Automatic Extraction of Performance Data from Recordings

Johanna Devaney
McGill University
devaney@music.mcgill.ca

DDMAL Distributed Digital Music Archives & Libraries Lab
CIRMMT Centre for Interdisciplinary Research in Music Media and Technology
McGill Schulich School of Music École de musique Schulich
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

Brief History of Performance Analysis

Challenges of Automatically Extracting Performance Data

Improved MIDI/Audio Alignment Technique for the Singing Voice

Case Study: Intonation in Schubert's 'Ave Maria'

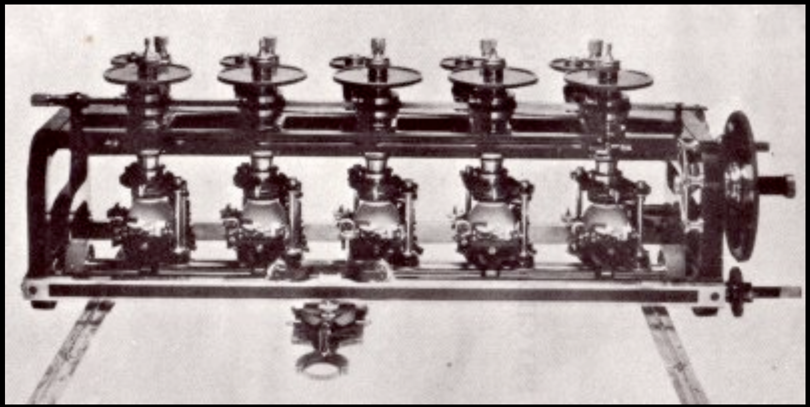
Conclusions

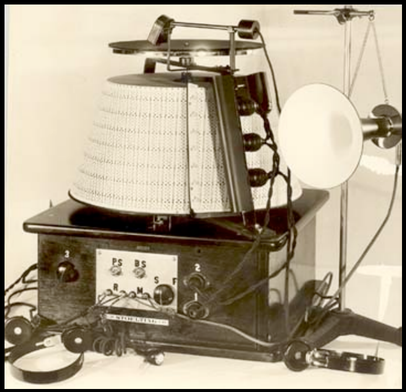
INTRODUCTION

- In 1938 Carl Seashore suggested that emotion is conveyed in performance through deviations from a norm
- In order to determine what is the ‘norm’ and what is ‘expression’ we need to examine a large number of performances
- Manual extraction of performance data of recordings is an arduous task
- Automatic extraction is a challenging, and as of yet unsolved, task
- This talk presents some work I have undertaken with Ichiro Fujinaga, Dan Ellis, and Michael Mandel towards this goal of automatic extraction

Brief History of Performance Analysis

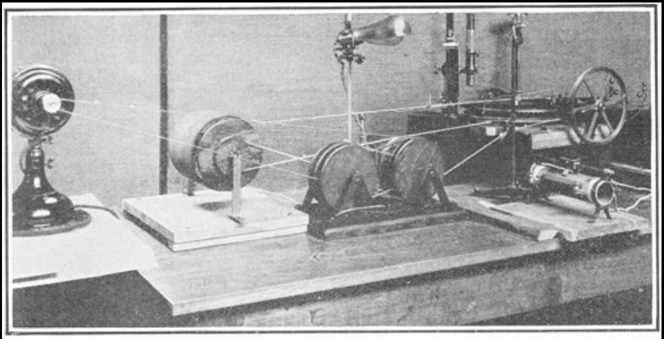
- Carl Seashore (1938) studied timing, dynamics, intonation, and vibrato in pianists, violinists, and singers
  - Equipment: piano rolls, films of the movement of hammers during performance, phono-photographic apparatus





Brief History of Performance Analysis

- Interest in empirical performance analysis diminished between the Second World War and 1980's, in part due to its labouriousness



Wave recorder for use with disk phonograph; the lever, acting like a pantograph, traces the waves on a revolving smoked drum

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

Brief History of Performance Analysis

- The resurgence in interest in the late 1970s/early 1980s coincided with
  - a movement by musicologists away from equating scores with music
  - an increased interest by cognitive psychologists in music

Brief History of Performance Analysis

- Ingemar Bengtsson and Alf Gabrielsson (1980) undertook a number of pioneering experiments on musical rhythm in performance
- Neil Todd (1985) studied both rubato and dynamics in piano performance
- Eric Clarke (1989) related rhythmic tendencies to both the structural hierarchy of the piece and note-level expressive gestures
- Bruno Repp (1992) also examined timing in piano performance and related it to phrase hierarchy
- Surveys are available in Palmer (1997) and Gabrielsson (1999, 2003)

Brief History of Performance Analysis

- AHRC Research Centre for the History and Analysis of Recorded Music (CHARM) and AHRC Research Centre for Musical Performance as Creative Practice (CMPCP)
  - Nicolas Cook, John Rink, Nicolas Cook, and Craig Sapp
- Machine Learning, Data Mining, and Intelligent Music Processing Group
  - Gerhard Widmer, Simon Dixon, and Werner Goebl
- Other researchers
  - Roger Dannenberg (Carnegie Mellon University)
  - Christopher Raphael (Indiana University)
  - Douglas Eck (Université de Montréal)

Brief History of Performance Analysis

- Piano performance is widely studied due to
  - the large amount of solo repertoire
  - the instrument's percussive nature
  - the ease with which one can acquire accurate, minimally intrusive performance measurements from a pianist via MIDI technology
  - the feasibility of using specially equipped pianos to measure performance data

Doug Eck's Bosendorfer SE Piano at BRAMS

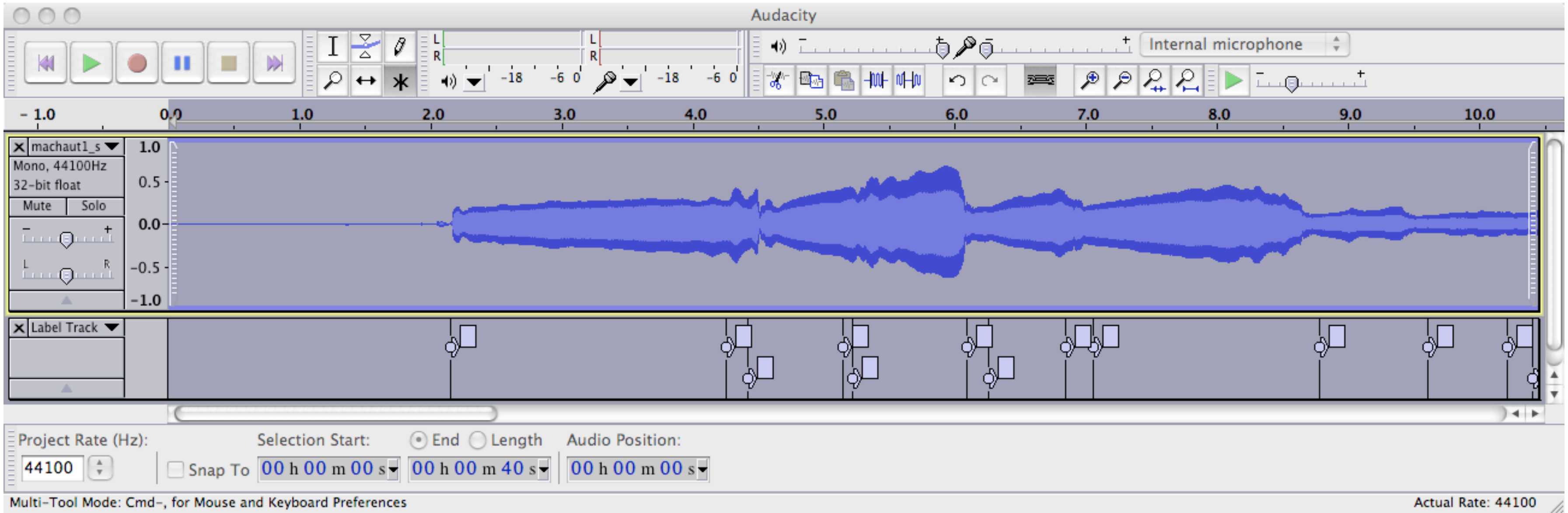
Brief History of Performance Analysis

- Issues with MIDI-based studies
  - require a MIDI-rigged piano
  - typically done in a lab environment
  - precision is limited for other instruments
- Signal processing techniques allow for extraction of performance data from recorded signals

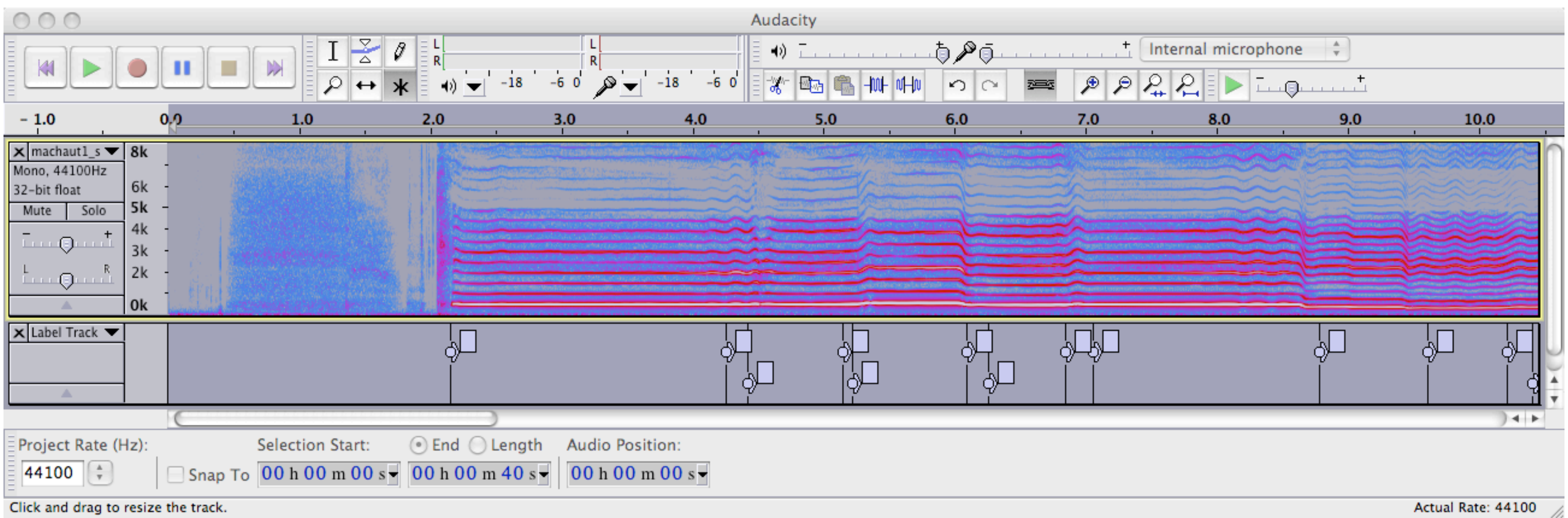
Extracting Performance Data

- Onsets and offsets
  - Onsets are needed for timing information
  - Onsets and offsets are needed for calculation of parameters over the duration of the note
- Fundamental frequency
  - Frame-wise fundamental frequency estimates are needed to calculate intonation and vibrato
- Power
  - Necessary to calculate dynamics

Time Domain Representation of Singing Voice in Audacity (with labels)

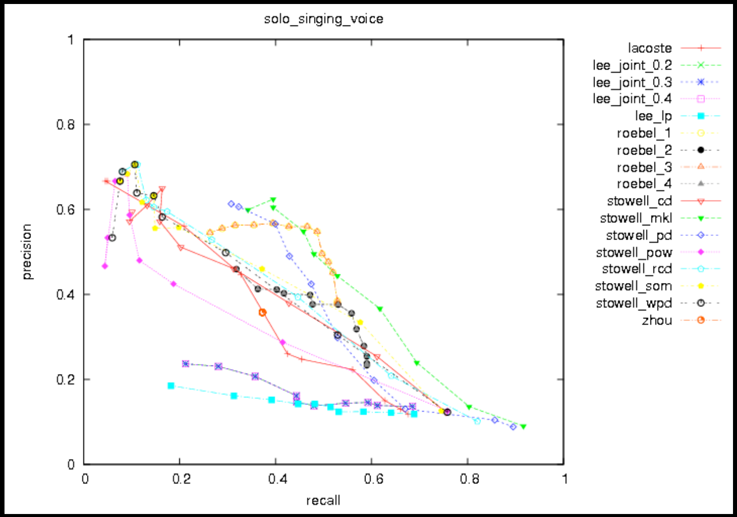


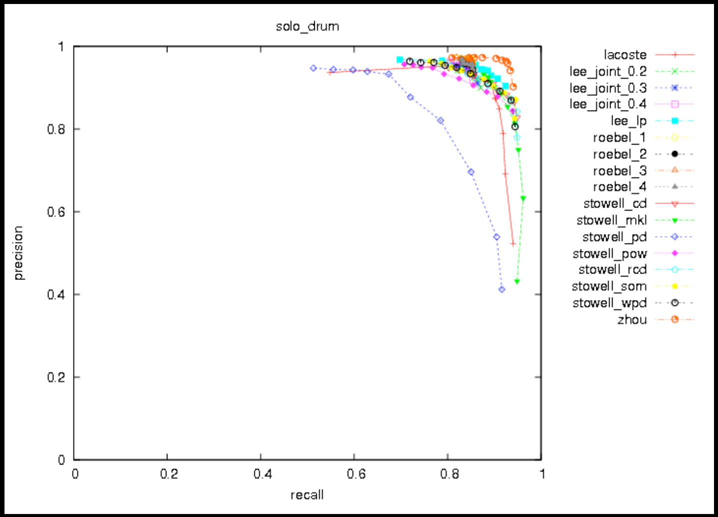
Frequency Domain Representation of Singing Voice in Audacity (with labels)



EXTRACTING ONSETS AND OFFSETS

- Existing onset detection methods work for instruments with percussive onsets, e.g., piano
  - they generally perform poorly for non-percussive instruments (MIREX Audio Onset Detection, 2007)





- Much less work has been done on offset detection

Extracting Onsets and Offsets

- MIDI/Audio alignment is another option for onset and offset detection
  - 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

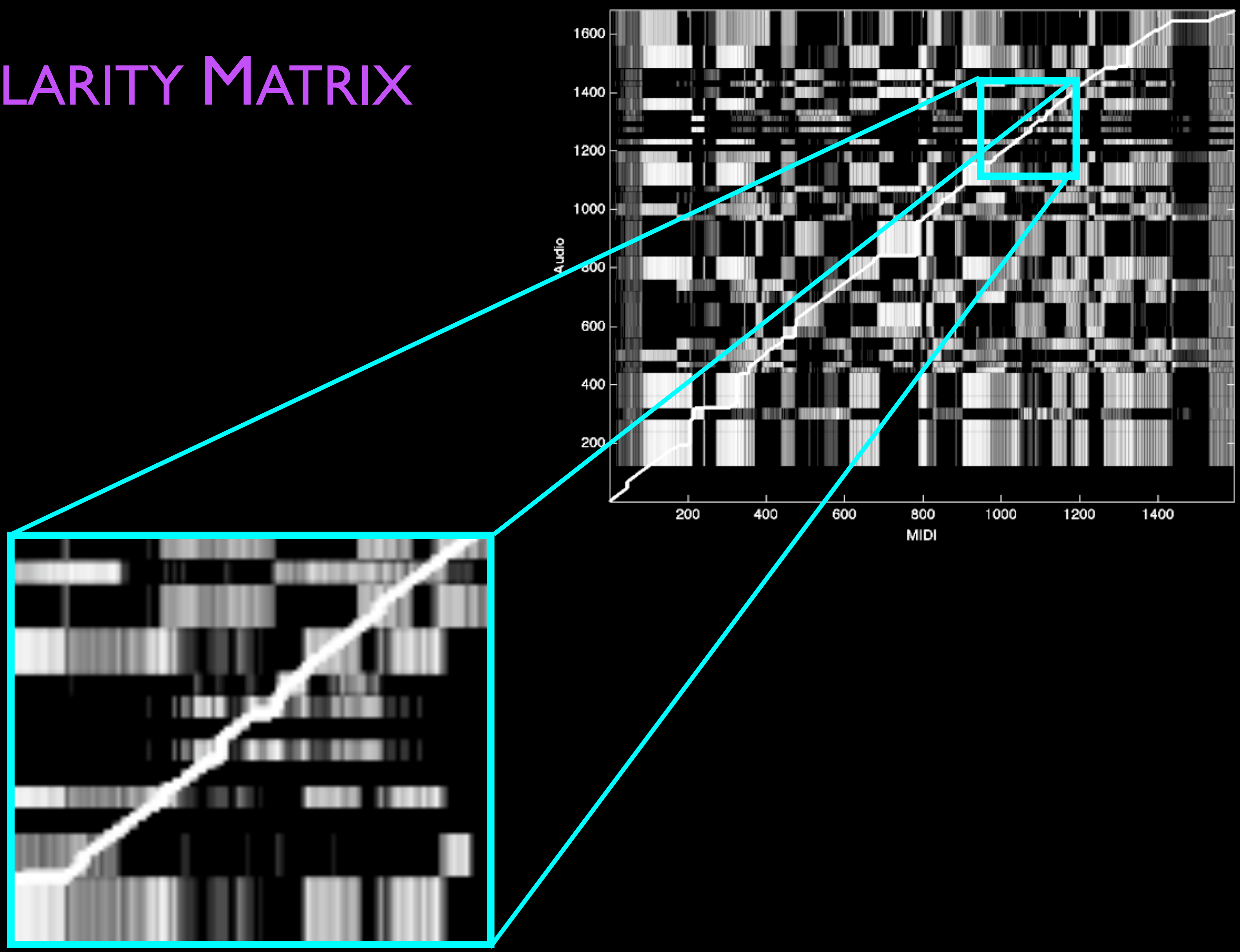
Extracting Onsets and Offsets

- A brief history of MIDI/audio alignment
  - 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

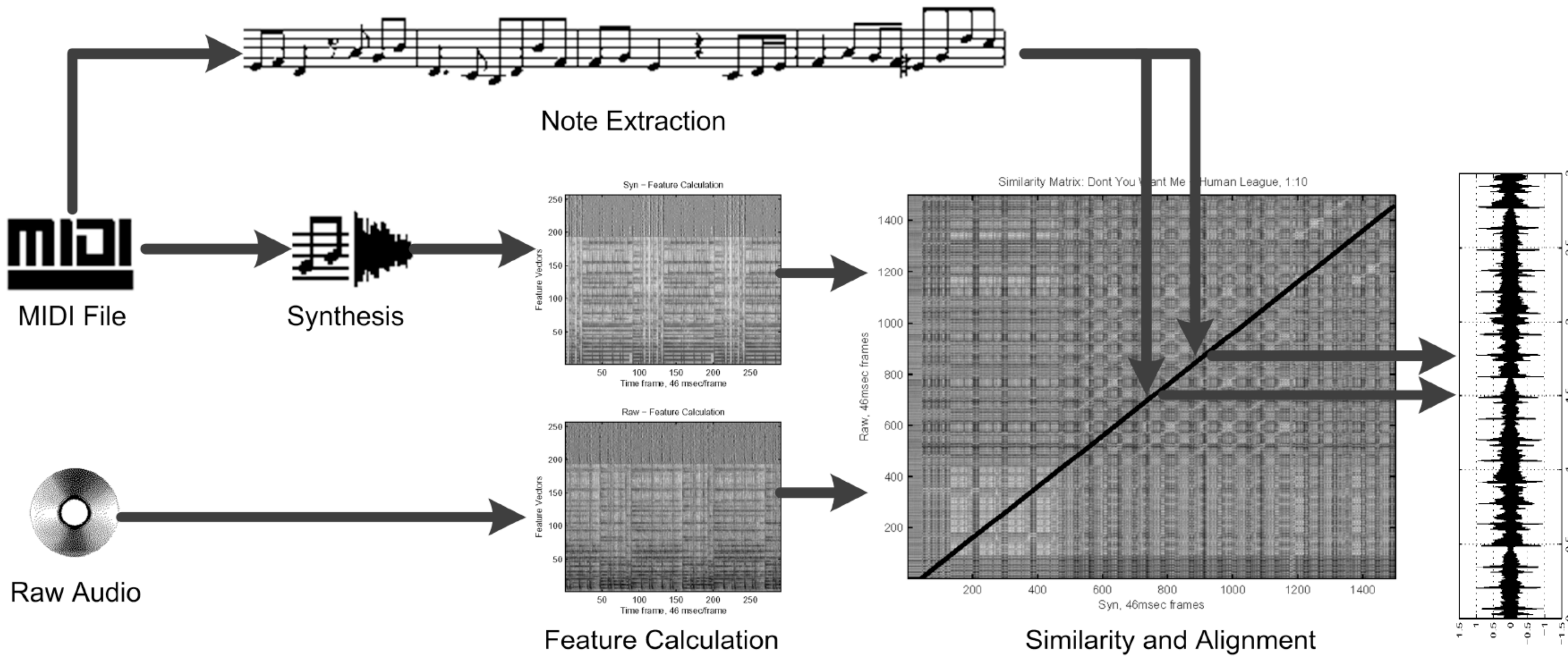
Dynamic Time Warping

- 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

Similarity Matrix



DYNAMIC TIME WARPING OVERVIEW



Turetsky and Ellis 2003

Extracting Pitch Data

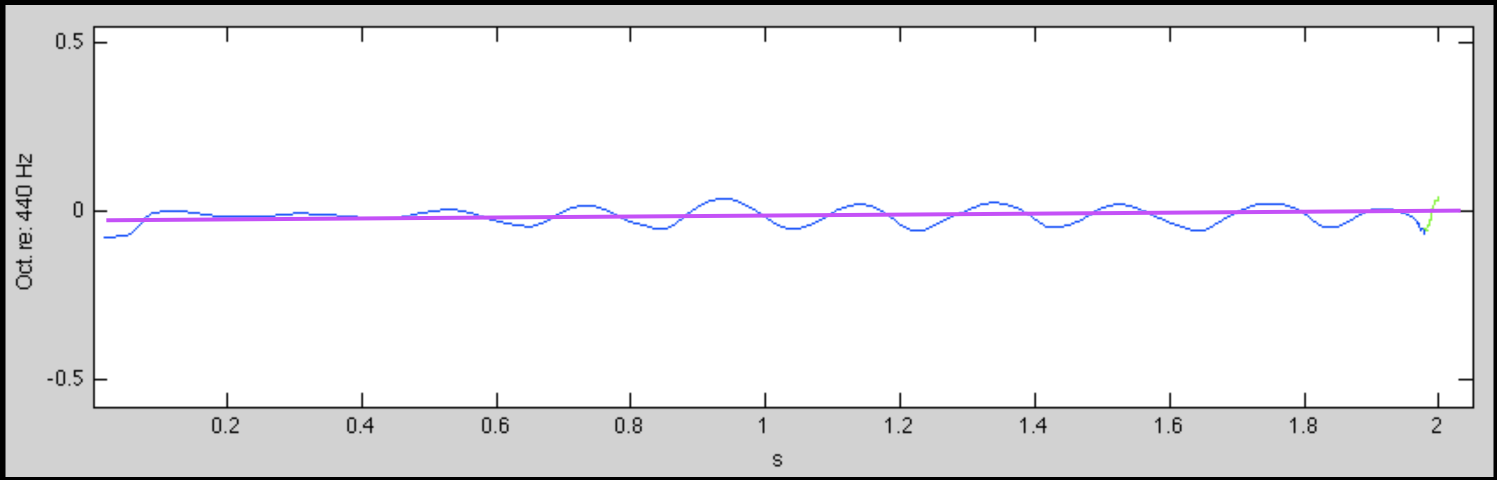
- Frame-wise fundamental frequency estimation for monophonic signals can be done in either the
  - frequency domain: peak picking, template matching
  - time domain: autocorrelation
- YIN (de Cheveigné and Kawahara 2002) is a time domain approach that, like autocorrelation, measures self-similarity over time

Extracting Pitch Data

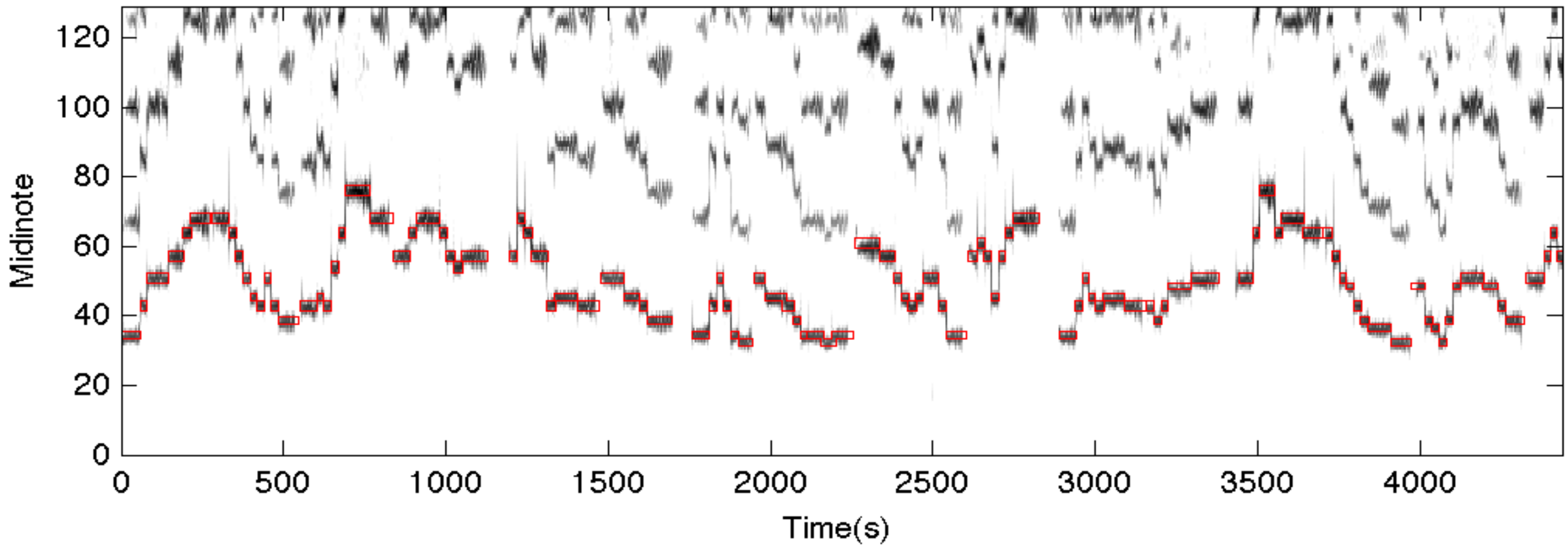
- Robust polyphonic transcription is still an unsolved problem
- However, there is a workaround when a score is available:
  - align the MIDI score to audio
  - use the MIDI score to guide the signal processing analysis to estimate accurate frequency information
- Work being undertaken by Dan Eillis with Christine Smit

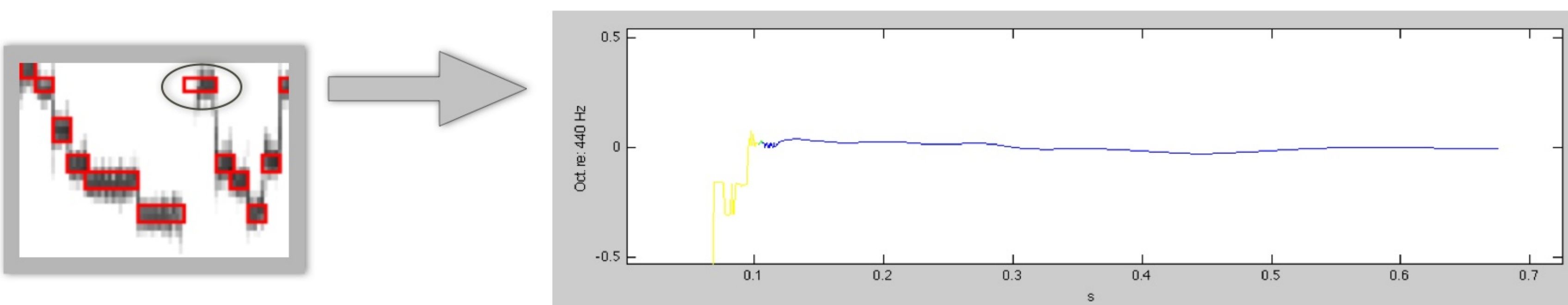
EXTRACTING PITCH DATA

- The perceived pitch over the duration of the note can be calculated as the geometric mean of the frame-wise fundamental frequency estimates (Brown and Vaughn 1996)

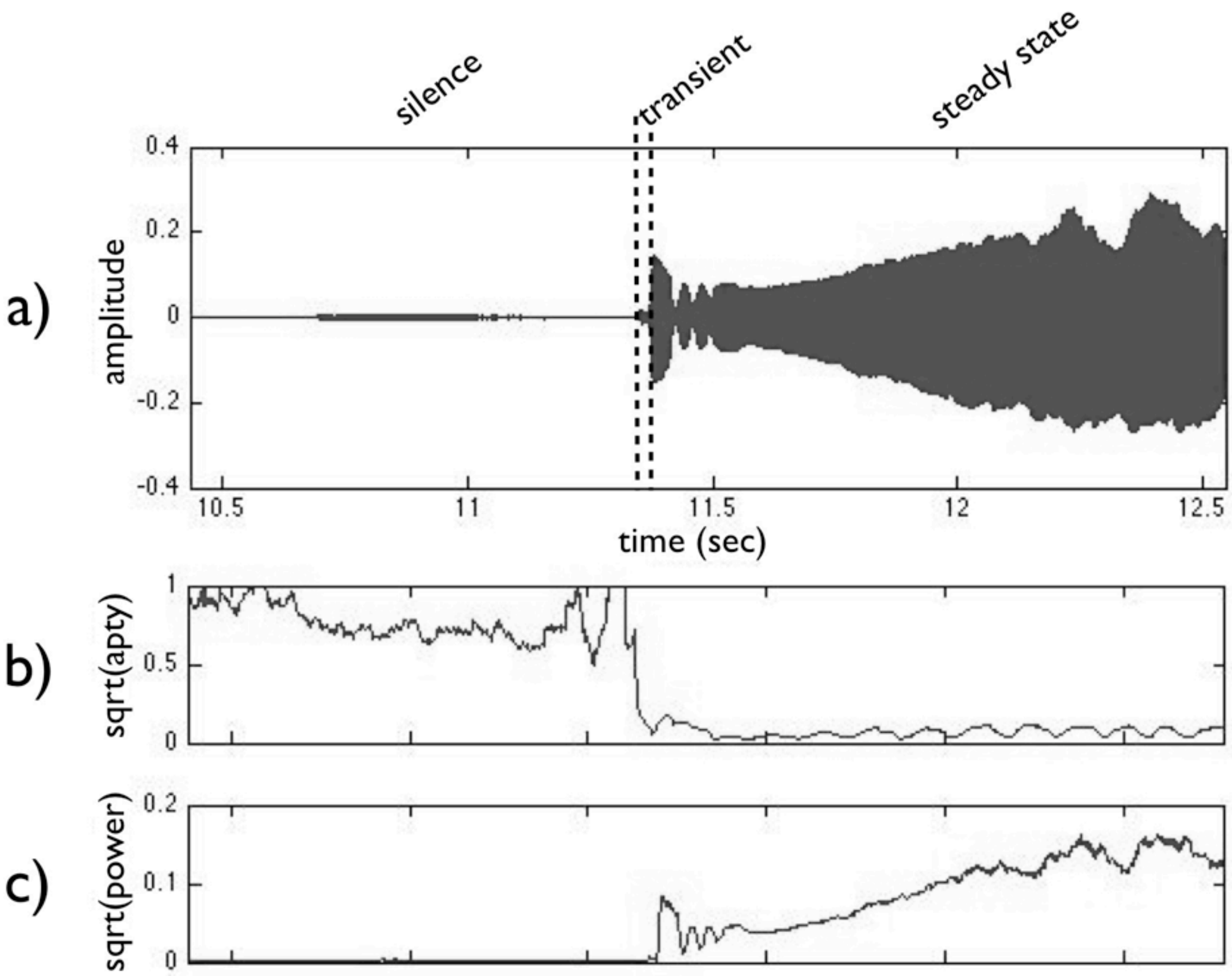


ISSUE WITH DYNAMIC TIME WARPING APPROACH

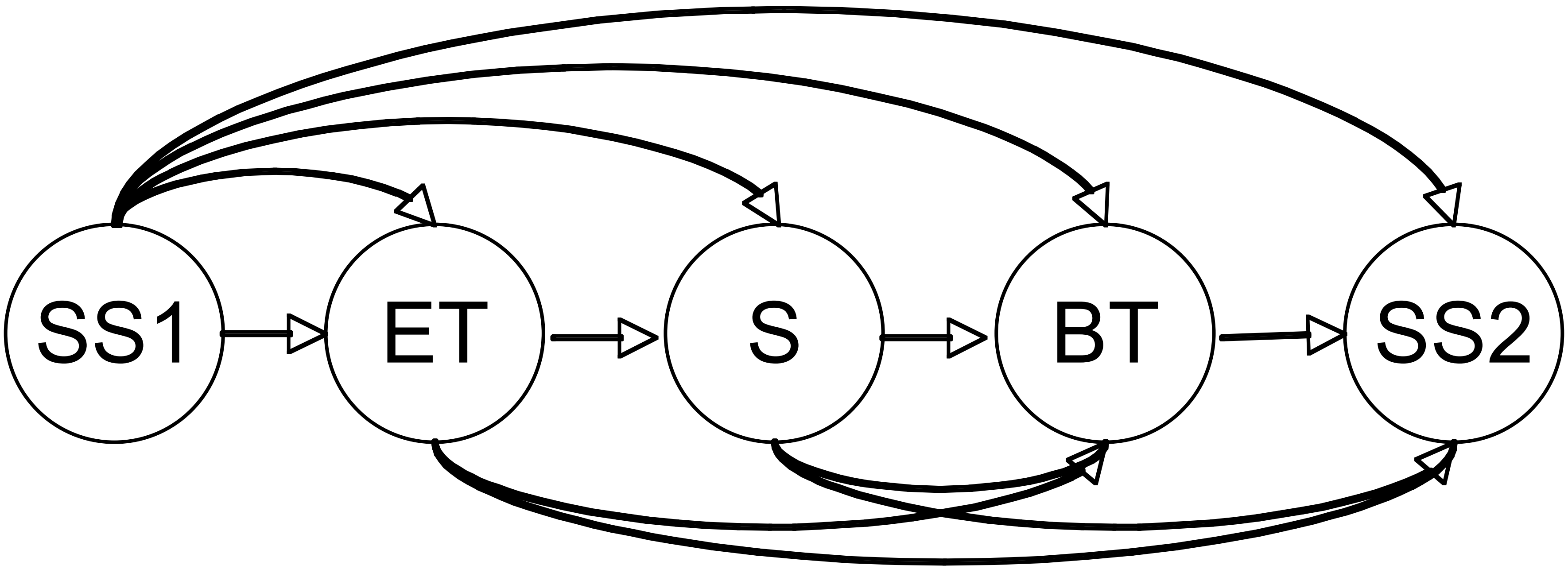




ACOUSTICAL FEATURES OF THE SINGING VOICE



State Sequence Diagram



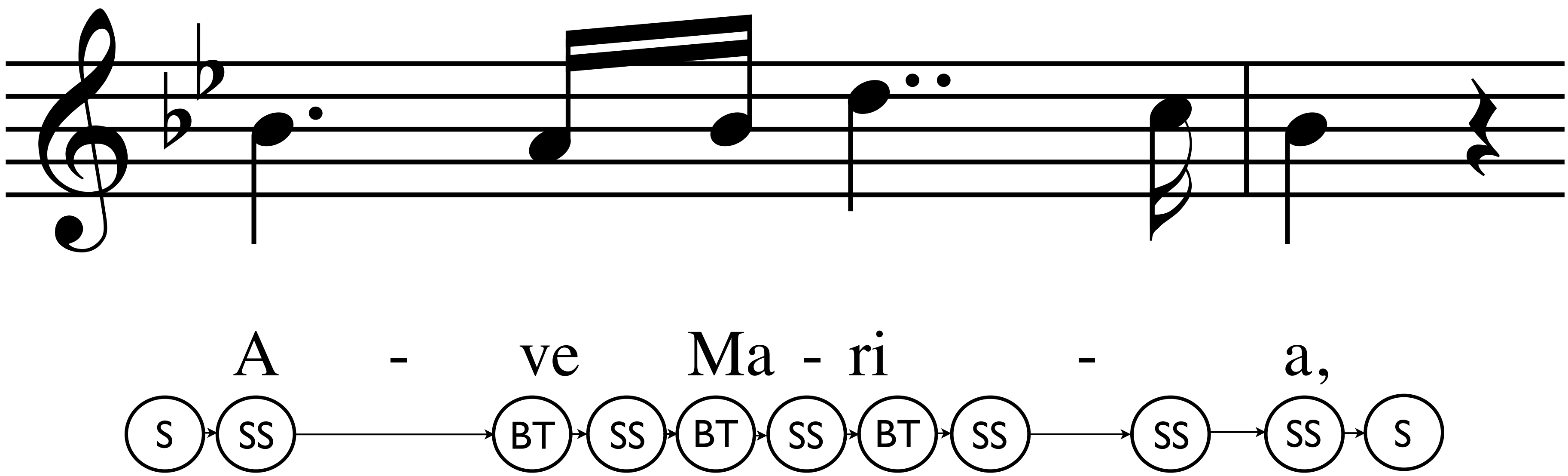
SS = Steady State

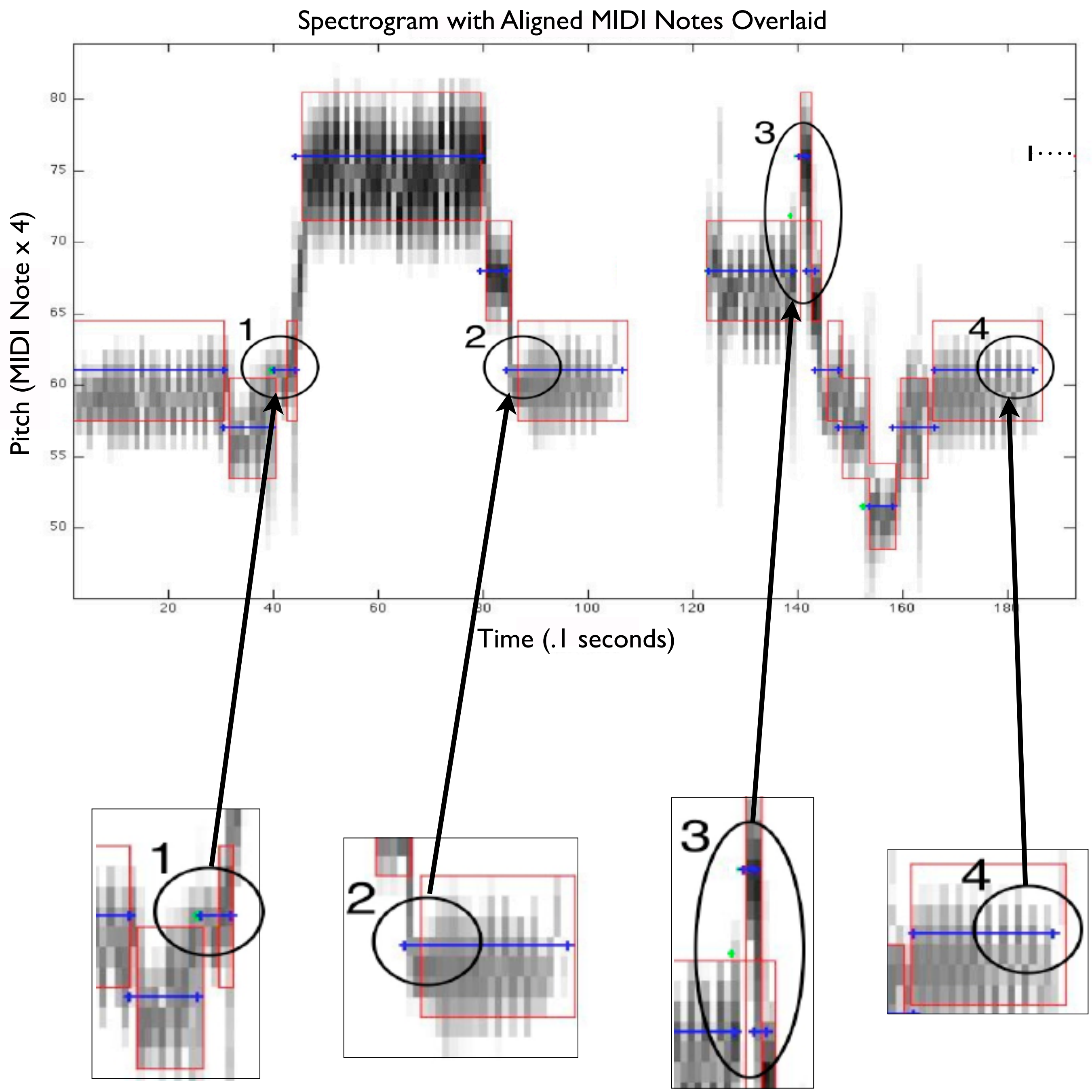
ET = Ending Transient

S = Silence

BT = Beginning Transient

Modified State Sequence Diagram





Spectrogram with Aligned MIDI Notes Overlaid
Pitch (MIDI Note x 4)
Time (.1 seconds)
80
75
70
65
60
55
50
20
40
60
80
100
120
140
160
180
1
2
3
4

Improvements to DTW-based Approach

- Mean error for note onsets and offsets
  - Dynamic Time Warping: 52 ms
  - General state sequence mean error: 48 ms
  - Modified state sequence mean error: 28 ms

- Details of this algorithm is available in Devaney et al. 2009

Vocal Intonation Studies

- Seashore and colleagues work at the University of Iowa (1920s and 30s)
- "Speech, Music, and Hearing" group, Royal Institute of Technology, Stockholm (1980s-present)
- Prame's study of vibrato and intonation in solo singers (1997)

INTONATION IN SOLO VOCAL PERFORMANCE

Original

INTONATION IN SOLO VOCAL PERFORMANCE

Original

Quantized

‘Ave Maria’ Experiment

- Purpose
  - To explore whether there is a relationship between melodic interval tuning in solo singing and harmonic function
- Subjects
  - Six undergraduate sopranos from McGill University
- Task
  - Three performances of the ‘Ave Maria’ *a cappella*
  - Three with recorded accompaniment

'Ave Maria' Experiment

- Analysis of singing
  - Mean fundamental frequency across duration of the note
  - Evolution of the fundamental frequency over the duration of the note
    - Slope (1st Discrete Cosine Transform Coefficient)
    - Curvature (2nd Discrete Cosine Transform Coefficient)

‘Ave Maria’ Experiment

- Analysis of singer’s self-consistency and intra-singer consistency under various conditions
    - A-Bb *a cappella* and accompanied
    - Bb-A *a cappella* and accompanied
    - other semitones ascending *a cappella* and accompanied
    - other semitone descending *a cappella* and accompanied


A - ve Ma - ri - - - - a, Gra - ti - a ple na Ma - ri - a gra ti - a
4
ple - - - na, Ma - ri - - - a gra - ti - a ple - na A - ve A - ve Do - mi-
6
nus, Do - mi - nus te - cum Be - ne - dic - ta tu in mu - li - e - ri - bus et
8
be - ne - di - - - - - ctus, et be - ne - dic - tus fru - ctus ven - tris, ven - tris
10
tu - i Je - - - - - sus. A - ve - Ma - ri - - - a!

'Ave Maria' Experiment

- Fundamental frequency analysis
  - Weak effects for singer identity and accompaniment
  - No effects were found leading tone function or intervallic direction
- Slope
  - Weak effects for direction, accompaniment, and singer identity
- Curvature
  - Weak effects for singer identity

'Ave Maria' Experiment

- Results
  - No observable effects for leading tone function
  - General tendency for small semitones
- Future Work
  - Extend study with a similarly sized group of professional singers

Conclusions

- Automatic extraction of performance data allows for a larger number of performances to be studied
- This talk presented an algorithm that automatically identifies pitch, onsets and offsets for recordings where a symbolic representation of the score is available
- It also described some results of a study of intonation in solo vocal performance that made use of this algorithm

ACKNOWLEDGEMENTS

- My collaborators
  - Ichiro Fujinaga (McGill University)
  - Dan Ellis (Columbia University)
  - Michael Mandel (Université de Montréal)
- Center 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)

Thank You

Questions?

Bengtsson, I., and A. Gabrielsson. 1980. Methods for analyzing performance of Musical rhythm. *Scandinavian Journal of Psychology.* 21: 257-68.

Brown, J. C., and K.V. Vaughn. 1996. Pitch center of stringed instrument vibrato tones. *Journal of the Acoustical Society of America*. 100:3, 1728-35.

Clarke, E. 1989. The perception of expressive timing in music. *Psychological Research.* 51. 2–9.

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

de Cheveigné, A. and H. Kawahara. 2002. YIN, a fundamental frequency estimator for speech and music. *Journal of the Society of the Acoustical Society of America.* 111 (4): 1917–30.

Devaney, J., M. I. Mandel, D. P. W. Ellis. 2009. Improving MIDI-audio alignment with acoustic features. In *Proceedings of the IEEE Workshop on Audio and Signal Processing to Audio and Acoustics*.

Gabrielsson, A. 1999. The performance of music. In D. Deutsch (Ed.) *The Psychology of Music* (2nd ed.). San Diego, CA: Academic Press. 501–602.

Gabrielsson, A. 2003. Music performance research at the millennium. *Psychology of Music*, 31, 221–72.

Grubb, L., and R. Dannenberg. 1997. A stochastic method of tracking a vocal performer. In *Proceedings of the 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.

Prame, E. 1997. Vibrato extent and intonation in professional western lyric singing. *Journal of the Acoustical Society of America.* 102(1): 616–21.

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.

Repp, B. H. 1992. Diversity and commonality in music performance: an analysis of timing microstructure in Schumann's 'Träumerei'. *Journal of the Acoustical Society of America.* 92: 2546–68.

Seashore, C. E. 1938. *Psychology of music*. New York: Dover Publications.

Todd, N.P.M. (1992). The dynamics of dynamics: A model of musical expression. *Journal of the Acoustical Society of America*. 91: 3540–50

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 International Computer Music Conference*. 199–200.

HMM Transition Probabilities

- Self-loop probabilities were estimated from the average duration of each state in hand-labeled recordings of vocal pieces with latin text
- Non-self probabilities were estimated from summary statistics of musical scores of vocal pieces with latin text
  - Transient state transition probabilities were set to reflect the likelihood syllables beginning and ending with consonants in latin text
  - Silence transition probabilities were based on the average frequency of rests - legato singing styles was assumed

Observations

- Observations were the square root of aperiodicity and power estimates from the YIN algorithm (de Cheveigné and Kawahara 2002)
- YIN was run on audio sampled at 44,100 with a frame size of 10ms and a hop size of 0.7ms
- Mean and covariance values were calculated by isolating examples of each state from recordings of different singers
  - 2.25s of audio were used to calculate the means and variances for silence, 13.4s for steady state, 0.47s for transients, and 3.83s for breath.

Accuracy of improved alignment method compared to a dynamic time-warping alignment in milliseconds

Percentile	2.5	25	50	75	97.5
Dynamic Time Warping	3.2	32.6	52.3	87.9	478.7
HMM (General state sequence)	1.6	13.1	41.8	88.8	564.1
HMM (State sequence. adapted to lyrics)	1.6	13.1	27.8	78.0	506.0