MIDI-Audio alignment for singing voice analysis using temporal models

Johanna Devaney
Assistant Professor of Music Theory and Cognition
School of Music, The Ohio State University

Introduction to Studying Musical Performance

Prior Work on MIDI-Audio Alignment

Improved Method for Monophonic Performances

Improved Method for Polyphonic Performances

Conclusions

Why study performance?

- Observing what performers actually do in performance
  - How individuals change as they gain more experience
  - How performance practice has evolved over time
- Developing models of “expressive” performance
  - For generating performances
  - For comparing individual performances to a baseline
- Providing a link between the score and the music that listeners hear

Why study the singing voice?

- In its most basic form singing is innate and universal
  - Training and enculturation refine specific practices of singing
- The voice is one of the most expressive instruments
- Singing research is complementary to speech research
- Singing is a topic of interest to psychologists

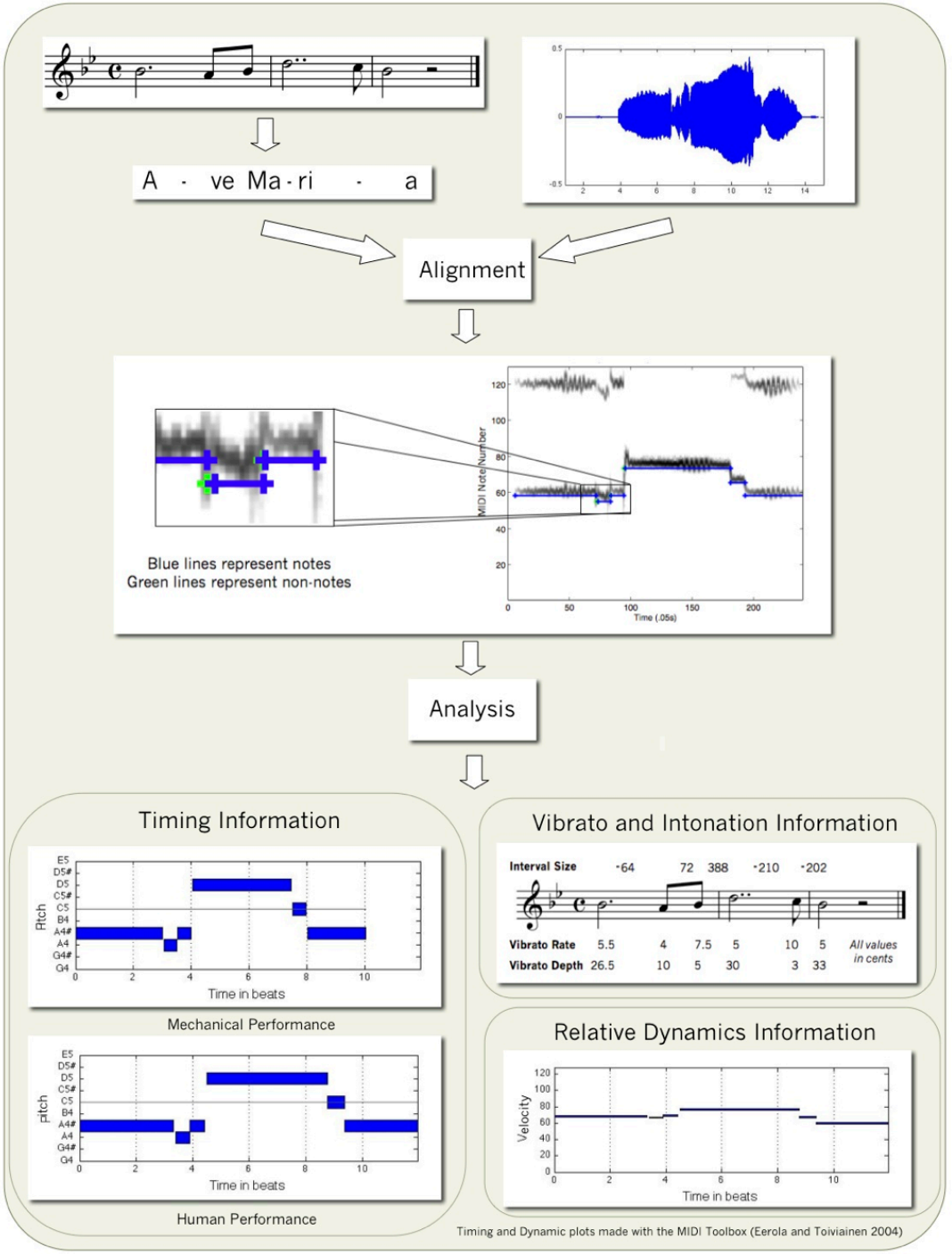
What is required to study performance?

1. Extract performance data from recordings reliably (and automatically)
2. Describe the extracted data in a perceptually meaningful way
3. Relate the collected data to the music materials

AMPACT

- Automatic Musical Performance Analysis and Comparison Toolkit (www.ampact.org and www.github.com/jcdevaney/AMPACT)
- MATLAB-based toolkit for extracting performance-related data from recordings and comparing the data across multiple performances

Devaney, Mandel, and Fujinaga (2012)

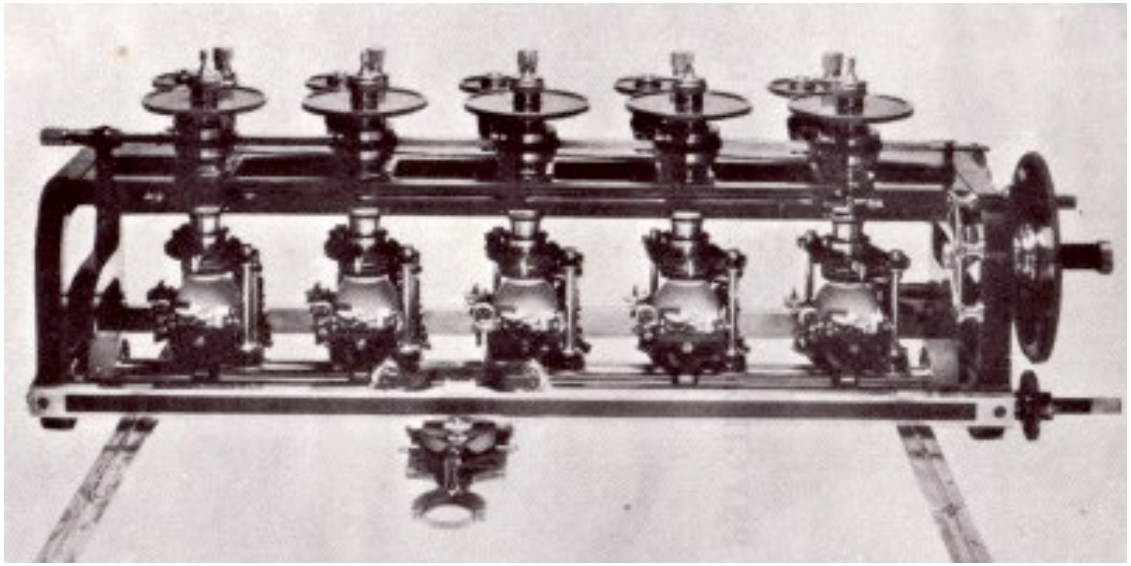


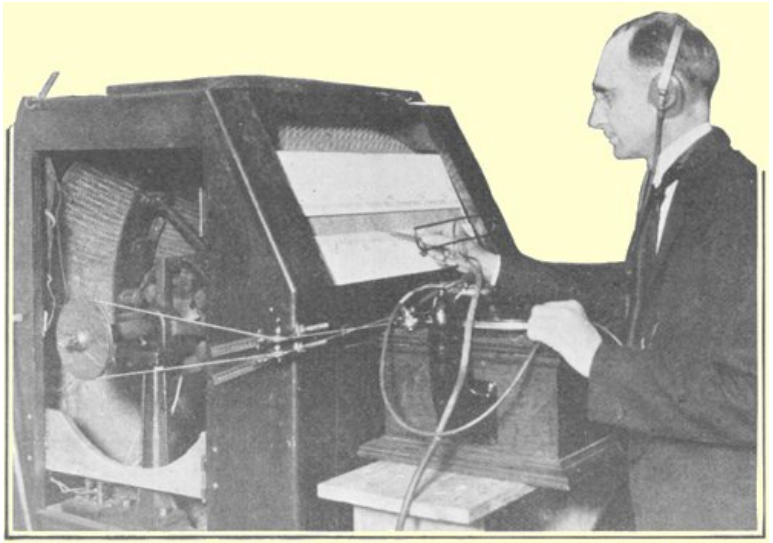
Methods for Extracting Timing Data

- Manual annotation with analogue equipment
- Use of specialized musical instruments
- Manually annotating onsets by tapping
- Manual annotation with digital software
- Audio onset detection
- MIDI-audio alignment

Manual Annotation with Analogue Equipment

- Carl Seashore (1938) and colleagues studied timing, dynamics, intonation, and vibrato in pianists, violinists, and singers
  - Equipment: piano rolls, films of the movement of hammers during performance, phono-photographic apparatus





The tonoscope for analyzing the pitch of the tones on a disk phonograph record

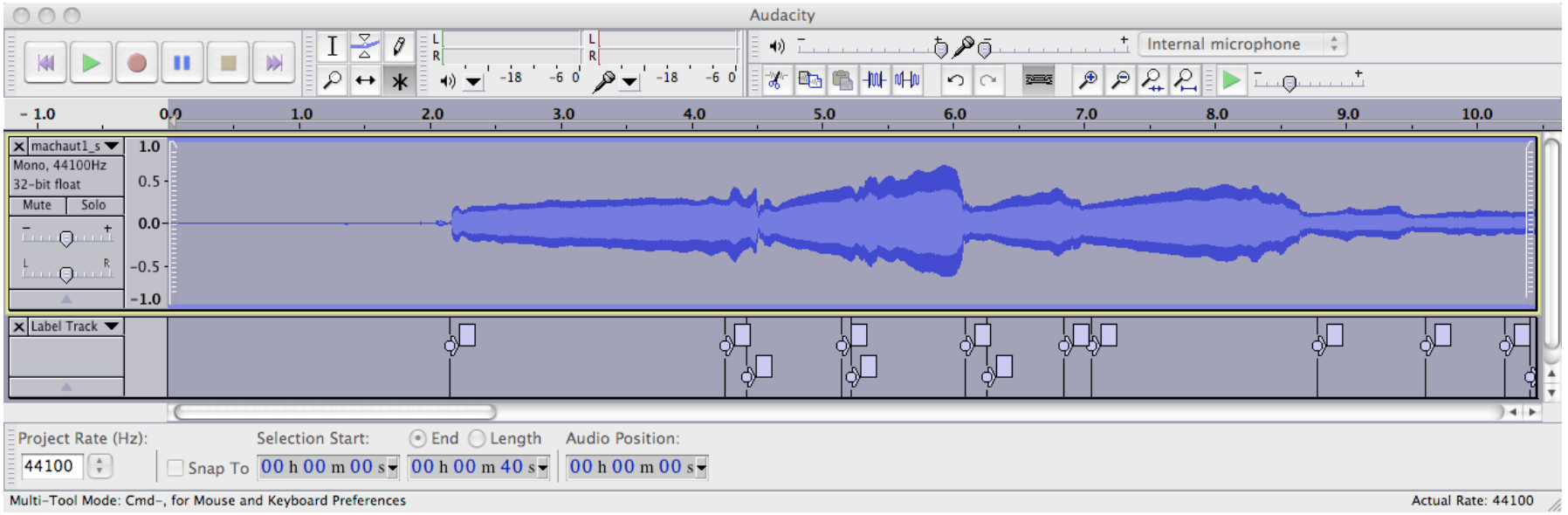
Use of Specialized Instruments

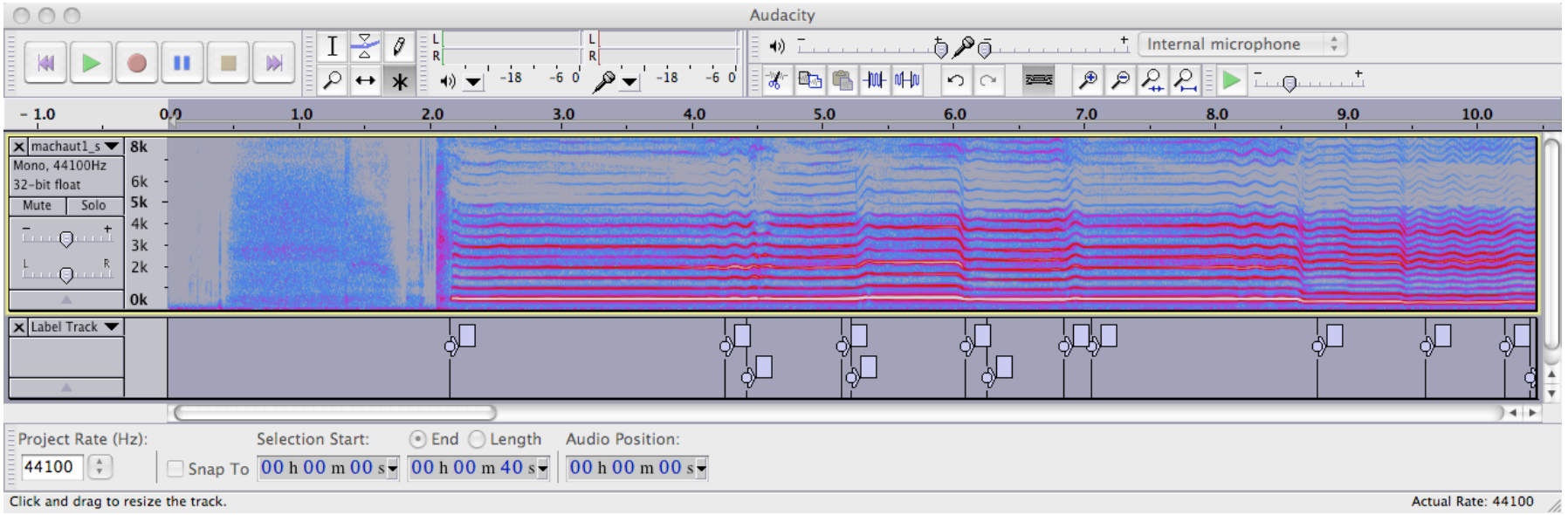
- Piano is particularly popular
- Some modified string instruments (e.g., guitar and cello) have also been studied
- Problems
  - cannot study existing recordings
  - new recordings are typically made in a lab environment
  - precision is limited for instruments other than the piano

Bosendorfer SE piano at BRAMS, Montreal

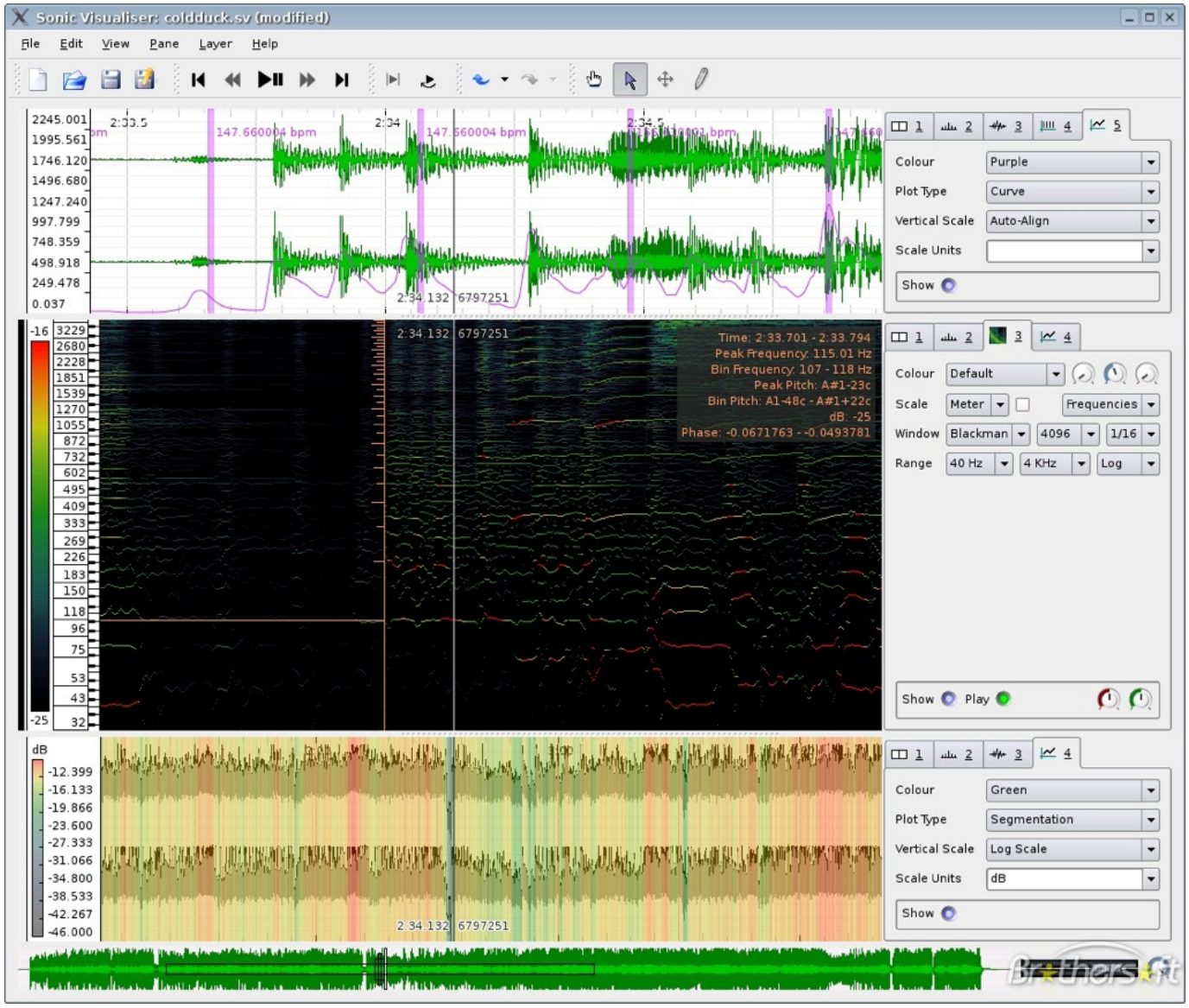
Manually Annotating Onsets byTapping

Manual annotation with software





Audio Onset Detection

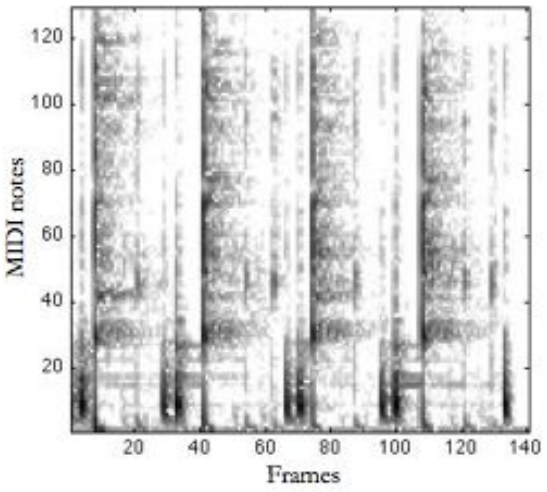


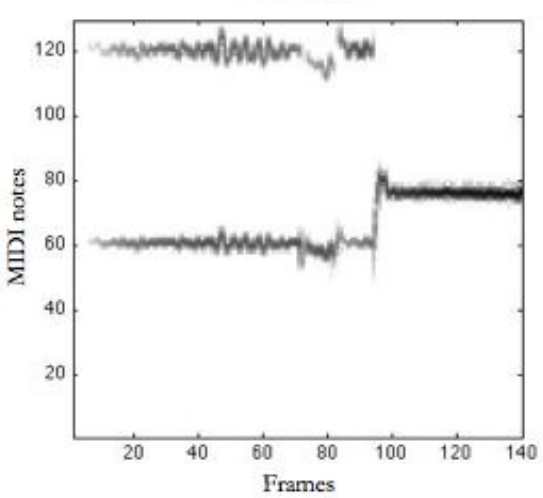
Audio Onset Detection

Solo Drum

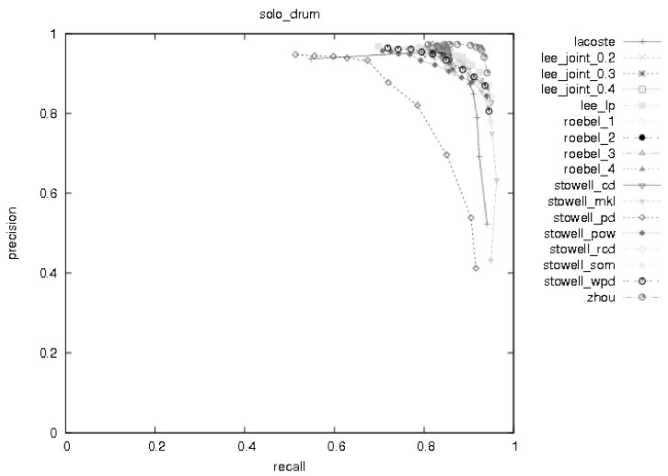
Solo Voice

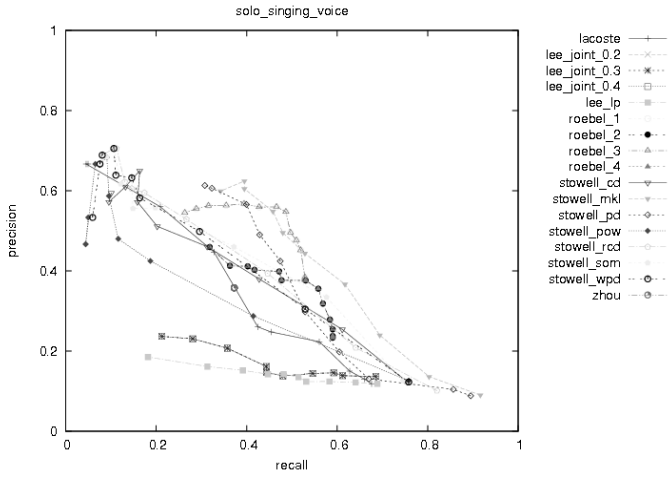
Example of spectrum





MIREX 2007 Results

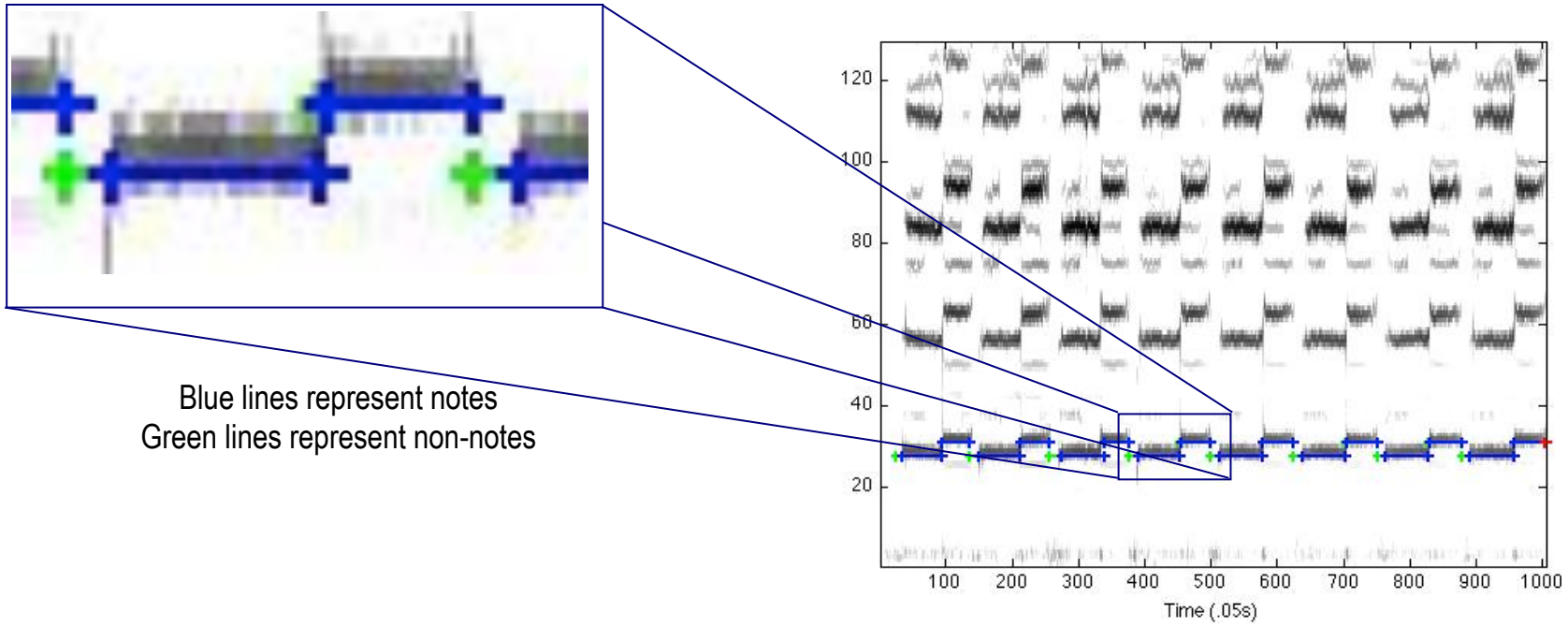




Precision = # of correctly detected onsets / (# of correctly detected onsets + # of false positive onsets)

Recall = # of correctly detected onsets / (# of correctly detected onsets + # of missed onsets)

MIDI-Audio Alignment



- Idea first put forth by Scheirer (1995)

Introduction to Studying Musical Performance

Prior Work on MIDI-Audio Alignment

Improved Method for Monophonic Performances

Improved Method for Polyphonic Performances

Conclusions

MIDI-Audio Alignment

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

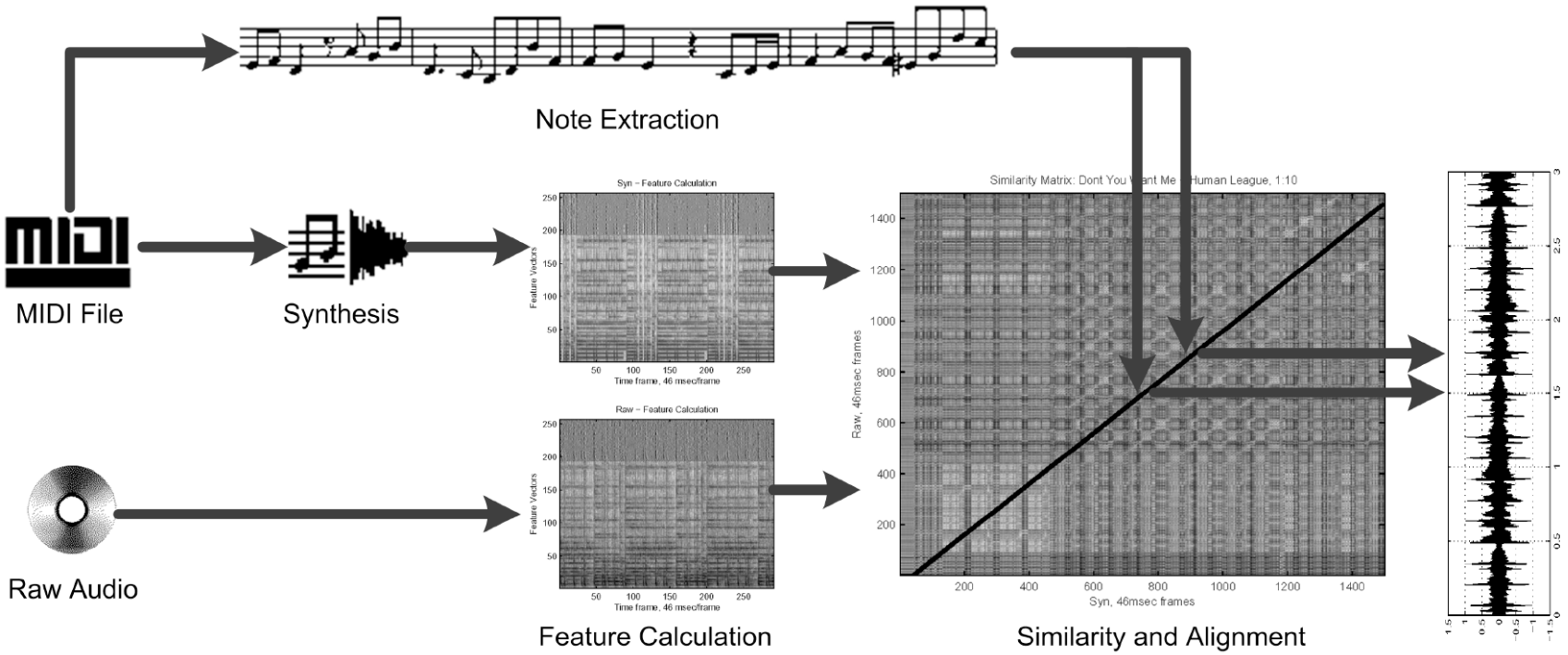
Brief History

- International Computer Music Conference
  - Dannenberg (1984) and Vercoe (1984)
    - Dannenberg made use of dynamic programming
- Dynamic Time Warping
  - Stammen and Pennycook (1993)
  - Orio and Schwarz (2001)
- Hidden Markov Models
  - Single-level (note-level)
    - Cano, Loscos, and Bonada (1999)
  - Multi-level (song- and note-levels)
    - Orio and Dechelle (2001)

Brief History

- More generalized graphical model (note- and tempo-levels)
  - Raphael (2004)
- Bayesian framework for score position pointer estimation
  - Peeling, Cemgil, and Godsil (2007)
  - Otsuka, Nakadai, Ogata, and Okuno (2011)
  - Duan and Pardo (2011)
- Semi-Markov model (dual-agent approach)
  - Cont (2010)

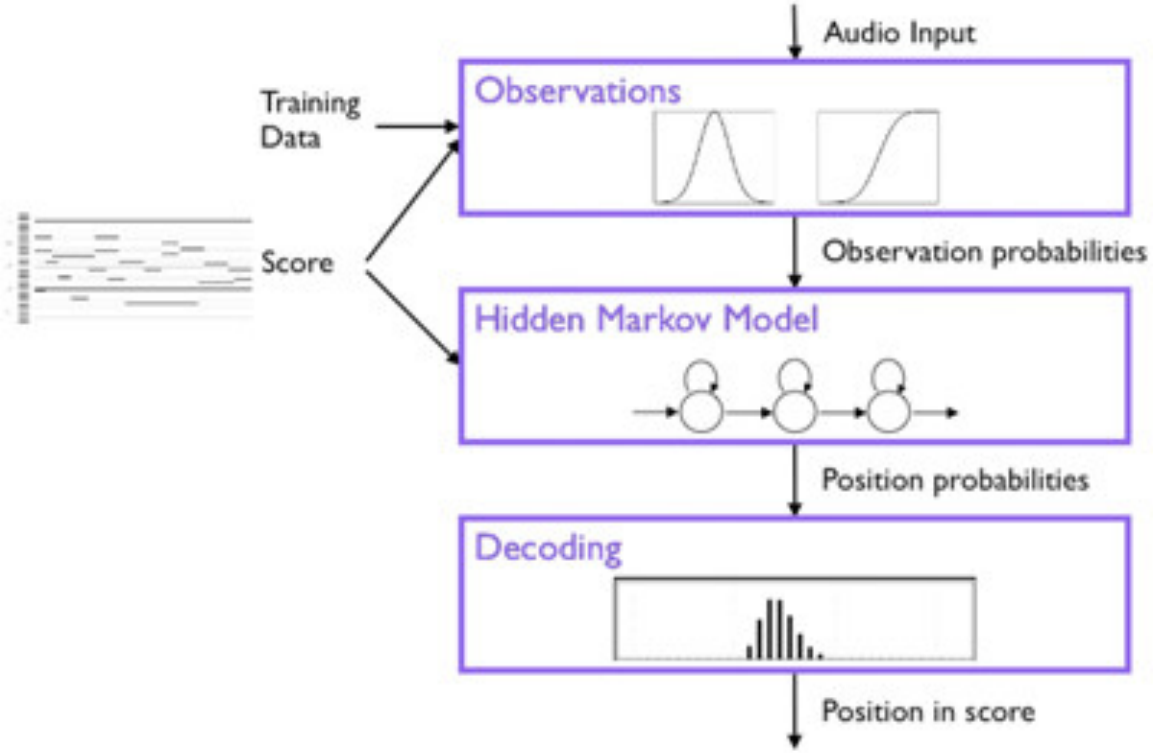
Dynamic Time Warping Overview



Turetsky and Ellis 2003

- Examples of Features
  - Peak structure distance - Orio and Schwartz (2001)
  - Chromagrams - Hu, Dannenberg, and Tzanetakis (2003)
  - Cosine distance - Turetsky and Ellis (2003)

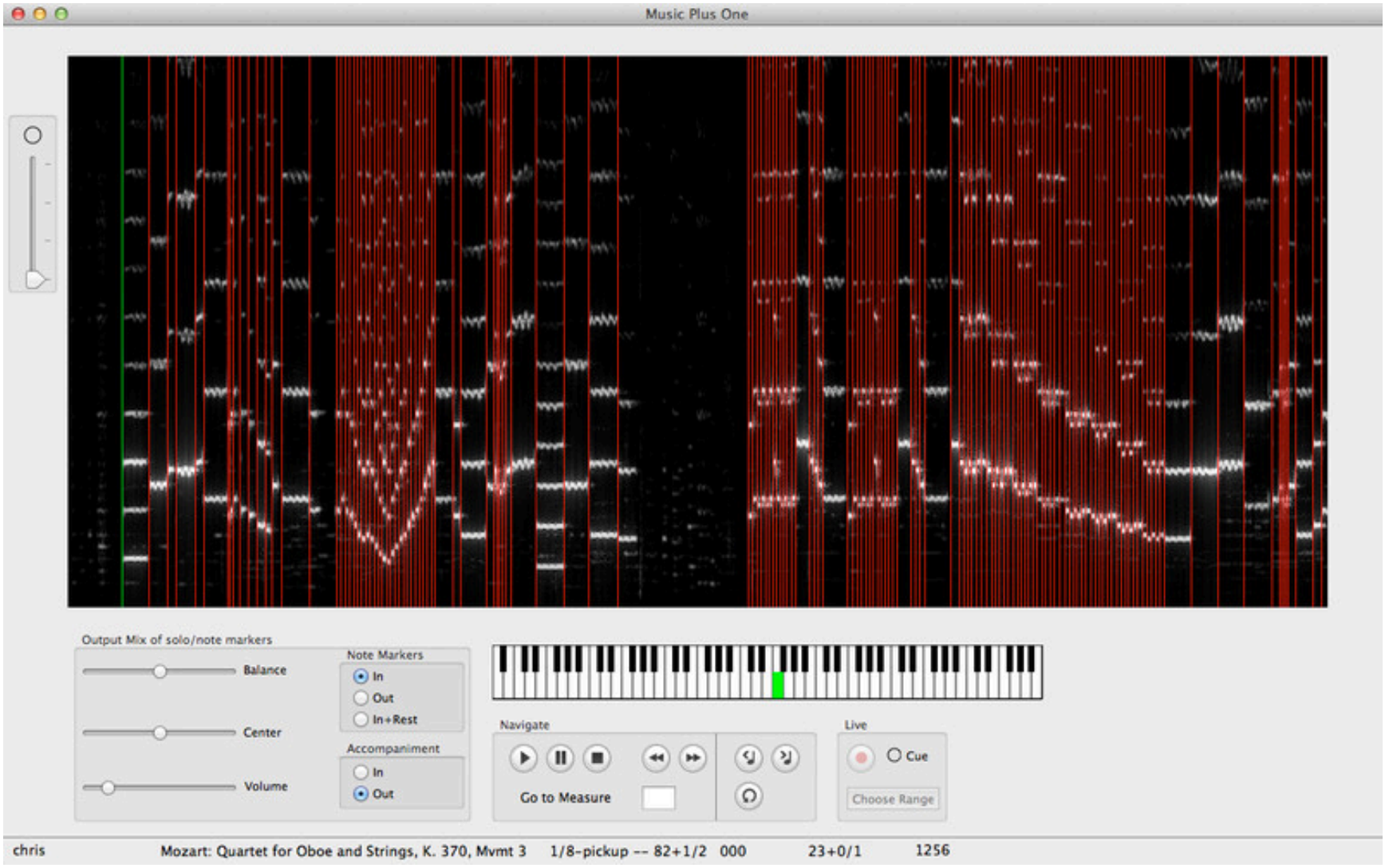
Hidden Markov Models



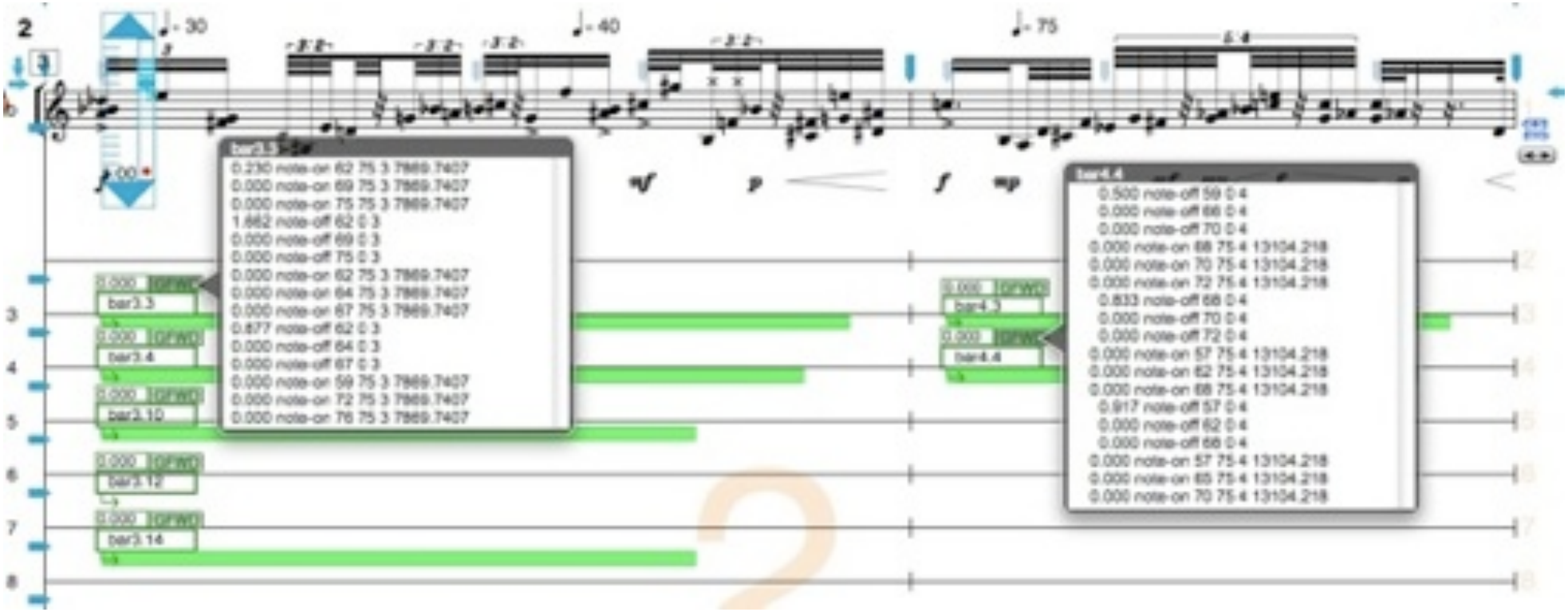
Cont and Shwarz 2006

- Examples of Observations/Features
  - Spectral Balance and Peak Structure Match - Cont and Schwarz (2006)
  - Power Spectrum of FFT - Raphael (2004)
  - Chroma and Pitch - Duan and Pardo (2011)

Raphael's Music Plus One (2001)



Cont's AnteScoFo (2010)



Multi-pass Approaches

- Devaney, Mandel, and Ellis (2009)
  - DTW + HMM
  - Monophonic recordings of the singing voice
- Nidermayer and Widmer (2011)
  - DTW + Non-negative Matrix Factorization (NMF)
  - Polyphonic piano pieces

Evaluation

- Required accuracy
  - Score Followers - estimates within 250 or 300 ms of the ground truth are considered correct
  - Digital Libraries - accuracy at either the note-level or bar-level
    - e.g., for a piece in 4/4 that is performed at 120 bpm
      - 500 ms for quarter note-level precision
      - 2 seconds for measure-level precision
  - Expressive Performance Studies - much more accuracy is required
    - e.g., asynchronies between polyphonic lines ranges 7–50 ms (Palmer 1997)

Evaluation

- Music Information Retrieval Evaluation eXchange (MIREX) Score Following Task (2006–12) - http://www.music-ir.org/mirex/
  - 47.38 minutes/8957 notes of primarily monophonic classical music audio for voice, clarinet, violin, and flute
  - Accuracy: 300 ms
  - Piecewise Precision: average precision across each piece
  - Total Precision: average precision across the whole database
  - Highest scoring system: 83% piecewise, 77% total

Introduction to Studying Musical Performance

Prior Work on MIDI-Audio Alignment

Improved Method for Monophonic Performances

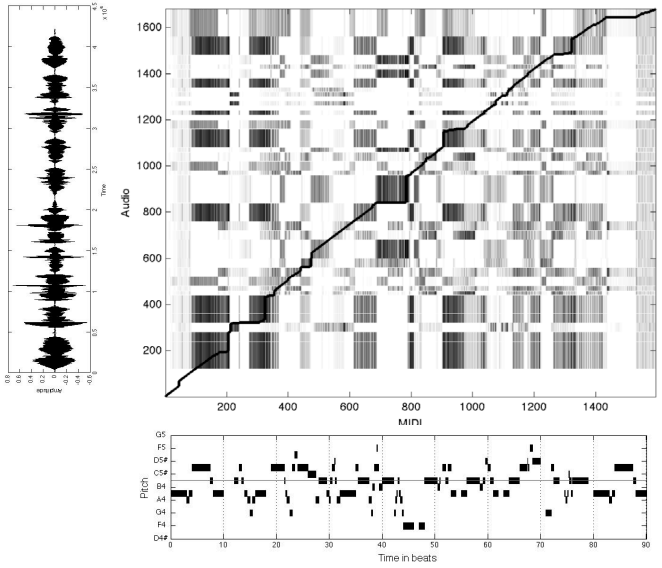
Improved Method for Polyphonic Performances

Conclusions

MIDI-Audio Alignment Algorithm

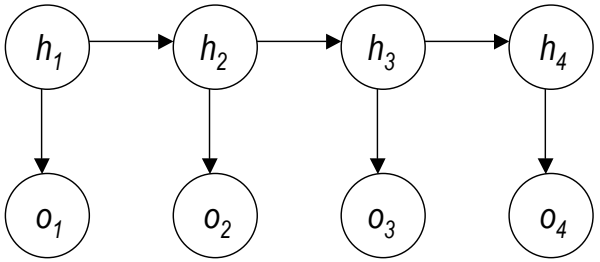
Step 1: Dynamic Time Warping alignment between MIDI and audio

Similarity Matrix



Orio and Schwarz (2001)

Step 2: Hidden Markov model adjustments of Dynamic Time Warping alignment

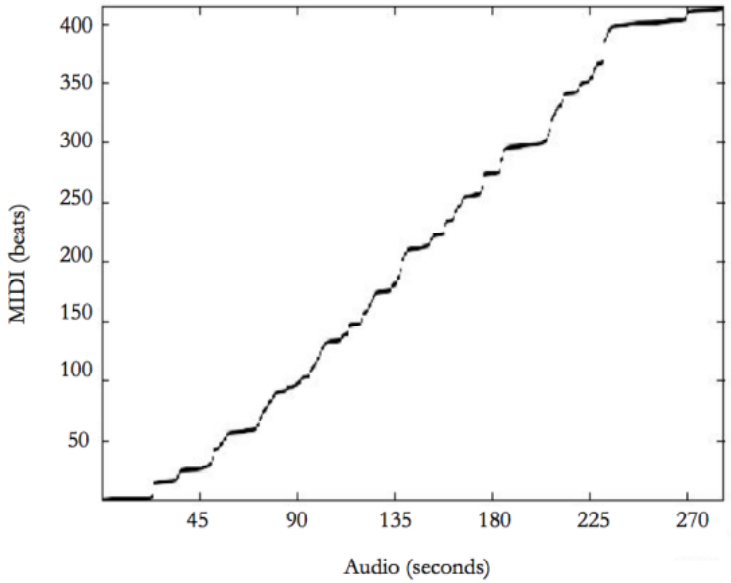


h = hidden states
o = observed states

Devaney, Mandel, and Ellis (2009)

Hidden Markov Model

Dynamic Time Warping alignment used as a prior for the HMM



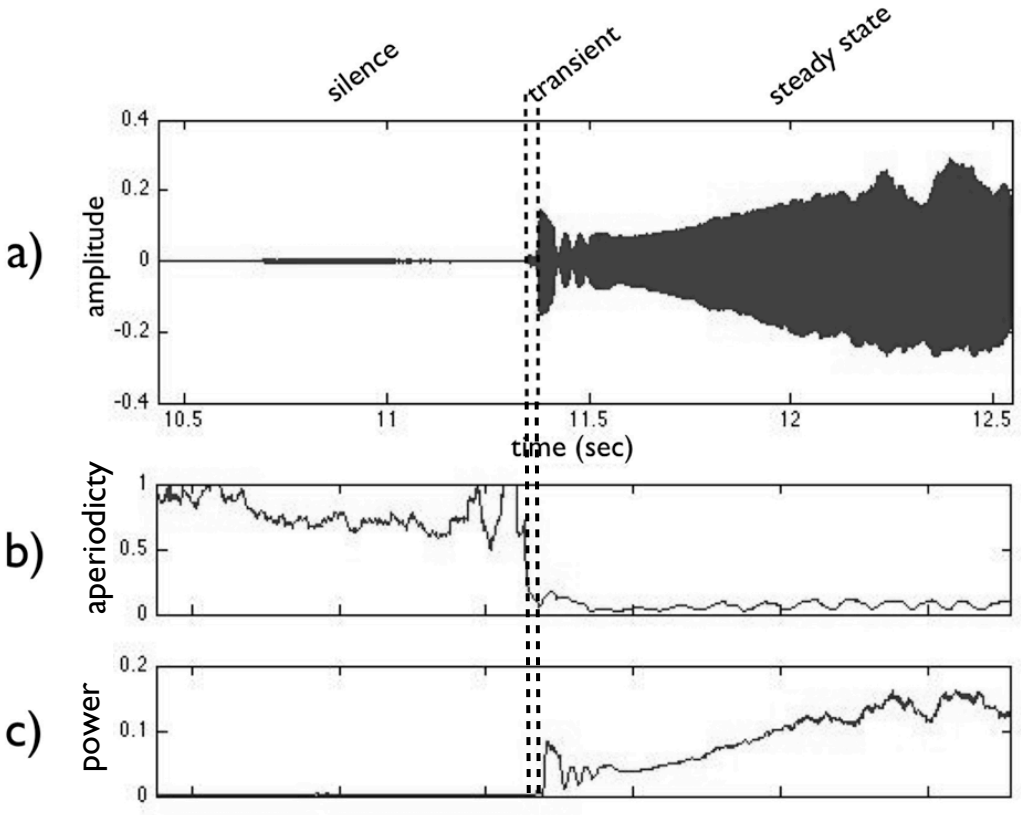
	5% start	100% start	100% end	5% end
Silence (and Breath)	50% between $N\text{-}1_{On}$ and $N\text{-}1_{Off}$	$N\text{-}1_{Off}$	N_{On}	50% between N_{On} and N_{Off}
Opening Transient	$N\text{-}1_{Off}$	75% between $N\text{-}1_{Off}$ and N_{On}	25% between N_{On} and N_{Off}	N_{Off}
Steady State	$N\text{-}1_{Off}$	N On	N Off	$N+1_{On}$
Closing Transient	N_{On}	75% between N_{On} and N_{Off}	25% between N_{Off} and $N+1_{On}$	$N+1_{On}$

Window functions were placed over the note positions in the DTW alignment

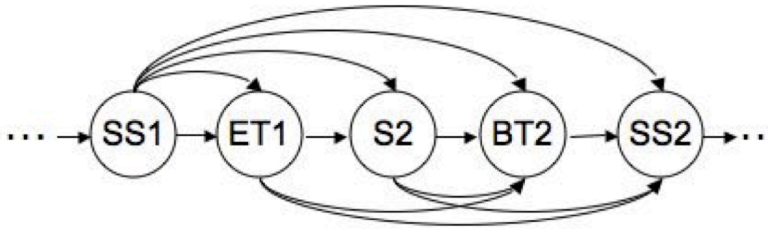
Devaney, Mandel, and Ellis (2009)

Hidden Markov Model

Acoustic Features of the Singing Voice

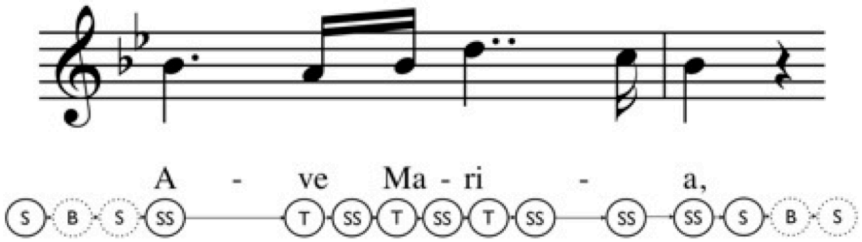


Basic State-Space Diagram



SS = Steady State
ET = Ending Transient
S = Silence
BT = Beginning Transient

Modified State-Space Diagram



Devaney, Mandel, and Ellis (2009)

Hidden Markov Model

Comparison of DTW to DTW+HMM Alignment

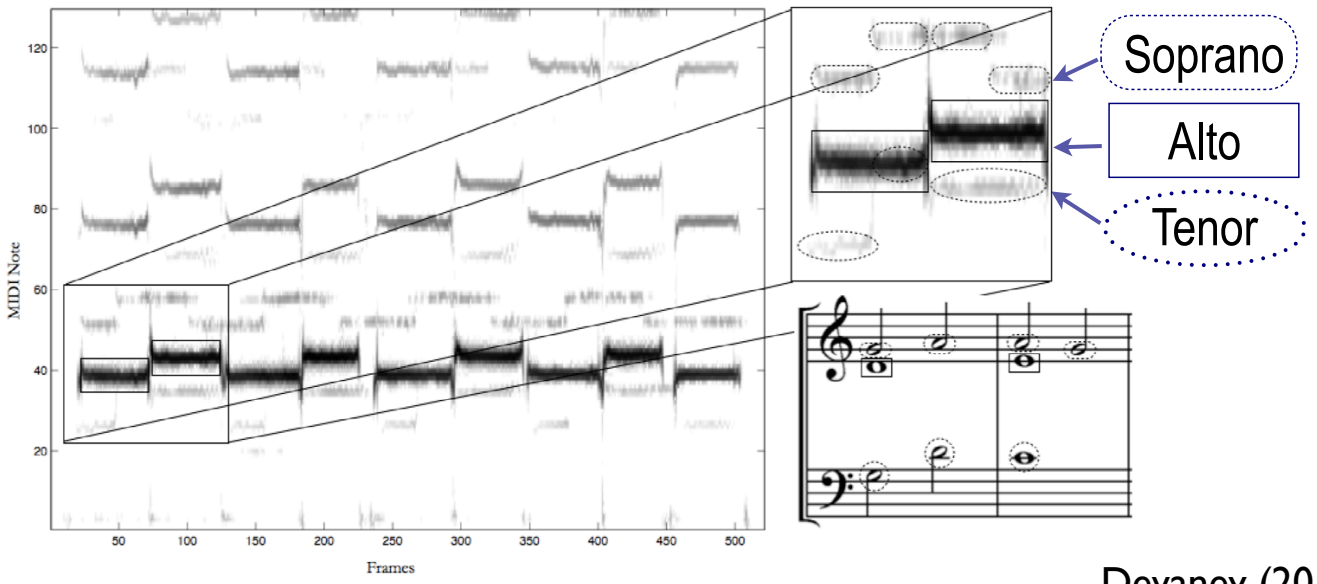
	Percentile				
	2.5	**25**	**50**	**75**	**97.5**
Dynamic Time Warping	3.2	32.6	52.3	87.9	**478.7**
Basic State Space	**1.6**	**13.1**	41.8	88.8	564.1
Modified State Space	**1.6**	**13.1**	**27.8**	**78.0**	506.0

Alignment errors in milliseconds

Devaney, Mandel, and Ellis (2009)

Quasi-Polyphonic Signals

- This algorithm also works on signals where there is a dominant voice and some bleed-through from other ensemble participants
- The alignment for such signals can be used to guide F_0 estimation
  - The YIN algorithm's ability to specify the minimum and maximum expected F_0 allows it to work for these quasi-polyphonic signals (de Cheveigné and Kawahara 2002)



Devaney (2011)

Introduction to Studying Musical Performance

Prior Work on MIDI-Audio Alignment

Improved Method for Monophonic Performances

Improved Method for Polyphonic Performances

Conclusions

Asynchrony in Polyphonic Performances

- There are typically asynchronies in musical performance for events that are notated as simultaneous in the score (Palmer 1996)
- Most polyphonic methods are unable to account for these asynchronies
  - Exception: Niedermeyer and Widmer's work using NMF on piano performances (2011)

Issues with 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 identify asynchronies
  - aligning individual lines
    - pros: timing of each line can vary independently
    - cons: highly prone to errors

Evaluation of Dynamic Time Warping

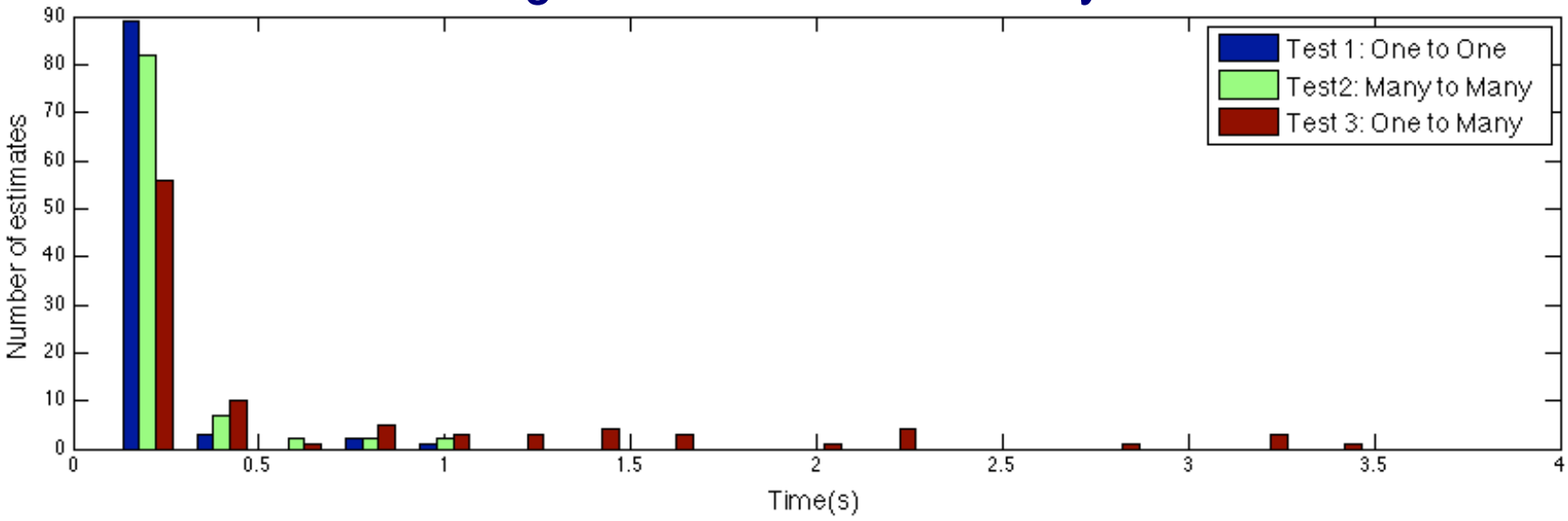
- Test data comprised of hand-annotated excerpts of four-track recordings from Machaut's Messe de Notre Dame

Soprano Line Mixdown

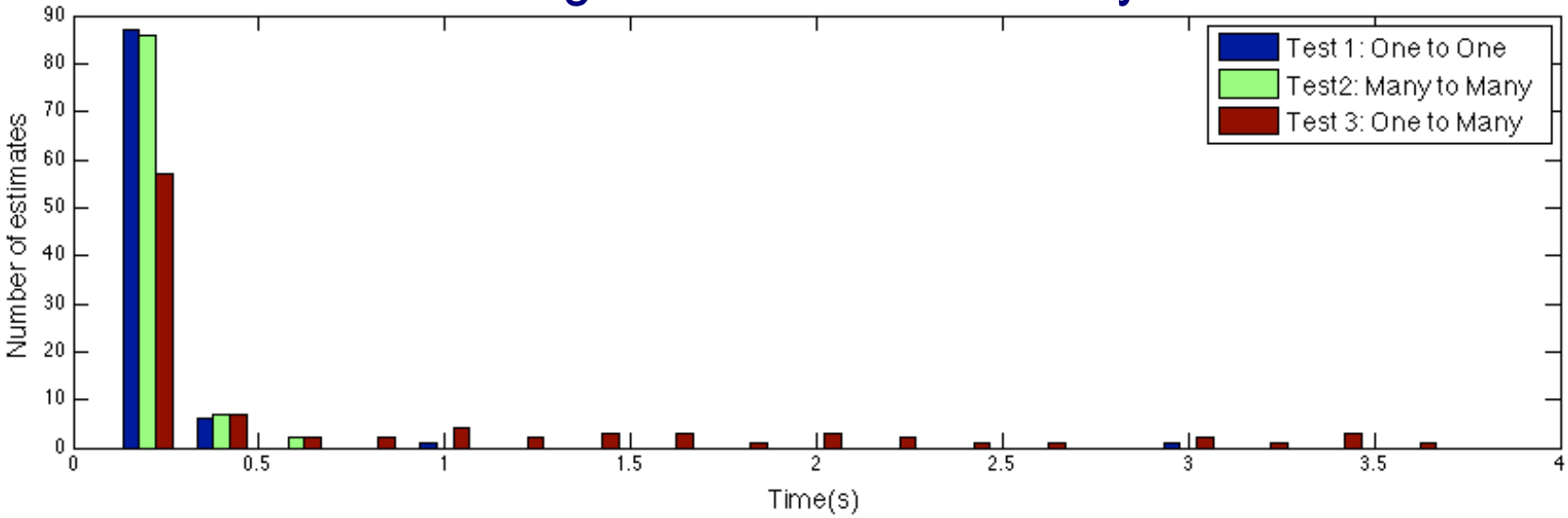
- Three alignment tests were performed on the test data
  - One to One - each voice to the recording of the individual track
  - Many to Many - the four voices simultaneously to the mixdown of the individual tracks
  - One to Many - each voice to the mixdown of the individual tracks

Devaney and Ellis (2008)

Histogram of Onset Accuracy



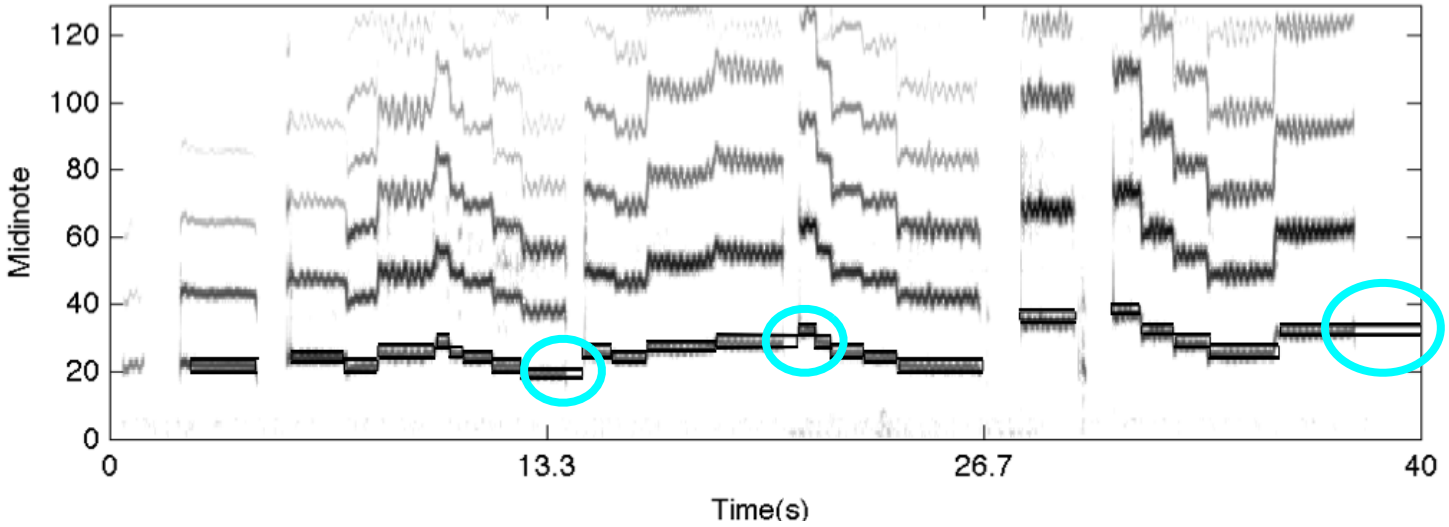
Histogram of Offset Accuracy



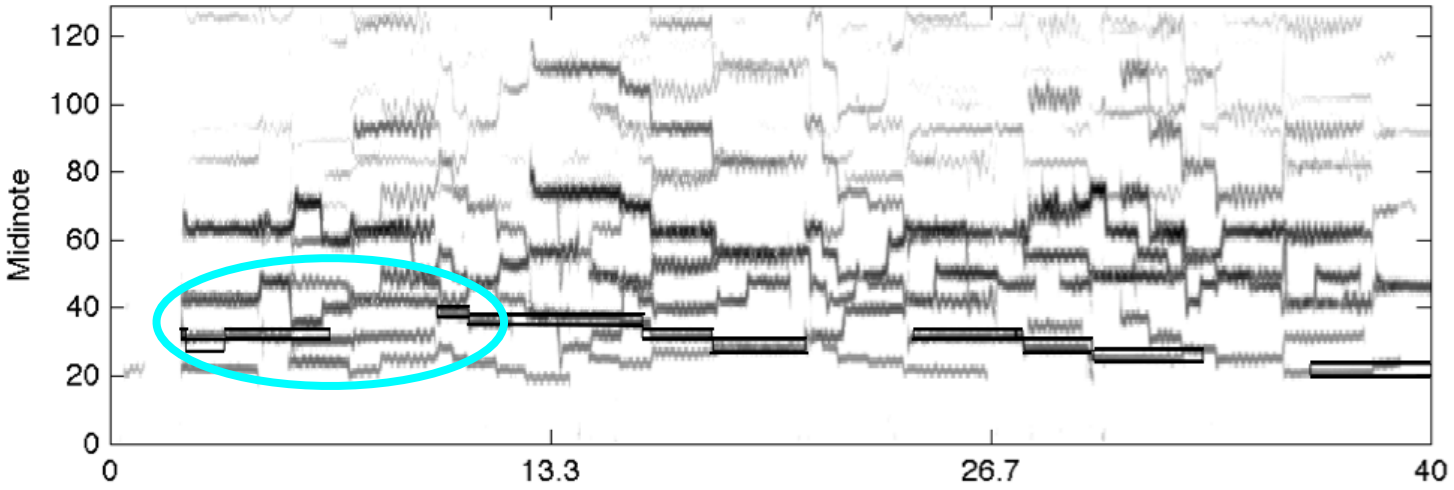
Devaney and Ellis (2008)

Evaluation of Dynamic Time Warping Approach

One to One



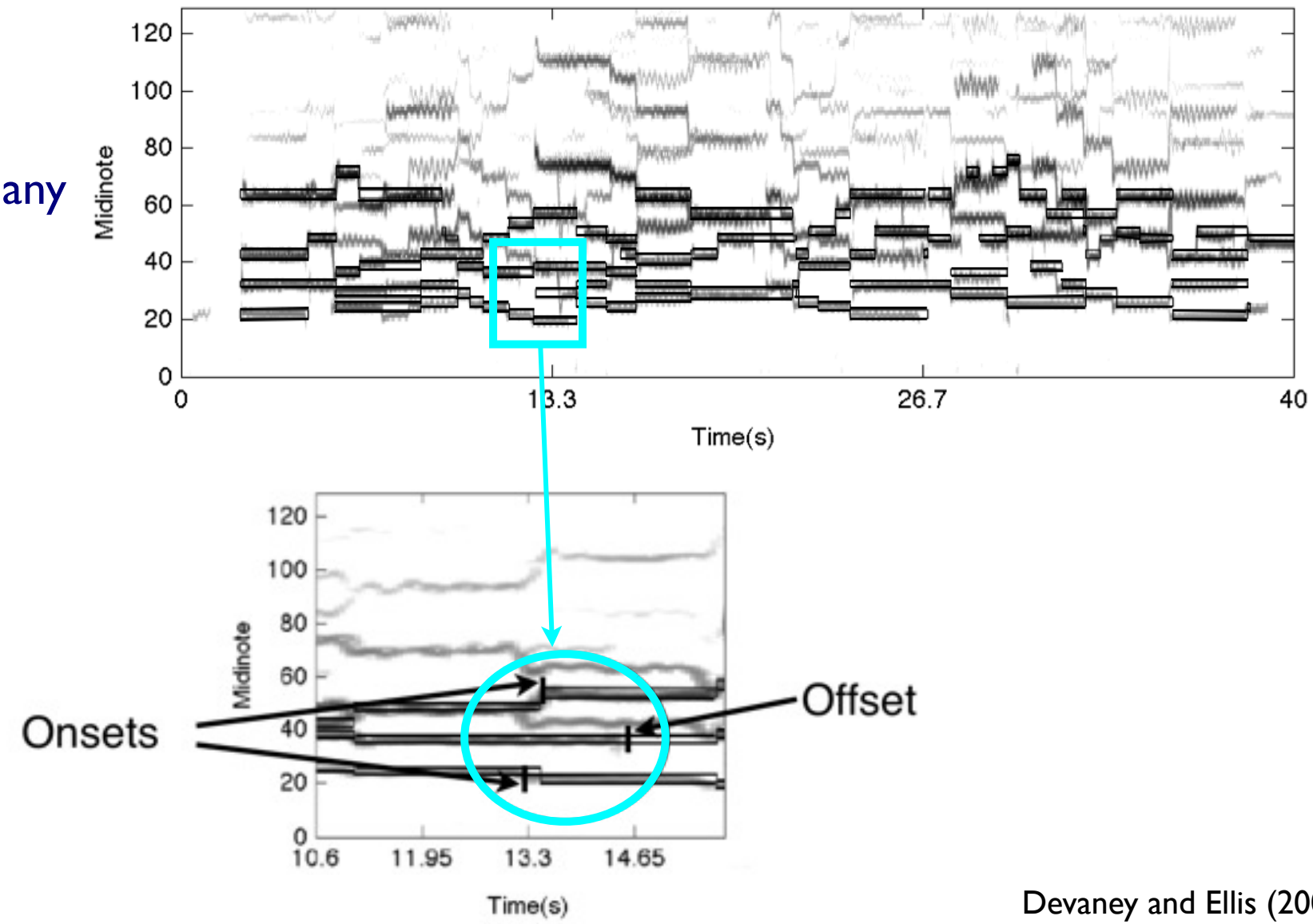
One to Many



Devaney and Ellis (2008)

Evaluation of Dynamic Time Warping Approach

Many to Many



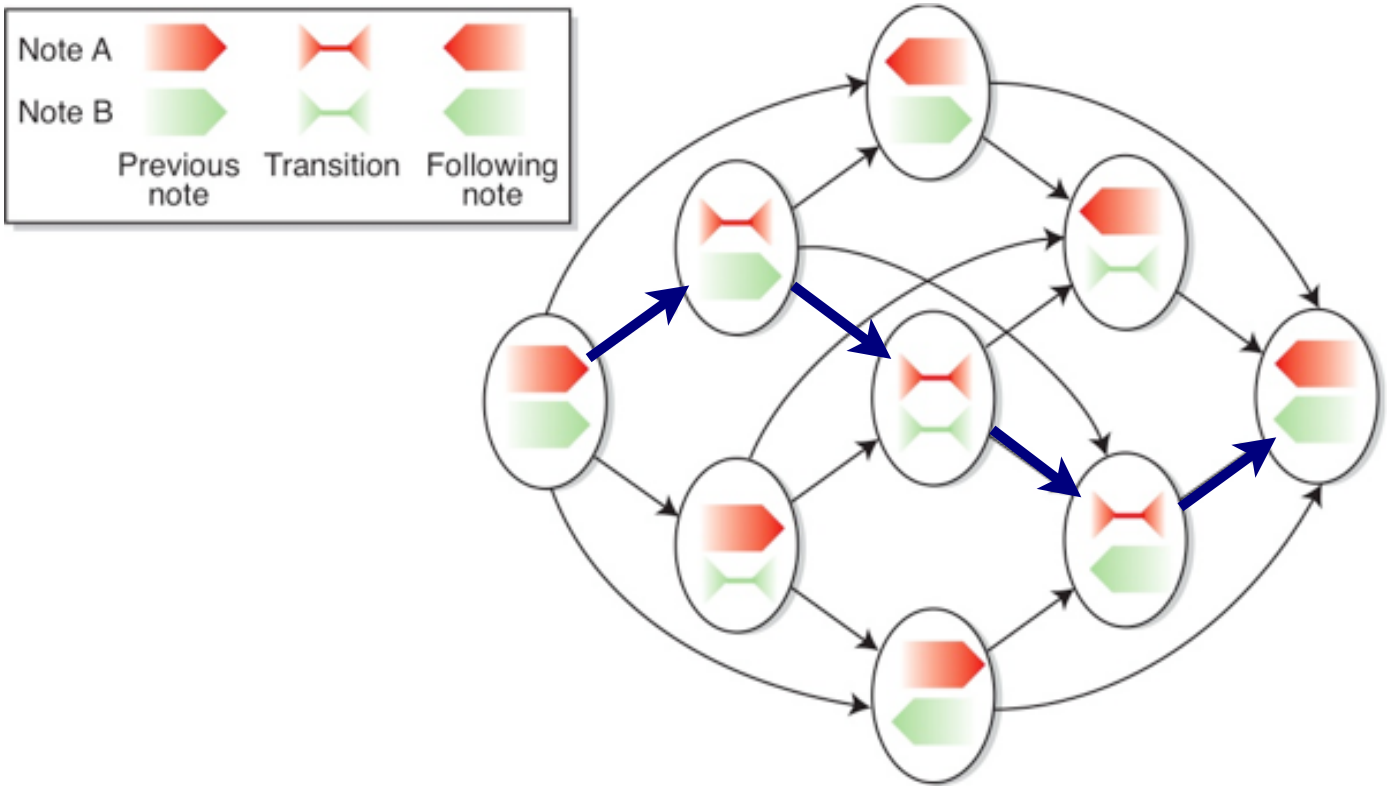
Devaney and Ellis (2008)

Multi-pass Approach

- As with monophonic algorithm
  - DTW on the full polyphonic score achieves a rough alignment
  - This can be refined by realigning the audio between the notes
- Each note goes through a three-state sequence
  - <End of first note> – <Silence> – <Start of second note>
  - The complexity of this is 3^N
    - where N is the number of simultaneous notes

Transition Matrix for Two Notes

- Note A1 ends
- Note B1 ends
- Note B2 begins
- Note A2 begins



Introduction to Studying Musical Performance

Prior Work on MIDI-Audio Alignment

Improved Method for Monophonic Performances

Improved Method for Polyphonic Performances

Conclusions

Conclusions

- Standard DTW-based approaches are generally robust for MIDI-audio alignment tasks
  - They lack the accuracy for annotating monophonic or polyphonic performances
- Multi-pass approaches capitalize on the general robustness of DTW while providing a greater level of accuracy

Acknowledgements

- Center for New Music and Audio Technologies (CNMAT)
- Distributed Digital Music Archives and Libraries (DDMAL)
- Centre for Research in Music Media and Technology (CIRMMT)
- Fonds de recherche sur la société et la culture (FQRSC)
- Social Sciences and Humanities Research Council of Canada (SSHRC)
- Advancing Interdisciplinary Research in Singing (AIRS)

Thank you!

References

Cano P., Loscos A., Bonada J. 1999. Score-performance matching using HMMs. In *Proceedings of ICMC*, 441–4.

Cont A. 2010. A coupled duration-focused architecture for realtime music to score alignment. *IEEE TPAML*. 32 (6):974– 87

Dannenberg R. 1984. An on-line algorithm for real-time accompaniment. In *Proceedings of ICMC*, 193–8.

de Cheveigné A.. and H. Kawahara. 2002. YIN: A fundamental frequency estimator for speech and music. *JASA*. 111: 1917–30.

Devaney, J. 2011. An empirical study of the influence of musical context on intonation practices in solo singers and SATB ensembles. Doctoral Thesis. McGill University.

Devaney J., M. Mandel, and D. Ellis. 2009. Improving MIDI-audio alignment with acoustic features. In *Proceedings of WASPAA*. 45–8.

Devaney J., Mandel M., Fujinaga I. 2012. A Study of Intonation in Three-Part Singing using the Automatic Music Performance Analysis and Comparison Toolkit (AMPACT). In *Proceedings of ISMIR*, 511–6.

Duan Z. and B. Pardo. 2011. A state space model for online polyphonic audio-score alignment. In *Proceedings of ICASSP*, 197–200.

Hu N., Dannenberg R., and G. Tzanetakis 2003. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of WASPAA*, 185–8.

Niedermayer B. and g. Widmer. 2010. A multi-pass algorithm for accurate audio-to-score alignment. In *Proceedings of ISMIR,* 417–22.

Orio N. and F. Déchelle. 2001. Score following using spectral analysis and hidden Markov models. *In Proceedings of ICMC*, 151–4.

Orio, N., and D. Schwarz. 2001. Alignment of monophonic and polyphonic music to a score. In *Proceedings of ICMC*, 155–8.

Otsuka T, K. Nakadai, T. Ogata, and H. Okuno 2011. Incremental Bayesian audio-to-score alignment with flexible harmonic structure models. In *Proceedings of ISMIR*, 525–30.

Palmer, C. 1997. Music performance. *Annual Review of Psychology.* 48: 115–38.

Peeling P., T. Cemgil, and S. Godsill. 2007. A probabilistic framework for matching music representations. *In Proceedings of ISMIR*, 267–72

Raphael, C. 2001. Music plus one: A system for expressive and flexible musical accompaniment. In *Proceedings of ICMC*, 159–62.

Raphael, C. 2004. A hybrid graphical model for aligning polyphonic audio with musical scores. In *Proceedings of ISMIR*, 387–94.

Scheirer E. 1995. Extracting expressive performance information from recorded music. Master's Thesis, Massachusetts Institute of Technology.

Stammen, D., and B. Pennycook. 1993. Real-time recognition of melodic fragments using the dynamic timewarp algorithm. In *Proceedings of ICMC*, 232–5.

Turetsky, R., and D.P.W. Ellis. 2003. Ground-truth transcriptions of real music from force-aligned midi synthesis. In *Proceedings of ISMIR*, 135– 42.

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