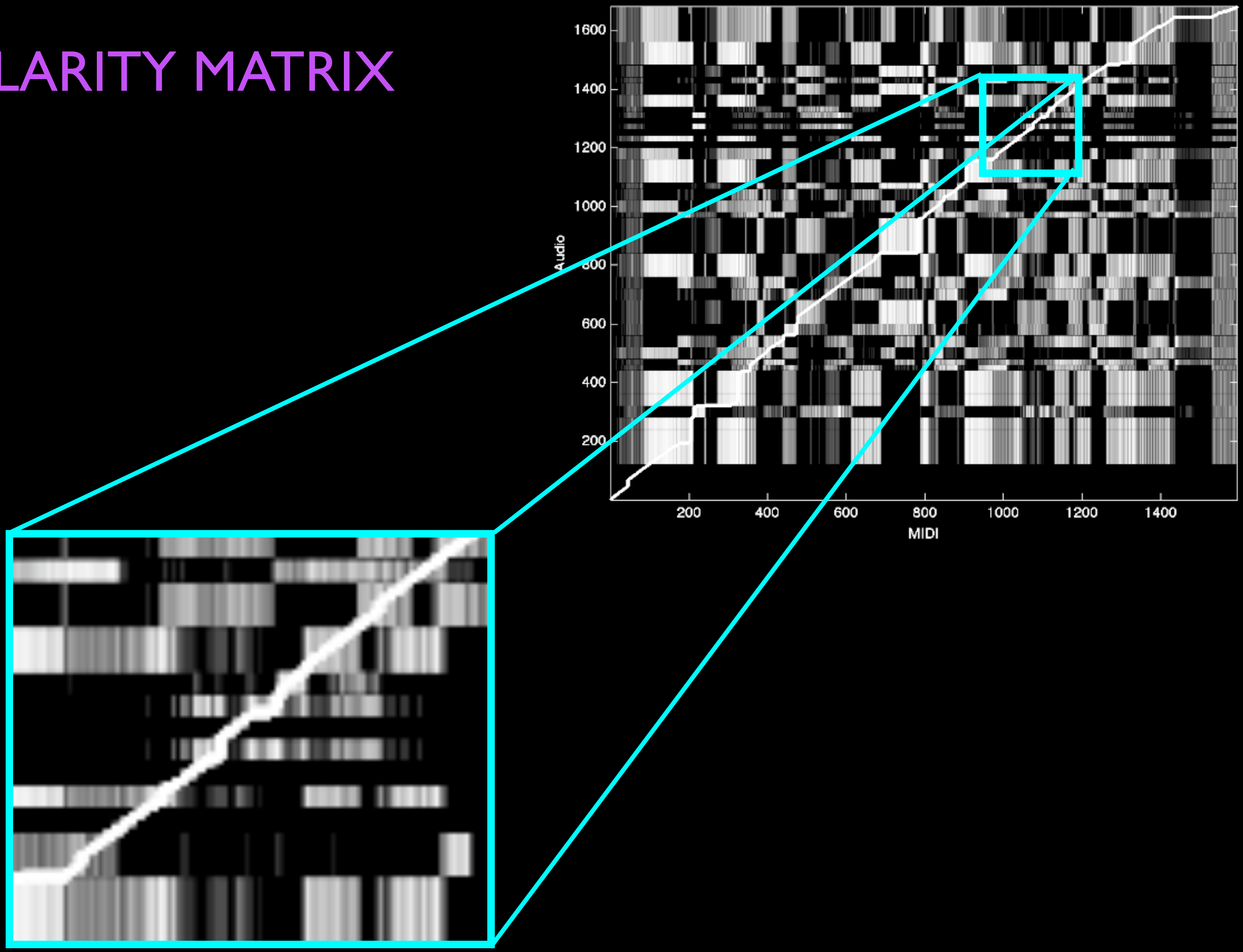
- Then the two sets of features are then compared in a similarity matrix

Dynamic Time Warping Overview



Turetsky and Ellis 2003

SIMILARITY MATRIX



Evaluation of Dynamic Time Warping

- Test data comprised of hand-annotated excerpts of four-track recordings from Machaut's *Messe de Notre Dame*

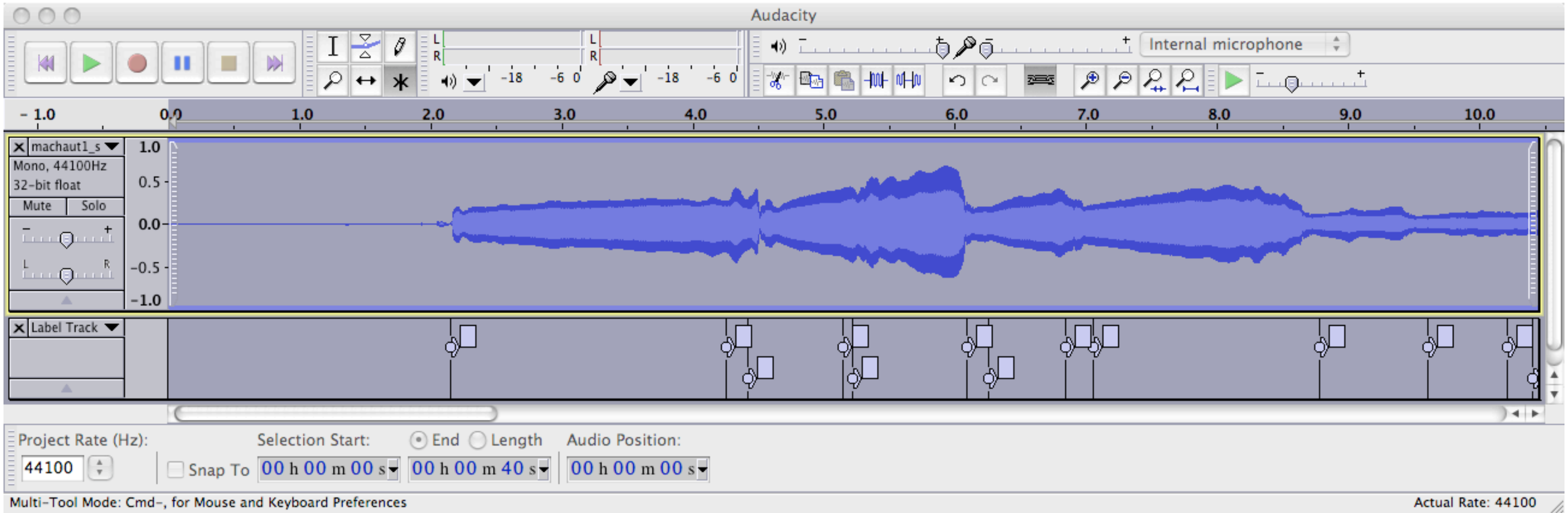
Soproano Line

Mixdown

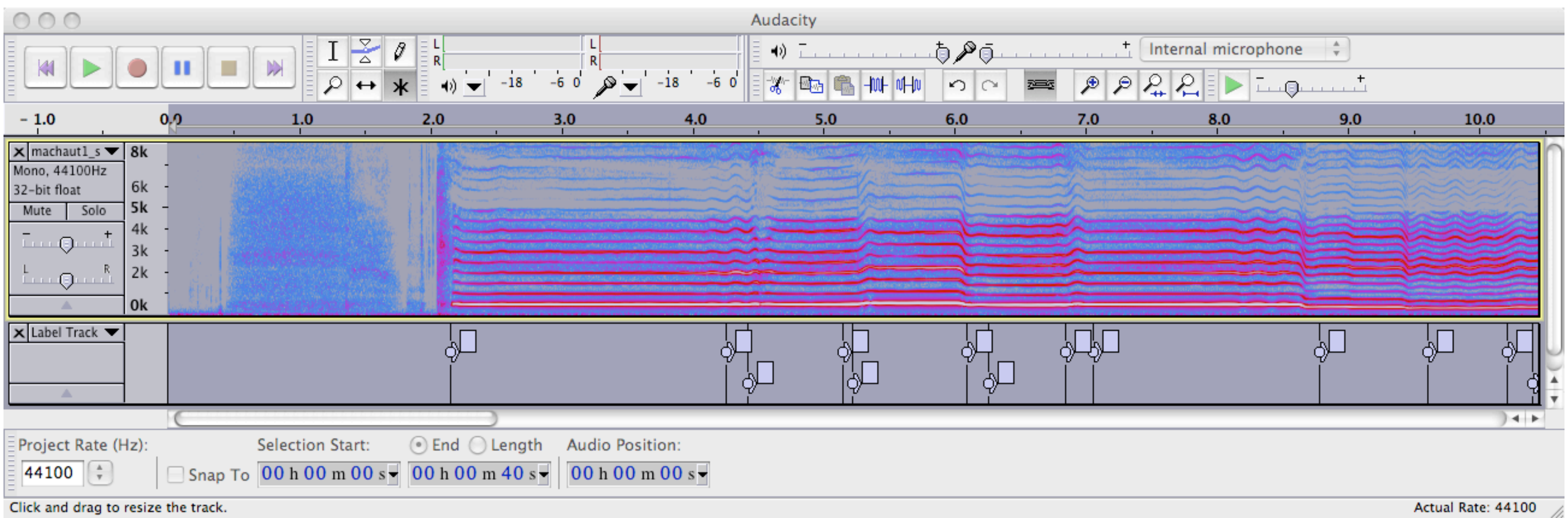
- Note onsets and offsets in the individual tracks could be manually annotated with a high degree of accuracy

- Tests were performed with individual tracks as well as a mixdown of the tracks

Time Domain Representation of Soprano in Audacity (with labels)



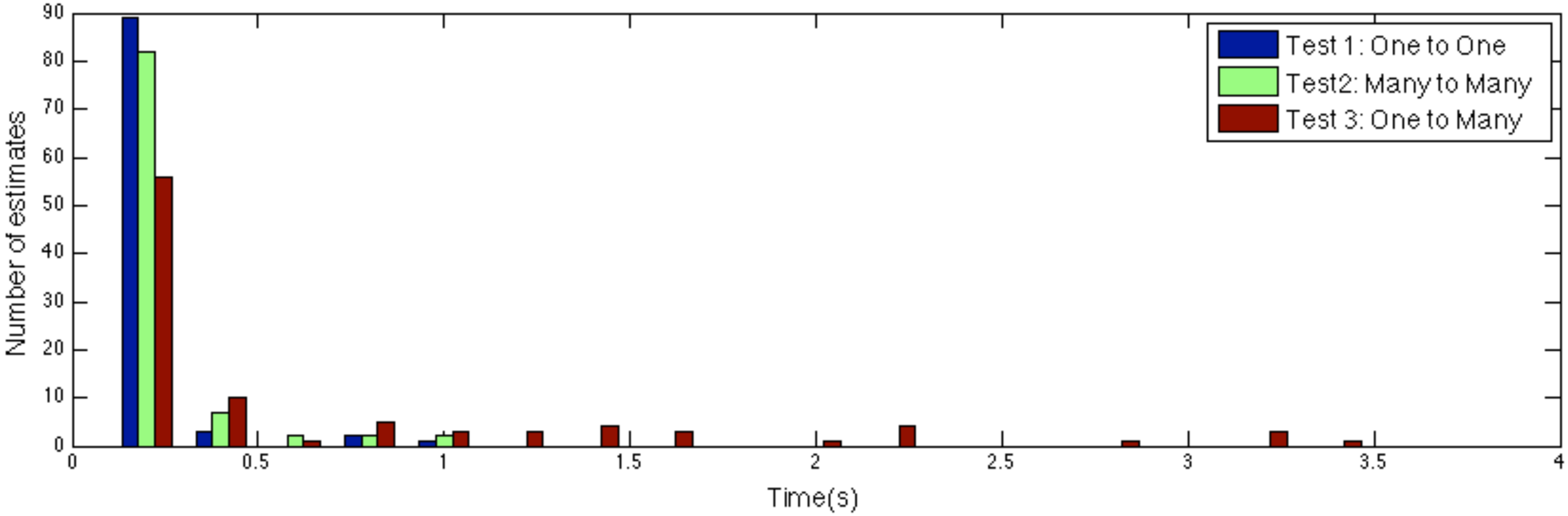
Frequency Domain Representation of Soprano in Audacity (with labels)



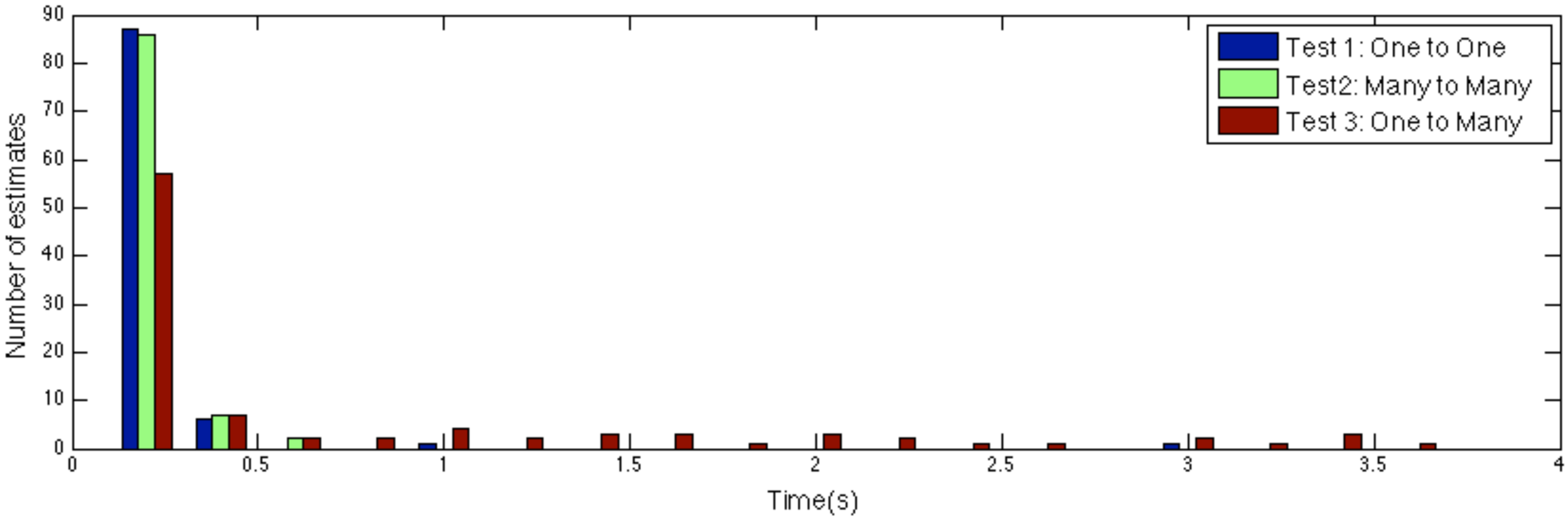
Evaluation of Dynamic Time Warping

- Three alignment tests were performed on the test data
  - One to One - each part is aligned to a recording of the corresponding individual track
  - Many to Many - the four voices are simultaneously aligned to a mixdown of the individual tracks
  - One to Many - each part is aligned to a mixdown of the individual tracks

HISTOGRAM OF ONSET ACCURACY

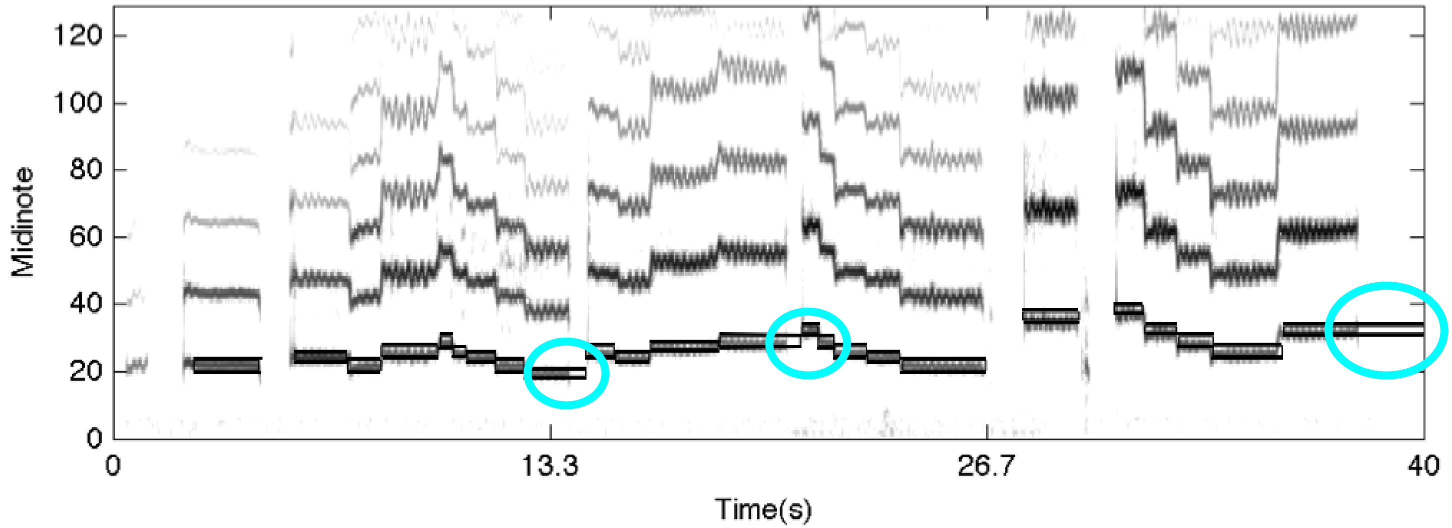


HISTOGRAM OF OFFSET ACCURACY

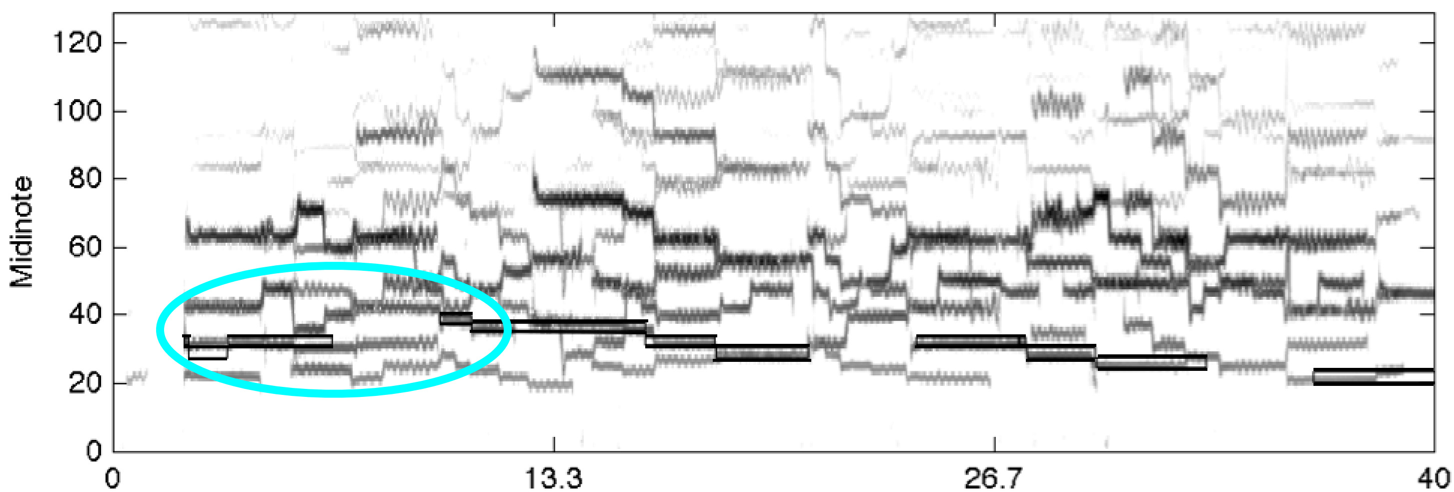


EVALUATION OF DYNAMIC TIME WARPING APPROACH

One to One

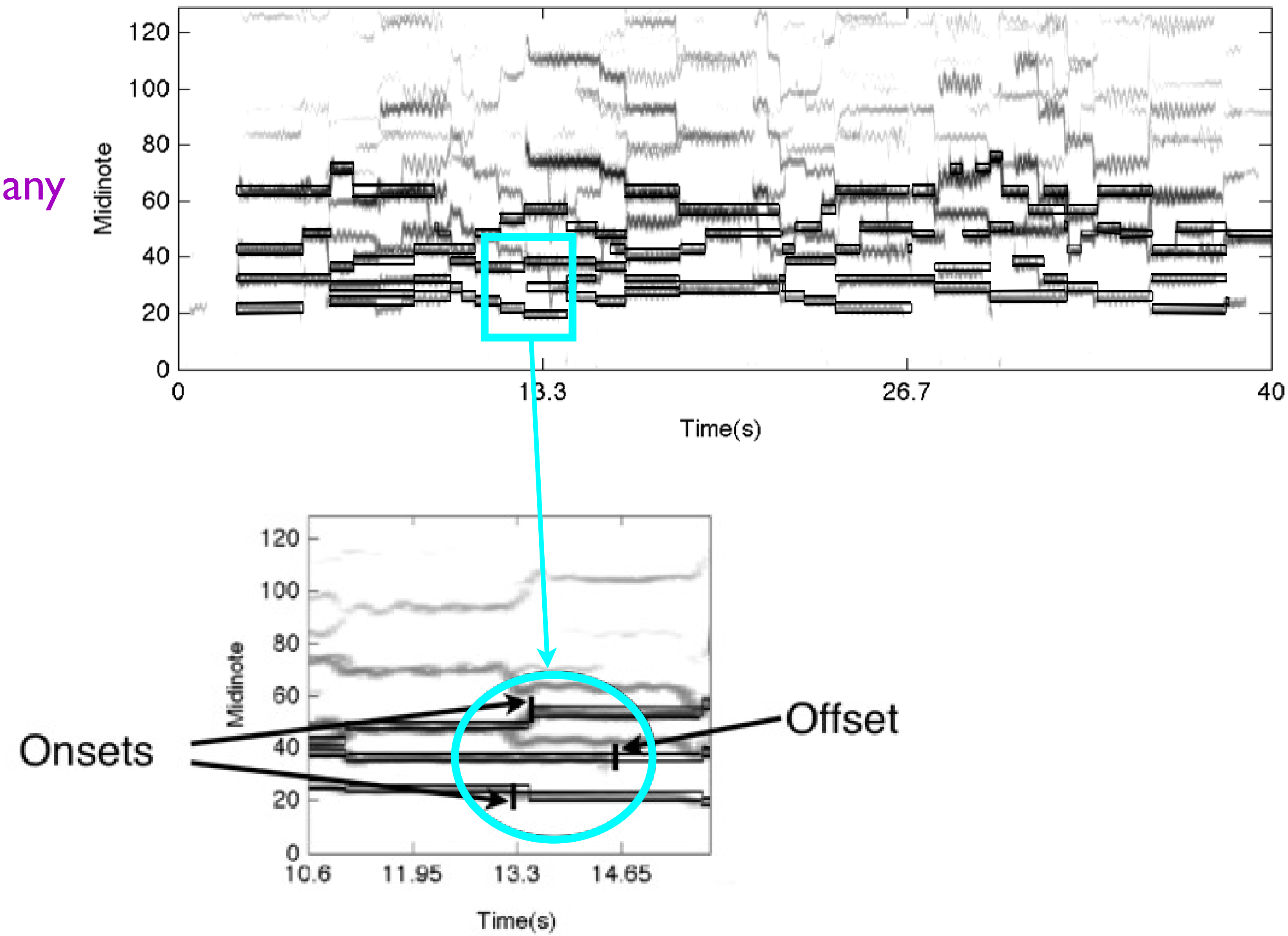


One to Many



Evaluation of Dynamic Time Warping Approach

Many to Many



Extensions to Dynamic Time Warping

- When using standard DTW on polyphonic audio there is a compromise between:
  - aligning the full polyphonic score
    - PROS: most likely to succeed
    - CONS: unable to account for asynchronies
  - aligning individual lines
    - PROS: timing of each line can vary independently
    - CONS: highly prone to errors

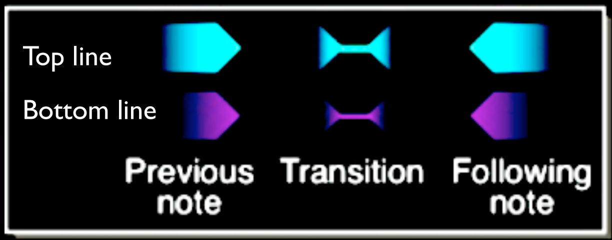
Future Work

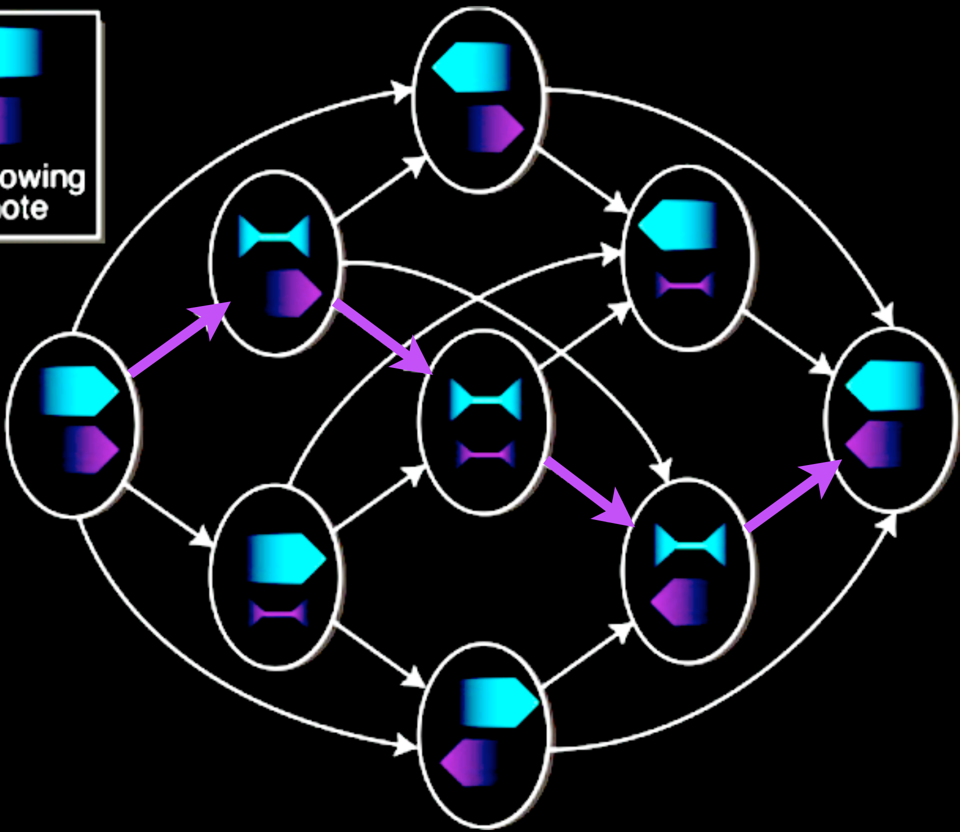
- DTW is applied to the full polyphonic score to get a rough alignment
- This is refined by realigning the portion of the audio in-between the notes
- Each note goes through a three-state sequence
  - initial note - silence - final note
- The complexity of this is 3^N, where N is the number of simultaneous notes
  - 2 voices would have 9 possible combinations
  - 3 voices would have 27 possible combinations
  - 4 voices would have 81 possible combinations

FUTURE WORK

TRANSITION MATRIX FOR TWO NOTES

- B ends (top line)
- D ends (bottom line)
- E begins (bottom line)
- C begins (top line)





Conclusions

- DTW-based approaches are the generally robust for aligning the particularly challenging idiom of polyphonic *a cappella* vocal recordings
- DTW-based approaches are unable to account for asynchronies in notated simultaneities
- Aligning one line at a time against a polyphonic signal with this technique is not a viable option
- Standard DTW-based approaches need to be extended in order to account for these asynchronies

ACKNOWLEDGEMENTS

- 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)

- We would also like to thank Ichiro Fujinaga for his feedback at various stages of this project

- We would like to thank Chris Raphael and Paul Peeling for providing code for evaluation purposes

Thank You

Questions?

References

Dannenberg, R. 1984. An on-line algorithm for real-time accompaniment. In *Proceedings of the 1984 International Computer Music Conference*. 193–8.

Grubb, L., and R. Dannenberg. 1997. A stochastic method of tracking a vocal performer. In *Proceedings of the 1997 International Computer Music Conference*. 301–8.

Hu, N., R., Dannenberg, & G. Tzanetakis. 2003. Polyphonic Audio Matching and Alignment for Music Retrieval. In *Proceedings of the IEEE Workshop on Audio and Signal Processing to Audio and Acoustics*. 185-8.

Orio, N., & D. Schwarz. 2001. Alignment of Monophonic and Polyphonic Music to a Score. In *Proceedings of the International Computer Music Conference*. 129-32.

Palmer, C. 1997. Music Performance. *Annual Review of Psychology*. 48. 115-38.

Puckette, M. 1995. Score following using the sung voice. In *Proceedings of the 1995 International Computer Music Conference*. 175–8.

Raphael, C. 1999. Automatic segmentation of acoustic musical signals using hidden Markov Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 21(4): 360–70.

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.* 135-41.

Vercoe, B. 1984. The synthetic performer in the context of live performance. In *Proceedings of the 1984 International Computer Music Conference*. 199–200.

Evaluation of Dynamic Time Warping

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

	Test 1 Individual		**Test 2** Composite Simultaneous		**Test 3** Composite Individual	
	Means	Std Dev	Means	Std Dev	Means	Std Dev
Ons	0.171	0.146	0.142	0.117	0.612	0.836
Offs	0.147	0.331	0.118	0.124	0.693	0.974

Evaluation of Dynamic Time Warping

Percentage of onsets and offsets predicted by the alignment within 100ms of the ground truth asynchrony for a notated simultaneity

	Test 1 Individual	**Test 2** Composite Simultaneous	**Test 3** Composite Individual
Ons	31%	40%	26%
Offs	64%	60%	46%